

Some statistical results in high-dimensional dependence modeling

Alexis Derumigny

▶ **To cite this version:**

Alexis Derumigny. Some statistical results in high-dimensional dependence modeling. Statistics Theory [stat.TH]. Université Paris Saclay (COmUE), 2019. English. NNT : 2019SACLG002 . tel-02144029

HAL Id: tel-02144029
https://pastel.hal.science/tel-02144029

Submitted on 29 May 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

université PARIS-SACLAY

ENSAE ParisTech
École nationale de la statistique et de l'administration économique

NNT : 2019SACLG002

Thèse de doctorat

Contributions à l'analyse statistique des modèles de dépendance en grande dimension

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Ecole nationale de la statistique et de l'administration économique (ENSAE ParisTech)

Ecole doctorale n°574 - École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 15 mai 2019, par

ALEXIS DERUMIGNY

Composition du Jury :

Jean-David Fermanian Professeur, CREST-ENSAE	Directeur de thèse
Anne-Laure Fougères Professeure, Université Claude Bernard Lyon 1	Présidente du jury
Ivan Kojadinovic Professeur, Université de Pau et des Pays de l'Adour	Rapporteur
Matthieu Lerasle Chargé de recherche CNRS, Université Paris-Sud	Examinateur
Dominique Picard Professeure, Université Paris 7	Examinateur
Alexandre Tsybakov Professeur, CREST-ENSAE	Co-directeur de thèse
Marten Wegkamp Professeur, Cornell University	Rapporteur

Contents

Résumé substantiel de la thèse en français **9**

1 Introduction **13**
1.1 Estimation of the conditional mean: linear regression and related methods 14
1.1.1 Least-squares estimators and penalization 14
1.1.2 Adaptivity to σ using two square-root estimators 15
1.1.3 Robustness to outliers using the Median-of-Means approach 16
1.2 Copulas and conditional dependence modeling 18
1.2.1 Distributions with given margins 18
1.2.2 Inference of copulas models 19
1.2.3 Conditional copulas and the simplifying assumption 20
1.2.4 Kendall's tau: a measure of dependence, and its conditional version 23
1.2.5 Estimation of the conditional Kendall's tau 25
1.3 Other topics in inference 28
1.3.1 Estimation of a regular conditional functional by conditional U-statistic regression . 28
1.3.2 About confidence intervals for ratios of means 29
Publications List 30

I Linear regression **31**

2 Improved bounds for Square-root Lasso and Square-root Slope **33**
2.1 Introduction 33
2.2 The framework 35
2.3 Optimal rates for the Square-Root Lasso 35
2.4 Adaptation to sparsity by a Lepski-type procedure 37
2.5 Algorithms for computing the Square-root Slope 41
2.6 Optimal rates for the Square-Root Slope 41
2.7 Proofs 43
2.7.1 Preliminary lemmas 43
2.7.2 Proof of Theorem 2.1 46
2.7.3 Proofs of the adaptive procedure 48
2.7.3.1 Proof of Theorem 2.3 48
2.7.3.2 Proof of Lemma 2.4 49
2.7.3.3 Proof of Lemma 2.5 50
2.7.4 Proof of Theorem 2.8 51

3 Robust-to-outliers simultaneous inference and noise level estimation using a MOM approach 55
3.1 Introduction . . . 55
3.2 Results in the high-dimensional linear regression framework . . . 56
3.3 A general framework . . . 59
3.4 Technical lemmas . . . 62
3.5 Control of the supremum of $T_{K,\mu}(g, \chi, f^*, \sigma^*)$ on each $F_i^{(\kappa)}$ . . . 65
3.5.1 Preliminaries . . . 65
3.5.2 Proof of the first assertion of Lemma 3.12 . . . 66
3.5.3 Proof of the second assertion of Lemma 3.12 . . . 67
3.6 Proof of Lemma 3.11 . . . 67
3.6.1 Bound on $F_1^{(\kappa)}$ . . . 67
3.6.2 Bound on $F_2^{(\kappa)}$ . . . 67
3.6.3 Bound on $F_3^{(\kappa)}$ . . . 68
3.6.3.1 Case $||f - f^*||_{L_P^2} \leq r(\rho_K)$ . . . 68
3.6.3.2 Case $||f - f^*||_{L_P^2} > r(\rho_K)$ . . . 69
3.6.4 Bound on $F_4^{(\kappa)}$ . . . 69
3.6.5 Bound on $F_5^{(\kappa)}$ . . . 70
3.6.6 Bound on $F_6^{(\kappa)}$ . . . 70
3.6.6.1 Case $||f - f^*||_{L_P^2} \leq r(\rho_K)$ . . . 71
3.6.6.2 Case $||f - f^*||_{L_P^2} > r(\rho_K)$ . . . 71
3.6.7 Bound on $F_7^{(\kappa)}$ . . . 72
3.6.8 Bound on $F_8^{(\kappa)}$ . . . 73
3.6.9 Bound on $F_9^{(\kappa)}$ . . . 73
3.6.9.1 Case $||f - f^*||_{L_P^2} \leq r(\rho_K)$ . . . 73
3.6.9.2 Case $||f - f^*||_{L_P^2} > r(\rho_K)$ . . . 74
3.7 Proofs of main results . . . 75
3.7.1 Proof of Theorem 3.4 . . . 75
3.7.2 Proof of Theorem 3.1 . . . 76

II Conditional copula estimation 77

4 About tests of the "simplifying" assumption for conditional copulas 79
4.1 Introduction . . . 79
4.2 Tests of the simplifying assumption . . . 83
4.2.1 "Brute-force" tests of the simplifying assumption . . . 83
4.2.2 Tests based on the independence property . . . 87
4.2.3 Parametric tests of the simplifying assumption . . . 89
4.2.4 Bootstrap techniques for tests of $\mathcal{H}_0$ . . . 90
4.2.4.1 Some resampling schemes . . . 91
4.2.4.2 Bootstrapped test statistics . . . 92
4.3 Tests with "boxes" . . . 95
4.3.1 The link with the simplifying assumption . . . 95
4.3.2 Non-parametric tests with "boxes" . . . 100
4.3.3 Parametric test statistics with "boxes" . . . 100

4.3.4 Bootstrap techniques for tests with boxes . . . 102
4.4 Numerical applications . . . 103
4.5 Conclusion . . . 111
4.6 Notation . . . 111
4.7 Proof of Theorem 4.14 . . . 114
4.7.1 Preliminaires . . . 114
4.7.2 Proof of Theorem 4.14 . . . 118
4.7.3 Proof of Proposition 4.16 . . . 122

5 About kernel-based estimation of conditional Kendall's tau: finite-distance bounds and asymptotic behavior 125
5.1 Introduction . . . 125
5.2 Definition of several kernel-based estimators of $\tau_{1,2|\mathbf{z}}$ . . . 127
5.3 Theoretical results . . . 129
5.3.1 Finite distance bounds . . . 129
5.3.2 Asymptotic behavior . . . 132
5.4 Simulation study . . . 133
5.5 Proofs . . . 134
5.5.1 Proof of Proposition 5.1 . . . 134
5.5.2 Proof of Proposition 5.2 . . . 140
5.5.3 Proof of Proposition 5.3 . . . 141
5.5.4 Proof of Proposition 5.4 . . . 144
5.5.5 Proof of Proposition 5.6 . . . 145
5.5.6 Proof of Proposition 5.7 . . . 147
5.5.7 Proof of Proposition 5.8 . . . 148
5.5.8 Proof of Proposition 5.9 . . . 149
5.5.9 Proof of Lemma 5.17 . . . 151

6 About Kendall's regression 153
6.1 Introduction . . . 153
6.2 Finite-distance bounds on $\hat{\beta}$ . . . 156
6.3 Asymptotic behavior of $\hat{\beta}$ . . . 158
6.3.1 Asymptotic properties of $\hat{\beta}$ when $n \to \infty$ and for fixed n' . . . 158
6.3.2 Oracle property and a related adaptive procedure . . . 160
6.3.3 Asymptotic properties of $\hat{\beta}$ when n and n' jointly tend to $+\infty$ . . . 161
6.4 Simulations . . . 162
6.4.1 Numerical complexity . . . 162
6.4.2 Choice of tuning parameters and estimation of the components of β . . . 163
6.4.3 Comparison between parametric and nonparametric estimators of the conditional Kendall's tau . . . 164
6.4.4 Comparison with the tests of the simplifying assumption . . . 165
6.4.5 Dimension 2 and choice of ψ . . . 167
6.5 Real data application . . . 169
6.6 Proofs of finite-distance results for $\hat{\beta}$ . . . 169
6.6.1 Technical lemmas . . . 169
6.6.2 Proof of Theorem 6.5 . . . 171

6.7 Proofs of asymptotic results for $\hat{\beta}_{n,n'}$ . . . 172
6.7.1 Proof of Lemma 6.7 . . . 172
6.7.2 Proof of Theorem 6.10 . . . 172
6.7.3 Proof of Proposition 6.11 . . . 173
6.7.4 Proof of Theorem 6.12 . . . 173
6.7.5 Proof of Theorem 6.13 . . . 174
6.8 Proof of Theorem 6.14 . . . 175
6.8.1 Proof of Lemma 6.18 : convergence of T_1 . . . 175
6.8.2 Proof of the asymptotic normality of T_4 . . . 179
6.8.3 Convergence of T_6 to 0 . . . 182
6.8.4 Convergence of T_7 to 0 . . . 183
6.8.5 Convergence of T_3 to 0 . . . 183
6.9 Technical results concerning the first-step estimator . . . 184
6.10 Estimation results for a particular sample . . . 185

7 A classification point-of-view on conditional Kendall's tau 187
7.1 Introduction . . . 187
7.2 Regression-type approach . . . 189
7.3 Classification algorithms and conditional Kendall's tau . . . 192
7.3.1 The case of probit and logit classifiers . . . 193
7.3.2 Decision trees and random forests . . . 193
7.3.3 Nearest neighbors . . . 195
7.3.4 Neural networks . . . 197
7.3.5 Lack of independence and its influence on the proposed algorithms . . . 197
7.4 Simulation study . . . 198
7.4.1 Choice of the functions $\{\psi_i\}, i = 1, \ldots, p'$. . . . 199
7.4.2 Comparing different copulas families . . . 200
7.4.3 Comparing different conditional margins . . . 200
7.4.4 Comparing different forms for the conditional Kendall's tau . . . 202
7.4.5 Higher dimensional settings . . . 203
7.4.6 Choice of the number of neurons in the one-dimensional reference setting . . . 203
7.4.7 Influence of the sample size n . . . 203
7.4.8 Influence of the lack of independence . . . 204
7.5 Applications to financial data . . . 206
7.5.1 Conditional dependence with respect to the Eurostoxx's volatility proxy σ . . . 206
7.5.2 Conditional dependence with respect to the variations $\Delta\sigma^I$ of the Eurostoxx's implied volatility index . . . 208
7.6 Conclusion . . . 210
7.7 Some basic definitions about copulas . . . 211
7.8 Proof of Theorem 7.3 . . . 212
7.9 Proof of Theorem 7.4 . . . 214

III Other topics in inference 219

8 Estimation of a regular conditional functional by conditional U-statistic regression 221

8.1 Introduction . . . 221
8.2 Theoretical properties of the nonparametric estimator $\hat{\theta}(\cdot)$ . . . 224
8.2.1 Non-asymptotic bounds for N_k . . . 225
8.2.2 Non-asymptotic bounds in probability for $\hat{\theta}$ . . . 225
8.2.3 Asymptotic results for $\hat{\theta}$ . . . 226
8.3 Theoretical properties of the estimator $\hat{\beta}$ . . . 228
8.3.1 Non-asymptotic bounds on $\hat{\theta}$ . . . 228
8.3.2 Asymptotic properties of $\hat{\beta}$ when $n \to \infty$ and for fixed n' . . . 229
8.3.3 Asymptotic properties of $\hat{\beta}$ jointly in (n, n') . . . 230
8.4 Applications and examples . . . 231
8.5 Notations . . . 233
8.6 Finite distance proofs for $\hat{\theta}$ and $\hat{\beta}$ . . . 233
8.6.1 Proof of Lemma 8.3 . . . 233
8.6.2 Proof of Proposition 8.5 . . . 234
8.6.3 Proof of Theorem 8.8 . . . 239
8.7 Proof of Theorem 8.14 . . . 240
8.7.1 Proof of Lemma 8.20 . . . 241
8.7.2 Proof of the asymptotic normality of T_4 . . . 244
8.7.3 Convergence of T_6 to 0 . . . 246
8.7.4 Convergence of T_7 to 0 . . . 247
8.7.5 Convergence of T_3 to 0 . . . 247

9 Confidence intervals for ratios of means: limitations of the delta method and honest confidence intervals 249
9.1 Introduction . . . 249
9.2 Our framework . . . 252
9.3 Limitations of the delta method . . . 253
9.3.1 Asymptotic approximation takes time to hold . . . 254
9.3.2 Asymptotic results may not hold for sequences of models . . . 254
9.3.3 Extension of the delta method for ratios of expectations in the sequence-of-models framework . . . 255
9.4 Construction of nonasymptotic confidence intervals . . . 258
9.4.1 An easy case: the support of Y is well-separated from 0 . . . 258
9.4.2 Nonasymptotic confidence intervals with no assumption on the support of P_Y . . . 259
9.5 Nonasymptotic CIs: impossibility results and practical guidelines . . . 259
9.5.1 An upper bound on testable confidence levels . . . 260
9.5.2 Practical methods and plug-in estimators . . . 260
9.5.3 A lower bound on the length of nonasymptotic confidence intervals . . . 261
9.6 Numerical applications . . . 261
9.6.1 Simulations . . . 261
9.6.2 Application to real data . . . 263
9.7 Conclusion . . . 264
9.8 Proofs of the results in Sections 9.3, 9.4 and 9.5 . . . 266
9.8.1 Proof of Theorem 9.1 . . . 266
9.8.2 Proof of Theorem 9.2 . . . 267

9.8.3 Proof of Theorem 9.3 . . . 268
9.8.4 Proof of Theorem 9.6 . . . 269
9.8.5 Proof of Theorem 9.5 . . . 270
9.9 Adapted results for Hoeffding framework . . . 271
9.9.1 Concentration inequality in the easy case . . . 271
9.9.2 Concentration inequality in the general case . . . 272
9.9.3 An upper bound on testable confidence levels . . . 273
9.9.4 Proof of Theorems 9.11 and 9.12 . . . 273
9.9.5 Proof of Theorem 9.13 . . . 273
9.10 Additional simulations . . . 275
9.10.1 Gaussian distributions . . . 275
9.10.2 Rule of thumb using $\overline{\alpha}_n$ . . . 275
9.10.3 Student distributions . . . 275
9.10.4 Exponential distributions . . . 280
9.10.5 Pareto distributions . . . 282
9.10.6 Bernoulli distributions . . . 282
9.10.7 Poisson distributions . . . 287

Remerciements **293**

Bibliography **295**

Résumé substantiel de la thèse en français

Dans cette thèse, nous étudions les propriétés de plusieurs modèles statistiques. Dans la majorité des cas, il s'agira de modèles de dépendance, c'est-à-dire de modèles statistiques qui spécifient les interactions entre différentes variables aléatoires.

Le premier modèle étudié est le modèle de régression linéaire en grande dimension. Classiquement, on observe n réplications indépendantes et identiquement distribuées d'une variable aléatoire réelle Y, que l'on cherche à expliquer par p variables explicatives $X_1, \dots, X_p$. On suppose qu'il existe un vecteur $\beta^* \in \mathbb{R}^p$ tel que $Y = \mathbf{X}^T\beta^* + \varepsilon$, où $\mathbf{X} = (X_1, \dots X_p)$ et ε est une variable aléatoire indépendante de $\mathbf{X}$. On s'intéresse plus particulièrement au cas dit "sparse", où le vecteur β^* est composé d'un grand nombre de composante nulles. Autrement dit, on suppose que la sparsité s est faible, où $s := Card\{i : \beta_i^* \neq 0\}$.

Dans ce cadre, il est habituel d'utiliser des estimateurs de type moindres-carrés pénalisés, qui peuvent permettent d'atteindre des vitesses optimales au sens minimax (cf. [12]). Néanmoins, ces estimateurs nécessitent la connaissance a priori de l'écart-type σ^* du bruit ε. Dans le Chapitre 2, nous prouvons que deux estimateurs, le *Square-root Lasso* et le *Square-root Slope* permettent d'atteindre les vitesses optimales à distance finie sans nécessiter la connaissance de σ^*. En outre, le *Square-root Slope* ne nécessite pas non plus la connaissance de la sparsité s. Nous détaillons également des algorithmes permettant le calcul du *Square-root Slope*.

Dans le Chapitre 3, nous proposons des modifications de ces estimateurs, qui leur permettent d'être davantage robustes. En effet, les estimateurs traditionnels à base de moindres carrés sont, tout comme la moyenne empirique, très sensibles aux valeurs aberrantes, et la présence d'une seule valeur aberrante peut considérablement réduire les performances d'un estimateur (cf. [91]). Dans cette optique, nous proposons une version dite MOM (pour médiane-de-moyennes) des estimateurs adaptatifs du précédent chapitre. Ces estimateurs MOM adaptatifs peuvent encore atteindre les vitesses optimales en même temps qu'estimer l'écart-type du bruit σ^*. Nous proposons également un algorithme permettant de calculer ces estimateurs.

La seconde partie de cette thèse est consacrée aux modèles de copules conditionnelles. Ces modèles permettent de mettre en valeur la façon dont la dépendance entre les différentes composantes d'un vecteur $\mathbf{X}$ varie en fonction d'une variable explicative $\mathbf{Z}$. Formellement, par le théorème de Sklar, on peut décomposer la fonction de répartition conditionnelle de $\mathbf{X}$ sachant $\mathbf{Z} = \mathbf{z}$ de la façon suivante

$$\forall \mathbf{x} \in \mathbb{R}^p,\ \forall \mathbf{z},\ F_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{Z} = \mathbf{z}) = C_{\mathbf{X}|\mathbf{Z}}\Big(F_{X_1|\mathbf{Z}}(x_1|\mathbf{Z} = \mathbf{z}), \dots, F_{X_p|\mathbf{Z}}(x_p|\mathbf{Z} = \mathbf{z}) \,\Big|\, \mathbf{Z} = \mathbf{z}\Big),$$

où $F_{\mathbf{X}|\mathbf{Z}}$ est la fonction de répartition conditionnelle jointe de $\mathbf{X}$ sachant $\mathbf{Z}$, pour $i \in \{1, \dots, n\}$, $F_{X_i|\mathbf{Z}}$ est la fonction de répartition conditionnelle de X_i sachant $\mathbf{Z}$ et $C_{\mathbf{X}|\mathbf{Z}}$ est la copule conditionnelle de $\mathbf{X}$

sachant $\mathbf{Z}$.

L'idée principale d'une telle décomposition est de séparer l'influence de $\mathbf{Z}$ sur les variables marginales X_i d'une part ; et l'influence de $\mathbf{Z}$ sur la dépendance entre les composantes de $\mathbf{Z}$, d'autre part. Cette dernière est représentée par la copule conditionnelle de $\mathbf{X}$ sachant $\mathbf{Z}$. Autrement dit, dans le cas général, pour chaque valeur $\mathbf{z}$ de la variable conditionnante $\mathbf{Z}$, il existe une copule conditionnelle $C_{\mathbf{X}|\mathbf{Z}}(\cdot|\mathbf{Z}=\mathbf{z})$ décrivant la dépendance entre les différentes composantes de $\mathbf{X}$ conditionnellement à l'évènement $\mathbf{Z}=\mathbf{z}$. Certains auteurs disent qu'un tel niveau de complexité n'est pas nécessaire, et affirment parfois qu'une modélisation fixe de la copule conditionnelle est préférable. Cela permettrait d'estimer une copule fixe, plutôt qu'une famille (infinie) de copules indexée par le paramètre $\mathbf{z}$. D'autres disent qu'un tel modèle est peu susceptible d'apparaître en pratique, et ne saurait être proche de la réalité en général. Dans le chapitre 4, on développe des tests de cette hypothèse, appelée "hypothèse simplificatrice".

Formellement, cette hypothèse peut s'écrire "$\mathbf{z} \mapsto C_{\mathbf{X}|\mathbf{Z}}(\cdot|\mathbf{Z}=\mathbf{z})$ est une fonction constante", c'est-à-dire que la fonction $C_{\mathbf{X}|\mathbf{Z}}(\cdot|\mathbf{Z}=\mathbf{z})$ ne dépend pas du choix de $\mathbf{z}$. Une première idée naturelle est de construire un test de $\mathcal{H}_0$ basé sur une comparaison entre la copule conditionnelle $C_{I|J}$ estimée avec et sans l'hypothèse simplificatrice. Ces estimateurs seront appelés respectivement $\hat{C}_{s,I|J}$ et $\hat{C}_{I|J}$. Ainsi, en introduisant une certaine distance $\mathcal{D}$ entre copules conditionnelles, un test peut être basé sur la statistique $\mathcal{D}(\hat{C}_{I|J}, \hat{C}_{s,I|J})$. Nous proposons des procédures de ré-échantillonage pour évaluer la loi limite de telles statistiques de test et prouvons la validité d'un schéma de ré-échantillonage semi-paramétrique spécifique.

Si l'hypothèse simplificatrice est rejetée, il faut modéliser la dynamique de la dépendance en fonction de $\mathbf{z}$. Pour ce faire, nous utilisons la version conditionnelle d'un indicateur de dépendance usuel, le tau de Kendall. Pour un entier p fixé, et pour chaque $\mathbf{z} \in \mathbb{R}^p$, le tau de Kendall conditionnel d'un vecteur bivarié $\mathbf{X} := (X_1, X_2)$ sachant un vecteur de covariables $\mathbf{Z} = \mathbf{z}$ est défini par

$$\tau_{1,2|\mathbf{Z}=\mathbf{z}} = \mathbb{P}\big((X_{2,1}-X_{1,1})(X_{2,2}-X_{1,2})>0\big|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}\big) - \mathbb{P}\big((X_{2,1}-X_{1,1})(X_{2,2}-X_{1,2})<0\big|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}\big),$$

où $(\mathbf{X}_1, \mathbf{Z}_1) = (X_{1,1}, X_{1,2}, Z_{1,1}, \ldots, Z_{1,p})$ et $(\mathbf{X}_2, \mathbf{Z}_2) = (X_{2,1}, X_{2,2}, Z_{2,1}, \ldots, Z_{2,p})$ sont deux copies indépendantes de $(\mathbf{X}, \mathbf{Z})$.

On peut noter que le tau de Kendall conditionnel appartient toujours à l'intervalle $[-1, 1]$ et reflète une dépendance positive (si $\tau_{1,2|\mathbf{Z}=\mathbf{z}} > 0$) ou négative (si $\tau_{1,2|\mathbf{Z}=\mathbf{z}} < 0$) entre X_1 et X_2, sachant $\mathbf{Z} = \mathbf{z}$. Contrairement aux corrélations, cette mesure a l'avantage d'être toujours bien définie, même si l'un des X_k, $k = 1, 2$, n'admet pas de moments d'ordre 2. C'est le cas lorsqu'il suit une loi de Cauchy par exemple.

Dans le chapitre 5, nous proposons des estimateurs du tau de Kendall conditionnel utilisant des techniques d'estimation à noyau. Nous prouvons leurs propriétés asymptotiques et à distance finie, sous des conditions faibles. Puis, dans le chapitre 6, nous proposons un modèle de type régression pour le tau de Kendall conditionnel, de la forme $\Lambda(\tau_{1,2|\mathbf{z}=\mathbf{z}}) = \boldsymbol{\psi}(\mathbf{z})^T\beta^*$, où β^* est un paramètre à estimer, Λ est une transformation connue et $\boldsymbol{\psi}$ est un dictionnaire de fonctions. Un tel modèle peut être utile également pour donner une estimation directe d'effets marginaux, du type $\partial\tau_{1,2|\mathbf{z}=\mathbf{z}}/\partial z_1$, qui quantifient l'influence locale d'une des variables sur la dépendance entre X_1 et X_2.

Dans le chapitre 7, nous montrons les liens existants entre l'estimation du tau de Kendall conditionnel et les problèmes de classification. Ainsi, soient $W := 2 \times \mathbb{1}\{(X_{2,1}-X_{1,1})(X_{2,2}-X_{1,2})>0\} - 1$ et $\mathbb{P}\big((X_{2,1}-X_{1,1})(X_{2,2}-X_{1,2})>0\big|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}\big) = \mathbb{P}\big(W=1\big|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}\big) =: p(\mathbf{z})$. Nous pouvons alors

remarquer que la prédiction de la concordance/discordance parmi les paires d'observations $(\mathbf{X}_1, \mathbf{X}_2)$ sachant $\mathbf{Z}$ peut être vue comme un problème de classification de ces paires. Si un modèle est capable d'estimer la probabilité conditionnelle d'observer des paires concordantes d'observations, alors il est capable d'estimer le tau de Kendall conditionnel. De telles probabilités conditionnelles sont justement les sorties données par la plupart des algorithmes de classification usuels. Ainsi, la plupart des classifieurs peuvent être utilisés ici (par exemple les classifieurs linéaires, les arbres de décision, les forêts aléatoires, les réseaux de neurones, et ainsi de suite), mais appliqués ici à des paires d'observations.

Dans la troisième partie de cette thèse, nous abordons deux autres thèmes liés à l'inférence des modèles statistiques. Dans le chapitre 8, nous montrons comment les techniques développées dans le chapitre 6 peuvent se généraliser au cas de n'importe quelle U-statistique conditionnelle. Finalement, dans le chapitre 9, nous étudions la construction d'intervalles de confiance pour des ratios de moyennes. Nous montrons que les estimateurs asymptotiques classiques construits à partir du théorème central limite et de la delta-méthode peuvent ne pas marcher dans certains régimes. Nous proposons des indicateurs fondés sur des théorèmes d'impossibilité pour détecter de tels cas, et proposons d'autres intervalles de confiance, uniformément valides sur des classes de distributions.

Chapter 1

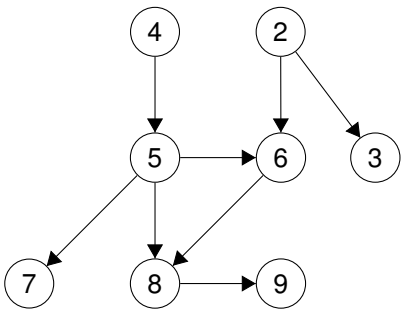
Introduction

In the era of big data, statistics and data science are increasingly useful and necessary to analyze the huge volume of information that is available. Of primary importance are the tasks of estimation and prediction: estimation because we want to infer some parameter or functional that determine the unknown data generating process, and prediction because we want to learn from some datasets to deduce some information about the future.

Often, several random variables $X_1, \dots, X_d$ are available and interact with each other. Therefore it is obvious that inference from the joint law of $\mathbf{X} := (X_1, \dots, X_d)$ should take into account the potential dependence between the different components of $\mathbf{X}$.

Indeed, in applications, many datasets are fundamentally multivariate. Let us describe shortly a few examples: in finance, we may want to model the joint distribution of the returns of several assets ; in hydrology, it is important to model the joint distribution of the characteristic of several rivers ; in social sciences, usually several variables are available, in the individual level (income, wealth, age and so on) as well as in the aggregated level (GDP, unemployment rate, average life expectancy and so on) ; in biostatistics, interactions between different genes or different characteristics of the patients may be of interest.

In this thesis, we study two main dependence modeling frameworks : the high-dimensional sparse linear regression, studied in Chapters 2 and 3, and the conditional dependence framework, studied in Chapters 4, 5, 6 and 7. Other related topics are discussed in the last part : estimation of conditional U-statistics in Chapter 8, and confidence intervals for ratios of means in Chapter 9.



2. Improved bounds for Square-root Lasso and Square-root Slope
3. Robust-to-outliers simultaneous inference and noise level estimation
4. About tests of the simplifying assumption
5. About kernel-based estimation of the conditional Kendall's tau
6. About Kendall's regression
7. A classification point-of-view of the conditional Kendall's tau
8. Estimation of a regular conditional functional
9. Confidence intervals for ratios of means

Figure 1.1: Links between the different chapters

1.1 Estimation of the conditional mean: linear regression and related methods

1.1.1 Least-squares estimators and penalization

One of the oldest well-known dependence model is the linear regression. In such cases, the statistician tries to explain specifically one of the variables, called for example Y, using the other variables $X_1, \dots, X_p$. As usual, we assume that there is a true parameter $\beta^* \in \mathbb{R}^p$ such that $Y = \mathbf{X}^T\beta^* + \varepsilon$, where $\mathbf{X} := (X_1, \dots, X_p)$ is the vector of explanatory variables and ε is a random variable representing the uncertainty. Indeed, in the general case, there is no reason why there would be enough information in $\mathbf{X}$ to completely explain the variable Y. We observe $n > 0$ i.i.d. replications of $(\mathbf{X}, Y)$, denoted by $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$, and we define the design matrix as $\mathbb{X} := (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$ and $\mathbf{Y} := (Y_1, \dots, Y_n)^T$. As our goal is to infer the unknown parameter β^*, we can invoke the usual least-squares estimator $\hat{\beta}^{LS} := \arg\min_{\beta \in \mathbb{R}^p} ||\mathbf{Y} - \mathbb{X}\beta||_n^2$, where $||u||_n^2 := (1/n)\sum_{i=1}^n u_i^2$ for a vector $\mathbf{u} \in \mathbb{R}^n$.

If $p > n$, the estimator $\hat{\beta}^{LS}$ may not be relevant. Often, it is no longer unique, and may lead to a perfect reconstruction of the signal Y as an artifact of lack of observations due to the small size n. In this case, additional assumptions have to be made about the model to make the task of inference easier. The most usual assumption is the so-called sparsity of the vector β^*: a vector β^* is said to be s-sparse for a given integer $s > 0$ if the number of nonzero coefficients of β^*, denoted by $|\beta^*|_0 \leq s$. In other words, we assume that Y is not affected by all the variables $X_1, \dots X_p$, but only by s of them. Obviously, the statistician does not have knowledge of the precise set of variables that are relevant. For example, in biostatistics, the statistician may assume that only a few genes are relevant without knowing which ones.

As a consequence of this sparsity assumption, we do not want to have β "too big", in the sense that large values of β should be seen as less appropriate, even if they seem to lead to a better fit. Therefore, we introduce penalized least-squares estimators of β^*, of the form

$$\hat{\beta}^{pen} := \arg \min_{\beta \in \mathbb{R}^p} ||\mathbf{Y} - \mathbb{X}\beta||_n^2 + pen(\beta),$$

where $pen(\cdot)$ is a real-valued function of $\beta \in \mathbb{R}^p$. Choice of the penalization function is not an easy task, and different alternatives are possible. The penalization $pen(\beta) := \lambda \times |\beta|_0$ with a tuning parameter λ seems to be a natural choice, but this estimator is NP-hard to compute, meaning that it is impossible to compute in polynomial time.

The Lasso estimator [135] uses a convex relaxation of this penalization, with $pen(\beta) := \lambda \times |\beta|_1$, where $|\cdot|_1$ is the l_1-norm, i.e. the sum of the absolute values of the coefficients. Bellec, Lecué and Tsybakov [12] have shown that it is possible to choose a tuning parameter $\lambda^{Lasso,\, opt.}$ such that the Lasso estimator with this choice of λ attains the optimal minimax rates of prediction and estimation in this sparse framework, under an assumption on the design matrix $\mathbb{X}$.

Nevertheless, the tuning parameter $\lambda^{Lasso,\, opt.}$ depends on the sparsity s, i.e. the number of non-zero coefficients of β, that is unknown in practice. Bellec, Lecué and Tsybakov [12] provide two possibilities to solve this issue. The first solution is to use a Lepski-type procedure to aggregate all the Lasso estimators for every choice of s. As a result, the aggregated estimator is adaptive to s. The second solution consists in using a different estimator, the Slope.

The Slope estimator [20] is another penalized least-squares estimator, whose penalty is defined by $|\beta|_* := \sum_{j=1}^p \lambda_j |\beta|_{(j)}$, with tuning parameters $\lambda_1 \geq \dots \geq \lambda_p > 0$, defining $|\beta|_{(j)}$ as the j-th largest

component of $|\beta|$. The intuition behind the Slope estimator is simple: larger components of β should be more penalized than smaller ones. Bellec, Lecué and Tsybakov [12] have shown that there exists optimal tuning parameters $\lambda_1^{Slope,\,opt.}, \dots, \lambda_p^{Slope,\,opt.}$ such that the Slope estimator also attains the optimal minimax rates of estimation and prediction.

1.1.2 Adaptivity to σ using two square-root estimators

Both the Lasso and the Slope estimators attains optimal rates, but the chosen tuning parameter depends on σ, the standard deviation of the noise ε. This may cause a problem since there is no reason why σ should be known in practice, making both estimators impossible to compute due to a lack of knowledge about σ. Looking closer at both estimators, it appears that the "optimal" tuning parameters are both proportional to σ. We can therefore rewrite those estimators as $\hat{\beta}^{pen} := \arg\min_{\beta\in\mathbb{R}^p} ||\mathbf{Y} - \mathbb{X}\beta||_n^2 + \sigma \times pen(\beta)$, where $pen(\beta)$ is a "standardized" version of the penalty.

If we knew the true β^*, we could replace σ by an oracle estimator $\hat{\sigma}^{oracle} := ||Y - \mathbb{X}\beta^*||_n$. It is also possible to replace σ in the minimization program by $||Y - \mathbb{X}\beta||_n$, which is an estimator of σ depending on β. The corresponding minimization problem would be $||Y - \mathbb{X}\beta||_n^2 + ||Y - \mathbb{X}\beta||_n \times pen(\beta)$. Simplifying this expression, we define a family of square-root penalized estimators by

$$\hat{\beta}^{sqrt,\,pen} := \arg\min_{\beta\in\mathbb{R}^p} ||\mathbf{Y} - \mathbb{X}\beta||_n + pen(\beta), \text{ where } ||\mathbf{Y} - \mathbb{X}\beta||_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^T\beta)^2}.$$

The name "square-root" follows from the first part of the objective function $||\mathbf{Y} - \mathbb{X}\beta||_n$, which is the square-root of the usual least-squares criteria. A first member of this family is the square-root Lasso estimator, defined as the square-root estimator using the $|\cdot|_1$ penalization. It was first introduced by Sun and Zhang [134] and Belloni et al. [14] under the name "scaled Lasso". In a similar way, we define the square-root Slope using the penalization $|\beta|_*$.

In Chapter 2, we show that the square-root Lasso estimator $\hat{\beta}^{SQL}$ attains the minimax rates of estimation and prediction under Gaussian noise and a restricted eigenvalue condition about the design matrix $\mathbb{X}$. Indeed, there exists universal constants $C_1, C_2 > 0$, such that for every n large enough, with $\mathbb{P}_{\beta^*}$-probability at least $1-(s/p)^s - (1+e^2)e^{-n/24}$, the estimator $\hat{\beta}^{SQL}$ with the tuning parameter chosen as $\lambda^{SQL,\,opt.} := \gamma\sqrt{(1/n)\log(2p/s)}$ and $\gamma \geq 16 + 4\sqrt{2}$ satisfies

$$||\mathbb{X}(\hat{\beta}^{SQL} - \beta^*)||_n \leq \frac{C_1}{\kappa^2}\sigma\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)},$$
$$\forall q \in [1,2],\ |\hat{\beta}^{SQL} - \beta^*|_q \leq \frac{C_2}{\kappa^2}\sigma s^{1/q}\sqrt{\frac{1}{n}\log\left(\frac{2p}{s}\right)},$$

where $|\cdot|_q$ is the l_q-norm and κ is a constant that only depend on the design $\mathbb{X}$. This estimator is adaptive to σ in the sense that $\lambda^{SQL,\,opt.}$ does not depend on σ anymore, as opposed to the tuning parameters of the usual Lasso and Slope.

Nevertheless, the square-root Lasso estimator is not adaptive to s, in the sense that knowledge of the true sparsity s is necessary to achieve this bound. Using a Lepski-type adaptation of this estimator, we prove that an aggregated estimator of the square-root Lasso achieves optimality in the prediction norm $||\cdot||_n$ or in an estimation norm $|\cdot|_q$, for a fixed $q \in [1,2]$ without any knowledge on s. In fact, we show a more general result: any family of estimators of β depending on s admits an adaptive to s aggregated version that achieve the same rate as the oracle estimator with the true s. Computation of this aggregated version is detailed in Algorithm 2.

Combining the ideas of the Slope and the square-root estimators, the Square-root Slope should be adaptive to both s and σ. We explain how any algorithm to compute the Slope estimator can be adapted to compute the Square-root Slope. Then, we show that there is a universal choice of tuning parameters $\lambda_j^{SQS,\,opt.} := \gamma\sqrt{(1/n)\log(2p/j)}$ independent of s and σ, and γ can be any real greater or equal to $16 + 4\sqrt{2}$. With such a choice of tuning parameters and under an assumption on the design matrix $\mathbb{X}$, there exists universal constants $C_1', C_2' > 0$, such that for every n large enough, with $\mathbb{P}_{\beta^*}$-probability at least $1 - (s/p)^s - (1+e^2)e^{-n/24}$, the estimator $\hat{\beta}^{SQS}$ satisfies

$$||\mathbb{X}(\hat{\beta}^{SQL} - \beta^*)||_n \leq \frac{C_1'}{\kappa'^2}\sigma\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)},$$
$$|\hat{\beta}^{SQL} - \beta^*|_2 \leq \frac{C_2'}{\kappa'^2}\sigma s^{1/2}\sqrt{\frac{1}{n}\log\left(\frac{2p}{s}\right)},$$

where κ' is a constant that only depends on the design $\mathbb{X}$. This shows that the Square-root Slope attains the minimax optimal rates of prediction and estimation in the l_2 norm.

1.1.3 Robustness to outliers using the Median-of-Means approach

Robustness is a natural idea in statistics: it seems interesting to ask what would happen to the estimator if the model is wrong, or if some observations have been "contaminated". Sometimes, the estimators are not robust, in the sense that only one outlier is enough to destroy the whole performance of an estimator. Following [44], we define the breakdown point of an estimator T by :

$$\epsilon^*(T, \mathcal{D}_I) = \min_{m\in\mathbb{N}}\Big\{\frac{m}{n+m} : \sup_{\mathcal{D}_\mathcal{O}:|\mathcal{D}_\mathcal{O}|=m}\big|T(\mathcal{D}_\mathcal{I}\cup\mathcal{D}_\mathcal{O}) - T(\mathcal{D}_\mathcal{I})\big| = +\infty\Big\}, \tag{1.1}$$

for an (uncontaminated) dataset $\mathcal{D}_\mathcal{I}$ of size $|\mathcal{D}_\mathcal{I}| = n > 0$. The breakdown point $\epsilon^*(T, \mathcal{D}_I)$ is the minimum proportion of outliers needed to "break the estimator" with the datset $\mathcal{D}_\mathcal{I}$. In other words, it is the minimum number of points that we need to add to the dataset $\mathcal{D}_\mathcal{I}$ of informative data such that the estimator T can lie arbitrary far away from its previous value $T(\mathcal{D}_\mathcal{I})$. The quantity $\epsilon^*(T, \mathcal{D}_I)$ can be therefore seen as a measure of the robustness of the estimators (see [91] for a review of such measures of robustness and new robust estimators).

Most of the times, the breakdown point of an estimator does not directly depends on the dataset $\mathcal{D}_\mathcal{I}$, but it is often only a function of its size $n = |\mathcal{D}_\mathcal{I}|$. For example, the breakdown point of the empirical mean is $1/(n+1)$, meaning that one needs only to add a single outlier to any dataset to make the empirical mean arbitrary high. The median is another location estimator, with a better breakpoint of $1/2$. Indeed, we need to double the size of the any dataset in order to be able to arbitrary modify its median.

The same problem happens with the Lasso and other least-squares estimators : a single outlier added to any dataset can reduce strongly the performance of the Lasso estimator, as proved in [91]. This problem affects in fact all the previous estimators since they are based on the minimization of the empirical risk $||\mathbf{Y} - \mathbb{X}\beta||_n = \sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i - \mathbf{X}_i^T\beta)^2}$, which is a kind of empirical mean.

In [91], Lecué and Lerasle propose to replace the empirical mean by an empirical median-of-mean (MOM) and use a minimaxisation trick as follows.

1. Given a dataset $\mathcal{D}$ of size n, we can divide it into K blocks $\mathcal{D}_1, \dots \mathcal{D}_K$ of size n/K (assumed to be an integer) corresponding to a partition $\{1, \dots, n\} = B_1 \sqcup \cdots \sqcup B_K$;

2. On each block B_k, we compute the criteria of β against $\tilde{\beta}$

$$P_{B_k}(l_\beta - l_{\tilde{\beta}}) := \frac{1}{|B_k|}\sum_{i\in B_k}(Y_i - \mathbf{X}_i^T\beta)^2 - (Y_i - \mathbf{X}_i^T\tilde{\beta})^2,$$

for every $\beta, \tilde{\beta} \in \mathbb{R}^p$;

3. The global MOM criteria of β against $\tilde{\beta}$ is defined by

$$MOM_K(l_\beta - l_{\tilde{\beta}}) := Median\Big\{P_{B_k}(l_\beta - l_{\tilde{\beta}}),\, k = 1, \dots, K\Big\};$$

4. Finally, the MOM-K estimator of β is defined by

$$\hat{\beta}^{MOM-K} := \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p} \max_{\tilde{\beta}\in\mathbb{R}^p} MOM_K\big(l_\beta - l_{\tilde{\beta}}\big) + \lambda\big(pen(\beta) - pen(\tilde{\beta})\big), \tag{1.2}$$

where λ is a tuning parameter and $pen(\cdot)$ is a penalty function, such as the Lasso or the Slope penalty.

They show that the resulting estimator, called the MOM-Lasso (resp. MOM-Slope) attains the optimal minimax rate of convergence while having a breakdown point of $(K/2)/(n + K/2)$. Indeed, both estimators are robust to up to $K/2$ outliers, because such a small number of outliers could not affect the median in step 3 above. Lecué and Lerasle also propose similar estimators for a more general framework where $Y_i = f(X_i) + \varepsilon$, with f belonging to a space of functions $\mathcal{F}$.

Nevertheless, their estimators are not adaptive to the moments of ε, meaning that, to attain the optimal rates, the tuning parameter λ has to be chosen in a way that depends on the standard deviation of ε, which is unknown in practice. We propose, in Chapter 3, an adaptive version of the MOM-Lasso and MOM-Slope. Moreover, it allows a simultaneous inference of the noise level, i.e. the standard deviation σ of ζ. With the Lasso penalty, this joint estimator is defined by

$$(\hat{\beta}, \hat{\sigma})^{Adaptive-MOM-Lasso} := \operatorname*{arg\,min}_{\beta\in\mathbb{R}^d,\, \sigma>\sigma_{\min}} \max_{\tilde{\beta}\in\mathbb{R}^d,\, \chi>\sigma_{\min}} MOM_K\left(\frac{l_\beta}{\sigma} + \sigma - \frac{l_{\tilde{\beta}}}{\chi} - \chi\right) + \mu\big(|\beta|_1 - |\tilde{\beta}|_1\big), \tag{1.3}$$

where

$$MOM_K\left(\frac{l_\beta}{\sigma} + \sigma - \frac{l_{\tilde{\beta}}}{\chi} - \chi\right) := Median\left\{P_{B_k}\left(\frac{l_\beta}{\sigma} + \sigma - \frac{l_{\tilde{\beta}}}{\chi} - \chi\right),\, k = 1, \dots, K\right\}$$

and, for all $k = 1, \dots, K$,

$$P_{B_k}\left(\frac{l_\beta}{\sigma} + \sigma - \frac{l_{\tilde{\beta}}}{\chi} - \chi\right) := \frac{1}{|B_k|}\sum_{i\in B_k}\frac{(Y_i - \mathbf{X}_i^T\beta)^2}{\sigma} - \frac{(Y_i - \mathbf{X}_i^T\tilde{\beta})^2}{\chi}.$$

We prove that this adaptive version is still robust to up to $K/2$ outliers and still attains the optimal rates of convergence in a non-asymptotic way. Moreover, the optimal choice of the tuning parameter μ that allows to attain these rates does not depend on σ anymore. This means our estimators is also adaptive to σ. We propose also a generalization of this estimator to the framework where $Y_i = f(X_i) + \varepsilon$, and prove a bound on the errors of the estimator $(\hat{f}, \hat{\sigma})$ where $\hat{f}$ is an estimator of f and $\hat{\sigma}$ is an estimator of the standard deviation of ε.

1.2 Copulas and conditional dependence modeling

1.2.1 Distributions with given margins

In the previous section, we studied different ways of estimating adaptively the conditional mean of a variable Y given other variables $X_1, \dots, X_p$. Nevertheless, more general models are needed if one wants to estimate a multivariate law without any specific separation between explained or explanatory variables. More precisely, in our framework, the statistician is given n i.i.d. replications $\mathbf{X}_1, \dots, \mathbf{X}_n$ of a random vector $\mathbf{X}$ in $\mathbb{R}^d$, and our goal is to estimate the law of $\mathbf{X}$. It can be convenient to model use a parametric model for the law of $\mathbf{X}$, so that the values of the estimated parameters can be easily interpreted in application. Often, it can be challenging to find a good parametric model for the data, especially in a multivariate framework.

We will distinguish two kind of parameters: on the first side, marginal parameters, i.e. parameters that only influence the univariate margins $X_1, \dots, X_d$; and on the other side, "pure" dependence parameters. For example, assume that $d = 2$ and that law of $\mathbf{X}$ is bivariate Gaussian with means μ_1, μ_2, standard deviation σ_1, σ_2 and correlation ρ. Then $(\mu_1, \mu_2, \sigma_1, \sigma_2)$ is a vector of marginal parameters and ρ is a pure dependence parameter, meaning that the distributions of X_1 and X_2 do not change with ρ. On the contrary, the covariance $Cov_{1,2} := \sigma_1\sigma_2\rho$ is a somehow mixed parameter that we would like to avoid, since it contained some information about the margins and about the dependence at the same time.

One fruitful idea for inference is a generalization of this idea. This is the fundamental concept of copula modeling, giving general and flexible estimation techniques such as *inference from margins* (see Algorithm 1), where the marginals are estimated first and the dependence is modelled in a second step.

The concept of copula itself comes from Sklar [125] in 1959 (see Sklar [126] and Scheizer [121] for historical references). Probabilists were interested in properties of several classes of distributions, in particular distributions with given margins. For example, if we have d continuous distributions $F_1, \dots, F_d$ on $\mathbb{R}$, how can we construct d-dimensional distributions $F_{1:d}$ whose margins are $F_1, \dots, F_d$?

Theorem 1.1 (Sklar, 1959)**.** *Let $d \geq 2$ be an integer.*

1. *Let $F_{1:d}$ be a distribution function on $\mathbb{R}^d$ with continuous margins $F_1, \dots, F_d$, and $\mathbf{X} \sim F_{1:d}$. Then there exists a distribution C on $[0,1]^d$ with uniform margins, named the copula of $X_1, \dots, X_d$ such that the following equation holds*

$$\forall \mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d,\ F_{1:d}(\mathbf{x}) = C\big(F_1(x_1), \dots, F_d(x_d)\big),$$

and C is given by

$$\forall \mathbf{u} = (u_1, \dots, u_d) \in [0,1]^d,\ C(u_1, \dots, u_d) = F_{1:d}\big(F_1^-(u_1), \dots, F_d^-(u_d)\big),$$

where F_i^- is the inverse function of F_i, for $i = 1, \dots, d$.

2. *Conversely, if $F_1, \dots, F_d$ are continuous distributions on $\mathbb{R}$, and C is a copula (i.e. a continuous distribution on $[0,1]^d$ with uniform margins), then $F_{1:d}$ defined by*

$$\forall \mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d, F_{1:d}(\mathbf{x}) := C\big(F_1(x_1), \dots, F_d(x_d)\big),$$

is a joint distribution on $\mathbb{R}^d$ whose margins are respectively distributed as $F_1, \dots, F_d$ and whose copula is C.

3. *Moreover, C is the joint distribution of $\mathbf{U} = (U_1, \dots, U_d)$ where $U_i := F_i(X_i)$ for $i = 1, \dots, d$ when $\mathbf{X} = (X_1, \dots, X_d) \sim F_{1:d}$.*

Therefore, we have a bijection between the joint cdf $F_{1:d}$ and the decomposition $(F_1, \dots, F_d, C)$. This allows to separate on the one hand the marginal distributions $F_1, \dots F_d$, that can be estimated separately, with possibly different models, and on the other hand the copula C, which summarizes the whole dependence between the components of $\mathbf{X}$. The copula C can be understood as a standardization of the law of $\mathbf{X}$ where all the information about the margins has been removed. Indeed, for every $j \in \{1, \dots, d\}$, $U_j := F_j(X_j)$ follows a uniform distribution on $[0, 1]$ and this is true as long as the marginal distributions F_i are continuous, that will be assumed everywhere in the following.

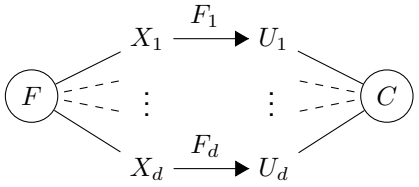


Figure 1.2: Decomposition of a vector $\mathbf{X} = (X_1, \dots, X_d)$ with multivariate cumulative distribution function F, margins $F_1, \dots F_d$ and copula C using Sklar's Theorem (Theorem 1.1).

1.2.2 Inference of copulas models

In practice, the univariate margins $F_1, \dots F_d$ are unknown and need to be estimated in a first, preliminary step. Let $\hat{F}_1, \dots, \hat{F}_d$ be their respective estimators. They can be of any type (parametric, semi-parametric or non-parametric), but a simple choice consists in using the empirical marginal cumulative distribution function, i.e. $\hat{F}_j(t) := (1/n) \sum_{i=1}^n \mathbb{1}\{X_{i,j} \leq t\}$.

From these estimators, we compute the pseudos-observations $\hat{U}_{i,j} := \hat{F}_j(X_{i,j})$ for every $1 \leq i \leq n$, $1 \leq j \leq d$. Indeed, the statistician cannot access the true values $U_{i,j} := F_j(X_{i,j})$ since they are not observe and depend on the unknown cdfs $F_1, \dots F_d$. Strictly speaking, for every $i \in \{1, \dots, n\}$, the random vector $(U_{i,1}, \dots, U_{i,d})$ follows the copula C, but the pseudo-observation $\hat{\mathbf{U}}_i := (\hat{U}_{i,1}, \dots, \hat{U}_{i,d})$ will not follow C. As $n \to \infty$, the estimated marginals cdfs $\hat{F}_1, \dots, \hat{F}_d$ should converge to the true univariate cdfs, so it is still relevant to use the pseudos-observations $(\hat{U}_{i,j})_{i,j}$ to estimate C.

The equivalent of the empirical cdf in this setting is the *empirical copula*, defined by

$$\forall \mathbf{u} := (u_1, \dots, u_d) \in [0, 1]^d, \hat{C}_n(u) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{\hat{\mathbf{U}}_i \leq \mathbf{u}\} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{\hat{U}_{i,1} \leq u_1, \dots, \hat{U}_{i,d} \leq u_d\}. \quad (1.4)$$

Note that the empirical copula $\hat{C}_n$ is not a copula since its margins are not continuous, and therefore not uniform on $[0, 1]$. The empirical copula has been well-studied and converges to the true copula under mild assumptions, see Fermanian et al. [51] and references therein.

Several parametric models of copulas are also available. One of the simplest copula is the Gaussian copula with correlation matrix Σ, defined as the copula of the Gaussian distribution $\mathcal{N}(0, \Sigma)$. One can similarly construct the Student copula as the copula of the multivariate Student distribution with correlation matrix Σ and degrees of freedom ν.

Another important class of copulas is the class of Archimedean copulas. A copula C of dimension 2 is called Archimedean if there exists a function φ convex, continuous, strictly decreasing from $[0, 1]$ to

$[0, +\infty[$ such that $\varphi(1) = 0$ and $\forall (u, v) \in [0, 1]^2$, $C(u, v) = \varphi^{(-1)}\big(\varphi(u) + \varphi(v)\big)$. Archimedean copulas also exist in dimensions strictly larger than 2, under stronger assumptions on the function φ. In practice, we can fix a family of generator $(\varphi_\theta)_{\theta\in\Theta}$ for a low dimension parameter space Θ, and a parameter θ that can be easily estimated. Several examples of such families are given in [106][Table 4.1].

More generally, assume that we have a finite-dimensional parameter space $\Theta \subset \mathbb{R}^p$, and a parametric family of copulas $\mathcal{C} := \{C_\theta, \theta \in \Theta\}$, with densities c_θ. As before, we assume that we are given n i.i.d. replications $\mathbf{X}_1, \dots, \mathbf{X}_n$ of a random vector $\mathbf{X} \in \mathbb{R}^d$, whose copula C_{θ^*} belongs to the parametric family $\mathcal{C}$, for a true unknown parameter $\theta^* \in \Theta$. One of the main techniques for estimating θ^* is the so-called Canonical Maximum Likelihood Estimation, which simply the pseudo-maximum likelihood estimation of the model using the pseudos-observations $\hat{U}_{i,j} := \hat{F}_j(X_{i,j})$, $i = 1, \dots, n$, $j = 1, \dots, p$ where $\hat{F}_j$ is the estimated univariate cdf of X_j. Usually $\hat{F}_j$ is chosen as the empirical cdf of X_j, that allow for any marginal distributions. This gives an estimator $\hat{\theta}$ defined by

$$\hat{\theta} := \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log c_\theta\big(\hat{U}_{i,1}, \dots \hat{U}_{i,n}\big). \tag{1.5}$$

Such an estimator is convergent and asymptotically normal under usual regularity conditions, see Tsukahara [137] for details and a proof.

Algorithm 1: Algorithm for the inference from margins

Input: n i.i.d. observations $\mathbf{X}_1, \dots, \mathbf{X}_n$ of dimension d

for $j \leftarrow 1$ **to** d **do**

estimate the marginal law F_j by the empirical cdf $\hat{F}_j$ or using a parametric model ;

for every $i = 1, \dots, n$, compute the pseudos-observation $\hat{U}_{i,j} := \hat{F}_j(X_{i,j})$;

end

estimate the copula $\hat{C}$ using the pseudos-observations $\hat{U}_{1:n,1:d}$

- either by the empirical copula, defined in Equation (1.4)
- or by Canonical Maximum Likelihood Estimation $\hat{C}_n := C_{\hat{\theta}}$ where $\hat{\theta}$ is defined by Equation (1.5) ;

Output: d marginals $\hat{F}_1, \dots \hat{F}_d$ and a copula $\hat{C}$.

1.2.3 Conditional copulas and the simplifying assumption

We study now a related framework, where the statistician observe i.i.d. replications of a vector $\mathbf{X} = (\mathbf{X}_I, \mathbf{X}_J)$ where $\mathbf{X}_I \in \mathbb{R}^p$ is a vector of conditioned variables and $\mathbf{X}_I \in \mathbb{R}^{d-p}$ is a vector of conditioning variables, in the sense that we want to model the law of $\mathbf{X}_I$ given $\mathbf{X}_J$. In the previous sections, we have separated marginal and dependence parameters of a given distribution F. Similarly, we would like to separate

- *"conditional marginal parameters"*, i.e. parameters linked to the conditional marginals cdfs $F_{j|J} := F_{X_j|J}$ of X_j given J for $j = 1, \dots, p$;
- *"conditional dependence parameters"*, i.e. parameters linked to the conditional copula $C_{I|J}$ of $\mathbf{X}_I$ given $\mathbf{X}_J$.

This conditional copula exists by the conditional version of Sklar's Theorem (Theorem 1.1), by which we can decompose the conditional multivariate cdf $F_{I|J}$ as follows

$$\forall \mathbf{x}_I \in \mathbb{R}^p,\ \forall \mathbf{x}_J,\ F_{I|J}(\mathbf{x}_I|\mathbf{X}_J = \mathbf{x}_J) = C_{I|J}\Big(F_{1|J}(x_1|\mathbf{X}_J = \mathbf{x}_J), \dots, F_{p|J}(x_p|\mathbf{X}_J = \mathbf{x}_J) \,\Big|\, \mathbf{X}_J = \mathbf{x}_J\Big). \quad (1.6)$$

These conditional copulas have been introduced by Patton [111, 112] and in a more general context by Fermanian and Wegkamp [52]. For example, in a time series context, we may have a sequence of random vectors $(\mathbf{X}_t)_t$ indexed by the time $t \in \mathbb{Z}$. To predict one observation using the previous one in a Markov-chain like model, we would need to estimate the conditional law of $\mathbf{X}_{t+1}$ given $\mathbf{X}_t$. This is close to the previous framework, with the formal choice $\mathbf{X}_I := \mathbf{X}_{t+1}$ and $\mathbf{X}_J := \mathbf{X}_t$. In this case, the conditional copula of $\mathbf{X}_{t+1}$ given $\mathbf{X}_t$ can be understood as the prediction of the dependence between the different components of $\mathbf{X}_{t+1}$ given $\mathbf{X}_t$.

Conditional copulas also naturally appear in the so-called vine framework, see [74, 77, 10]. Let us detail this idea. Using Bayes' theorem, one can show that any (unconditional) copula of dimension d can be decomposed using $d(d-1)/2$ bivariate conditional copula. By this term, we mean conditional copulas where the conditioned vector $\mathbf{X}_I$ is of dimension 2, while the conditioning vector $\mathbf{X}_J$ has a dimension between 0 and $d-2$. This decomposition, also called pair-copula construction [1], allows a a very flexible way of constructing any multivariate copula.

Getting back in the classical framework, the conditional copula of an explained random vector $\mathbf{X}_I$ given an explanatory vector $\mathbf{X}_I$ can be used to explain how the dependence among the components of $\mathbf{X}$ can change with the values of the conditioning variable. Indeed, in the general case, the conditional copula of $\mathbf{X}_I$ given $\mathbf{X}_J = \mathbf{x}_J$ does depend on the conditioning variable $\mathbf{x}_J$. Sometimes, to make the inference easier, people assume that the conditional copula is constant with respect to the conditioning variable $\mathbf{x}_J$. This is called the *"Simplifying Assumption"* for a given conditional copula model and may or may not be satisfied in practice, i.e. with a given data-generating process. A visual representation of the simplifying assumption when $d = 2$ is given on Figure 1.3. The general case, where the conditional copula does depend on the conditioning variable $\mathbf{Z}$, is illustrated on Figure 1.4.

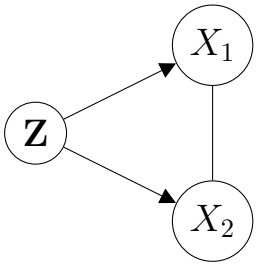


Figure 1.3: The "simplifying assumption": $\mathbf{Z}$ has an influence on the conditional margins X_1 and X_2, but not on the conditional dependence between them.

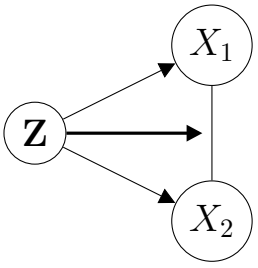


Figure 1.4: The general case: $\mathbf{Z}$ has an influence on the conditional margins X_1 and X_2, and also on their conditional dependence.

In Chapter 4, we provide some tests of this Simplifying Assumption $\mathcal{H}_0$. Formally, it is defined as follows:

$$\mathcal{H}_0 : \text{ the function } \mathbf{x}_J \mapsto C_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J) \text{ is constant.} \quad (1.7)$$

Let us define $C_{s,I|J}$ be the cdf of the vector $\mathbf{Z}_{I|J} := \big((F_{1|J}(X_1|X_J), \dots, F_{p|J}(X_p|X_J)\big)$, called the simplified, or partial copula. We prove in Proposition 4.4 that $\mathcal{H}_0$ is equivalent to the independence be-

tween $\mathbf{Z}_{I|J}$ and $\mathbf{X}_J$, and that if the conditional copula is constant with respect to the conditioning variable, then it is constant to the copula $C_{s,I|J}$. Note that the simplifying assumption does not imply that $C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J) = C_I(\mathbf{u}_J)$ where C_I is the usual (unconditional) copula of $\mathbf{X}_I$ (see Remark 4.1). This means that the simplified copula $C_{s,I|J}$ has no reason to be equal to the usual copula C_I.

Nevertheless, we can have an estimate of $\hat{C}_{s,I|J}$ by averaging kernel-based estimators of the conditional copula $\hat{C}_{I|J}$, or as the empirical cdf or empirical copula of $\mathbf{Z}_{I|J}$. This allows us to adapt any test statistics for the constancy of a function to the test of our simplifying assumption. For any norm $||\cdot||$ on the space generated by the class of all conditional copulas, the test statistics $||\hat{C}_{I|J} - \hat{C}_{s,I|J}||$ can be used to test $\mathcal{H}_0$. For example, we can think of Kolmogorov-Smirnov-type statistics

$$\mathcal{T}^0_{KS,n} := \|\hat{C}_{I|J} - \hat{C}_{s,I|J}\|_\infty = \sup_{\mathbf{u}_I \in [0,1]^p} \sup_{\mathbf{x}_J \in \mathbb{R}^{d-p}} \left|\hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}_{s,I|J}(\mathbf{u}_I)\right|,$$

or Cramer von-Mises-type test statistics

$$\mathcal{T}^0_{CvM,n} := \int \left(\hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}_{s,I|J}(\mathbf{u}_I)\right)^2 w(d\mathbf{u}_I, d\mathbf{x}_J),$$

for some weight function of bounded variation w, that could be chosen as random. Other tests statistics can be constructed using the fact that $\mathcal{H}_0$ is equivalent to the independence between $\mathbf{Z}_{I|J}$ and $\mathbf{X}_J$, recycling usual independence test statistics.

Similar test statistics can be constructed under parametric assumptions. Let $\mathcal{C} := \{C_\theta, \theta \in \Theta\}$ be a family of copulas and assume that for every $\mathbf{x}_J \in \mathbb{R}^{p-d}$, there exists a conditional parameter $\theta(\mathbf{x}_J) \in \Theta$ such that $C_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J) = C_{\theta(\mathbf{x}_J)}$. Then $\mathcal{H}_0$ is equivalent to the existence of a constant parameter θ_0 such that $\forall \mathbf{x}_J$, $\theta(\mathbf{x}_J) = \theta_0$. Test statistics can in this case be constructed as $||\hat{\theta}(\cdot) - \hat{\theta}_0||$, using weighted versions of the CMLE estimators defined in Equation 1.5 and a norm $||\cdot||$ on the space of functions $\mathbb{R}^{d-p} \to \Theta$.

In both cases, asymptotic distributions of these tests statistics are complicated to obtain and may not pivotal, in the sense that they may depend on unknown parameters. Therefore, we propose to use the bootstrap to estimate quantiles of the asymptotic distribution of a given test statistic $\mathcal{T}$. We propose different bootstrap schemes in general nonparametric framework, as well as under parametric assumptions.

The simplifying assumption that we have studied before concerns the constancy of conditional copulas with respect to pointwise conditioning events, i.e. conditioning events of the form $\mathbf{X}_J = \mathbf{x}_J$, for some fixed value of $\mathbf{x}_J$. One can wonder what would happen for different conditioning events, for example if we would conditioned by an event of the form $\mathbf{x}_J \in A_J$ for a borelian subset $A_J \subset \mathbb{R}^{d-p}$. Similarly as in Theorem 1.1 and Equation 1.6, we have the decomposition

$$F_{I|J}(\mathbf{X}_I \leq \mathbf{x}_I|\mathbf{X}_J \in A_J) = C^{A_J}_{I|J}\Big(F_{1|J}(x_1|\mathbf{X}_J \in A_J), \ldots, F_{p|J}(x_p|\mathbf{X}_J \in A_J) \,|\, \mathbf{X}_J \in A_J\Big),$$

for every point $\mathbf{x}_I \in \mathbb{R}^p$ and every subset $A_J \in \mathcal{A}_J$, where $\mathcal{A}_J$ is the set of all Borel subsets of $\mathbb{R}^{d-p}$. This defines implicitly a conditional copula $C^{A_J}_{I|J}$ that depends of the conditioning set A_J in the general case. So, it is tempting to replace $\mathcal{H}_0$ by

$$\widetilde{\mathcal{H}}_0 : C^{A_J}_{I|J}(\mathbf{u}_I|\mathbf{X}_J \in A_J) \text{ does not depend on } A_J \in \mathcal{A}_J, \text{for any } \mathbf{u}_I.$$

Surprisingly and counter-intuitively, $C^{A_J}_{I|J}$ still depends on A_J even under the simplifying assumption $\mathcal{H}_0$! This is due to the non-linear transformation in Sklar's theorem. In fact, we prove that $\tilde{\mathcal{H}}_0$ is

equivalent to a test of independence between $\mathbf{X}_I$ and $\mathbf{X}_J$, which is much stronger than the simplifying assumption $\mathcal{H}_0$. Indeed, $\mathcal{H}_0$ means that $\mathbf{X}_J$ has no influence on the conditional dependence between the components of $\mathbf{X}_I$; if $\mathbf{X}_I$ and $\mathbf{X}_J$ were independent, $\mathbf{X}_J$ would have no influence on the conditional distribution of $\mathbf{X}_J$, meaning no influence on its conditional margins nor on its conditional dependence.

Nevertheless, one can weaken the latter assumption, and restrict oneself to a **finite** family $\overline{\mathcal{A}}_J$ of subsets with positive probabilities. For such a family, we could test the assumption

$$\overline{\mathcal{H}}_0 : A_J \mapsto C_{I|J}^{A_J}(\cdot\,|\mathbf{X}_J \in A_J) \text{ is constant over } \overline{\mathcal{A}}_J.$$

We propose similar test statistics of $\overline{\mathcal{H}}_0$ in both non-parametric and parametric frameworks, as well as adapted bootstrap procedures. We prove the validity of a particular bootstrap resampling schemes. Finally, we illustrate the relevance of all these test statistics and related methods on simulated data.

1.2.4 Kendall's tau: a measure of dependence, and its conditional version

Since the copula is a cdf, it lives in an infinite-dimensional space and can be hard to represent, store (in the memory of a computer) or interpret in applications. Therefore, it may be useful to model the dependence by a number, rather than by a function. One can invoke the usual (Pearson's) coefficient of correlation, but it is not always defined. More precisely, the correlation coefficient does not exists when one of the marginal distribution does not belongs to L^2, for example, if it is a Cauchy distribution. Moreover, the correlation coefficient is no invariant with respect to increasing transformations of the margins, such as a logarithmic transformation.

Several margin-free measures of dependence have been proposed, one of the best-known among them is Kendall's tau [80]. For a bivariate random vector $\mathbf{X} = (X_1, X_2)$, it is defined as

$$\tau_{1,2} := \mathbb{P}\big((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0\big) - \mathbb{P}\big((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) < 0\big), \tag{1.8}$$

where $\mathbf{X}_1 := (X_{1,1}, X_{1,2})$ and $\mathbf{X}_2 := (X_{2,1}, X_{2,2})$ are two i.i.d. replications of $\mathbf{X}$. This can be interpreted as the probability of observing a concordant pair ("the two variables move in the same direction") minus the probability of observing a discordant pair ("the two variables move in opposite direction"). Note that Kendall's tau is always defined for any distribution on $\mathbb{R}^2$ without any moment assumption and lies in the interval $[-1, 1]$. It is invariant by increasing transformations of the marginal distribution and therefore only depends on the copula of $\mathbf{X}$. The link between the Kendall's tau of a given distribution and its copula is in fact explicit and given by $\tau_{1,2} = 4\int_{[0,1]^2} C(u,v)dC(u,v) - 1$. Further properties of Kendall's tau and related dependence measures are detailed in [106].

Kendall's tau can be estimated easily by the empirical proportion of concordant pairs minus the empirical proportion of discordant pairs, giving an estimator $\hat{\tau}$. For most bivariate families of copulas $\mathcal{C} = \{C_\theta, \theta \in \Theta \subset \mathbb{R}\}$, there exists a bijection Ψ between the Kendall's tau and the parameter θ, such that $\tau = \Psi(\theta)$. Then a natural estimator for θ is given by the technique called "Inversion of Kendall's tau", that is $\hat{\theta} := \Psi^{(-1)}(\hat{\tau})$, where $\Psi^{(-1)}$ denotes the inverse of Ψ.

In a bivariate framework, the inference procedure for the law of $\mathbf{X}$ can be therefore divided in four independent steps:

1. Estimation of the first marginal distribution F_1 ;
2. Estimation of the second marginal distribution F_2 ;

3. Choice / selection of the family of copula $\mathcal{C}$, which determines the *shape of the dependence* ;
4. Estimation of the Kendall's tau τ and the corresponding parameter θ, which determines the *strength of the dependence*.

By using Sklar's theorem, we can construct an estimator of the cdf of $\mathbf{X}$ as

$$\hat{F}_{\mathbf{X}}(\mathbf{x}) := C_{\Psi^{(-1)}\left(\hat{\tau}\right)}\Big(\hat{F}_1(x_1)\,,\,\hat{F}_{2|\mathbf{Z}}(x_2)\Big),$$

where $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$.

Assume now that we observe $n > 0$ i.i.d replications $(\mathbf{X}_1, \mathbf{Z}_1), \ldots, (\mathbf{X}_n, \mathbf{Z}_n)$ of a vector $(\mathbf{X}, \mathbf{Z})$ where $\mathbf{X}$ is a bivariate random vector and $\mathbf{Z}$ is a vector of explanatory variables. The conditional law of $\mathbf{X}$ given $\mathbf{Z}$ can be decomposed using the conditional version of Sklar's Theorem, as in Equation 1.6, using the conditional marginals $F_{1|\mathbf{Z}}$ and $F_{2|\mathbf{Z}}$ and the conditional copula $C_{\mathbf{X}|\mathbf{Z}}$. Assume that the two conditional margins $F_{1|\mathbf{Z}}$ and $F_{2|\mathbf{Z}}$ have already been estimated parametrically or non-parametrically (for example, using kernel smoothing techniques). If the simplifying assumption holds, the conditional copula $C_{\mathbf{X}|\mathbf{Z}}$ is constant with respect to $\mathbf{Z}$ and usual (unconditional) copula models can be used to estimate it.

On the contrary, if the simplifying assumption does not hold, then the statistician has to estimate the conditional copula $C_{\mathbf{X}|\mathbf{Z}}$. This corresponds to the estimation of a model of the thick arrow in Figure 1.4. It is always possible to use nonparametric estimators of a conditional copula, for example by kernel-smoothing, but they are very hard to visualize - and a fortiori to interpret. Indeed, even when the dimension of $\mathbf{Z}$ is one, the graph of a conditional copula is an hyper-surface in dimension 4.

A simple idea is the use of the conditional Kendall's tau. Remember that Kendall's tau for a copula C is $\tau(C) = 4\int C dC - 1$. We can therefore extend this definition in a straightforward way as

$$\begin{aligned}\tau_{1,2|\mathbf{Z}=\mathbf{z}} &:= \tau(C_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}) = 4\int C_{\mathbf{X}|\mathbf{Z}=\mathbf{z}} dC_{\mathbf{X}|\mathbf{Z}=\mathbf{z}} - 1 \\ &= \mathbb{P}\big((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) - \mathbb{P}\big((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) < 0 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big),\end{aligned} \tag{1.9}$$

where $(\mathbf{X}_1, \mathbf{Z}_1) = (X_{1,1}, X_{1,2}, Z_{1,1}, \ldots, Z_{1,p})$ and $(\mathbf{X}_2, \mathbf{Z}_2) = (X_{2,1}, X_{2,2}, Z_{2,1}, \ldots, Z_{2,p})$ are two independent versions of $(\mathbf{X}, \mathbf{Z})$. This can be interpreted as the probability of observing a concordant pair conditionally to $\mathbf{Z} = \mathbf{z}$ minus the probability of observing a discordant pair conditionally to $\mathbf{Z} = \mathbf{z}$. We note that the conditioning event $\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}$ in Equation 1.9 is unusual, but it is necessary. Indeed, conditional probabilities of concordance/discordance are only relevant when the two observations follow the same distribution. In this case, it corresponds to the constraint $\mathbf{Z}_1 = \mathbf{Z}_2$.

In a parametric framework, one can use the conditional Kendall's tau to estimate the conditional parameter of a conditional copula. Indeed, let $C_{1,2|\mathbf{Z}}(\cdot|\mathbf{Z} = \mathbf{z})$ be the conditional copula between X_1 and X_2 given $\mathbf{Z} = \mathbf{z}$, $\mathcal{C} := \{C_\theta, \theta \in \Theta\}$ be a family of copulas and assume that for every $\mathbf{z}$, there exists a conditional parameter $\theta(\mathbf{z}) \in \Theta$ such that $C_{1,2|\mathbf{Z}}(\cdot|\mathbf{Z} = \mathbf{z}) = C_{\theta(\mathbf{z})}$. Assume that there is a bijection Ψ such that $\tau(C_\theta) = \Psi(\theta)$, where $\tau(C_\theta)$ denotes the Kendall's tau of the copula C_θ. Then if we are given an estimator $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}$ of the conditional Kendall's tau between X_1 and X_2 given $\mathbf{Z} = \mathbf{z}$, we can estimate the conditional copula between X_1 and X_2 given $\mathbf{Z} = \mathbf{z}$ by

$$\hat{C}_{1,2|\mathbf{Z}}(\,\cdot\,|\,\mathbf{Z} = \mathbf{z}) := C_{\Psi^{(-1)}\left(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}\right)}\,,$$

where $\Psi^{(-1)}$ denotes the inverse function of Ψ. This allows one to separate the inference procedure for the joint law of $(\mathbf{X}, \mathbf{Z})$ in five independent steps:

1. Estimation of the first conditional marginal distribution $F_{1|\mathbf{Z}}$ which determines the influence of the covariate $\mathbf{Z}$ on X_1 ;
2. Estimation of the first conditional marginal distribution $F_{2|\mathbf{Z}}$ which determines the influence of the covariate $\mathbf{Z}$ on X_2 ;
3. Choice / selection of the family of conditional copula $\mathcal{C}$, which determines the *shape of the conditional dependence* ;
4. Estimation of the conditional Kendall's tau $\tau_{1,2|\mathbf{Z}=\cdot}$ and the corresponding parameter $\theta(\mathbf{z})$, which determines the *strength of the conditional dependence* ;
5. Estimation of the law of the conditioning variable $\mathbf{Z}$.

As a consequence, an estimator of the joint cdf $F_{\mathbf{X},\mathbf{Z}}$ can be constructed using these five pieces, by

$$\hat{F}_{\mathbf{X},\mathbf{Z}}(\mathbf{x},\mathbf{z}) := \hat{F}_{\mathbf{Z}}(\mathbf{z}) \times C_{\Psi^{(-1)}\left(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}\right)}\Big(\hat{F}_{1|\mathbf{Z}}(x_1|\mathbf{Z}=\mathbf{z})\,,\,\hat{F}_{2|\mathbf{Z}}(x_2|\mathbf{Z}=\mathbf{z})\Big),$$

where $\mathbf{x}=(x_1,x_2)\in\mathbb{R}^2$ and $\mathbf{z}\in\mathbb{R}^p$.

1.2.5 Estimation of the conditional Kendall's tau

To estimate the conditional Kendall's tau, we will follow three different approaches. In Chapter 5, we show how kernel-based estimators can be used to estimate the conditional Kendall's tau. In Chapter 6, we propose to estimate the conditional Kendall's tau by a regression-type model. Finally, in Chapter 7, we show how classification-based methods can be used for the estimation of the conditional Kendall's tau.

The conditional Kendall's tau can be rewritten as $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = \mathbb{E}[g(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}]$ for a well-chosen choice of g and two i.i.d. replications $(\mathbf{X}_1,\mathbf{Z}_1),(\mathbf{X}_2,\mathbf{Z}_2)$ of a random vector $(\mathbf{X},\mathbf{Z})\in\mathbb{R}^{2+p}$, where $dim(\mathbf{Z})=p>0$. Therefore, in Chapter 5, we propose a corresponding kernel-based estimator of the conditional Kendall's tau. It is given by $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} := \sum_{i=1}^n\sum_{j=1}^n w_{i,n}(\mathbf{z})w_{j,n}(\mathbf{z})g(\mathbf{X}_i,\mathbf{X}_j)$, where $w_{i,n}(z)=K_h(\mathbf{Z}_i-\mathbf{z})$, K is a kernel on $\mathbb{R}^p$ and $h>0$ is the so-called bandwidth.

Our first result is a concentration inequality for the estimated conditional Kendall's tau: we show that for a given $\mathbf{z}$, $t>0$ small enough and $t'>0$, with probability larger than $1-\exp\big(-nh^pt^2/(C_1+C_2t)\big)-\exp\big(-nh^{2p}t'^2/(C_3+C_4t')\big)$, we have $\big|\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}-\tau_{1,2|\mathbf{Z}=\mathbf{z}}\big|<\phi(t,t',h)$, for some explicit constants $C_1,\ldots,C_4$ and an explicit function ϕ, under some regularity conditions. Then, we show its consistency, uniform consistency and asymptotic normality under different regularity assumptions on the data-generating process.

This kernel-based can be nevertheless costly to compute when we want to estimate the whole curve $\mathbf{z}\mapsto\tau_{1,2|\mathbf{Z}=\mathbf{z}}$. We may also want to have a parametric form for the conditional Kendall's tau that would be easier to interpret in applications. Therefore, in Chapter 6, we propose the following regression-type model. For a given link function $\Lambda:[-1,1]\to\mathbb{R}$, and a given basis (ψ_i), we can always do the decomposition

$$\Lambda\big(\tau_{1,2|\mathbf{Z}=\mathbf{z}}\big)=\sum_i\psi_i(\mathbf{z})\beta_i^*=\boldsymbol{\psi}(\mathbf{z})^T\beta^*, \tag{1.10}$$

where the β_i^* are real coefficients, assuming that only a finite number of functions is necessary to reconstruct this transformation of the conditional Kendall's tau. It seems difficult to have intuition about

what kind of functions should be incorporated in the family ψ_i. As a consequence, we advise to include different transformations, such as polynomials, trigonometric or indicator functions, keeping in mind that only a few of them might be relevant to the analysis. This means that the vector β^* is sparse, in the sense that most of its coefficients may be equal to 0. As in the previous section, we would like to use a penalized regression to estimate Model 1.10. A direct estimation of it seems difficult since we do not observe realizations of the conditional Kendall's tau itself. Therefore, we will use a two-step based procedure :

1. Fixing some design points $\mathbf{z}'_1, \dots, \mathbf{z}'_{n'}$ with $n' > 0$, we estimate the conditional Kendall's tau using a kernel-based estimator at each of those points, giving estimators $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i}$ for $i = 1, \dots, n'$.
2. Then we estimate β^* by the following penalized least-squares estimator

$$\hat{\beta}^{PLS} := \arg\min_{\beta \in \mathbb{R}^{p'}} \Big[\frac{1}{n'} \sum_{i=1}^{n'} \big(\Lambda(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i}) - \boldsymbol{\psi}(\mathbf{z}'_i)^T \beta \big)^2 + \lambda |\beta|_1 \Big], \tag{1.11}$$

   where $|\cdot|_1$ is the l_1 norm and λ is a tuning parameter.

Finally, for any point $\mathbf{z}$, the conditional Kendall's tau itself can be estimated by $\Lambda^{-1}(\mathbf{z}^T \hat{\beta}^{PLS})$. As a by-product, the coefficient $\hat{\beta}^{PLS}$ can also be used to estimate marginal effects of the form $\partial \tau_{1,2|\mathbf{Z}=\mathbf{z}} / \partial z_j$ for $j = 1, \dots, p$. This can be useful to have a quantity that measures how the conditional dependence between X_1 and X_2 change with the variable Z_j, for example.

We prove a concentration bound for $\hat{\beta}^{PLS}$ with explicit constants that holds with high probability under regularity assumptions on the data-generating process and on the design matrix. Then we prove the consistency and the asymptotic normality of $\hat{\beta}^{PLS}$, when $n \to \infty$ and n' is fixed. In this situation, our estimator $\hat{\beta}^{PLS}$ does not fulfill the oracle property in the sense that it fails to estimate the true set of relevant variables with probability tending to 1.

We show that a related adaptive procedure can recover the true set of relevant variable. It is defined with the same l_1-penalized criteria, but it uses a random tuning parameter $\lambda = \mu / |\tilde{\beta}|^\delta$, where $\tilde{\beta}$ is a consistent estimator of β^* and $\mu = \mu_{n,n'}$ is a deterministic sequence satisfying some rate assumption.

Getting back to the previous estimator $\hat{\beta}$, we show that, under some regularity assumptions, it is consistent in the framework where both n and n' tend to the infinity. We also compute its asymptotic distribution, and show that the rate significantly improved in this double-asymptotic framework: we obtain $\hat{\beta}^{PLS}_{n,n'} - \beta^* \asymp (nn'h^p)^{-1/2}$, that can be compared with the equivalent $\hat{\beta}^{PLS}_{n,n'} - \beta^* \asymp (nh^p)^{-1/2}$ obtained when n' is fixed.

In Chapter 7, we show how the problem of estimating the conditional Kendall's tau can be rewritten as a classification task. Indeed, remember that it is defined as

$$\tau_{1,2|\mathbf{Z}=\mathbf{z}} = \mathbb{P}\big((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) - \mathbb{P}\big((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) < 0 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big).$$

We now introduce the variable W defined by

$$W = \mathbb{1}\big\{(X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0\big\} - \mathbb{1}\big\{(X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) < 0\big\}.$$

As the distribution of $\mathbf{X}$ is continuous, W belongs to the set $\{-1, 1\}$ almost surely, where $W = 1$ corresponds to the observation of a concordant pair and $W = -1$ to a discordant pair. Using this new variable W, we can rewrite the conditional Kendall's tau as

$$\begin{aligned} \tau_{1,2|\mathbf{Z}=\mathbf{z}} &= \mathbb{P}\big(W = 1 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) - \mathbb{P}\big(W = -1 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) \\ &= 2\mathbb{P}\big(W = 1 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) - 1 = 2p(\mathbf{z}) - 1, \end{aligned} \tag{1.12}$$

where $p(z) := \mathbb{P}\big(W = 1 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big)$. Therefore, estimating the conditional probability of the variable W being equal to 1 or -1 is equivalent to estimating the conditional Kendall's tau. Actually, the prediction of concordance/discordance among pairs of observations $(\mathbf{X}_1, \mathbf{X}_2)$ given $\mathbf{Z}$ can be seen as a classification task of such pairs. If a model is able to evaluate the conditional probability of observing concordant pairs of observations, then it is able to evaluate conditional Kendall's tau, and the former quantity is one of the outputs of most classification techniques.

For an i.i.d. sample $\mathcal{D} := (\mathbf{X}_i, \mathbf{Z}_i)_{i=1,\dots,n}$, we define $W_{(i,j)}$ as

$$W_{(i,j)} := 2 \times \mathbb{1}\{(X_{j,1} - X_{i,1})(X_{j,2} - X_{i,2}) > 0\} - 1 = \begin{cases} 1 & \text{if } (i,j) \text{ is a concordant pair,} \\ -1 & \text{if } (i,j) \text{ is a discordant pair,} \end{cases} \tag{1.13}$$

for every $1 \leq i, j \leq n$, $i \neq j$. A classification technique will allocate a given couple (i,j) into one of the two categories $\{1, -1\}$ (or "concordant versus discordant", equivalently), with a certain probability, given the value of the common covariate $\mathbf{Z}$.

We first consider a parametric approach, assuming a single-index model of the form $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = g\big(\boldsymbol{\psi}(\mathbf{z})^T \beta^*\big)$, similarly as in Model 1.10, where $g = \Lambda^{(-1)}$. We propose to use maximum-likelihood-type methods to estimate the parameter β^*. For one observation $(W_{(i,j)}, \mathbf{Z}_i, \mathbf{Z}_j)$, given $\mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}$, its log-likelihood is

$$\ell_\beta(W_{(i,j)}, \mathbf{z}) := \left(\frac{1 + W_{(i,j)}}{2}\right) \log \mathbb{P}_\beta\Big(W_{(i,j)} = 1 \Big| \mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}\Big) + \left(\frac{1 - W_{(i,j)}}{2}\right) \log \mathbb{P}_\beta\Big(W_{(i,j)} = -1 \Big| \mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}\Big).$$

In practice, when the underlying law of $\mathbf{Z}$ is continuous, there is virtually no couple for which $\mathbf{Z}_i = \mathbf{Z}_j$. Therefore, we will consider a localized "approximated" log-likelihood, based on $(W_{(i,j)}, \mathbf{Z}_i, \mathbf{Z}_j)$ for all pairs (i,j), $i \neq j$. It will be defined as the double sum

$$L_n(\beta) := \frac{1}{n(n-1)} \sum_{i,j;i \neq j} K_h(\mathbf{Z}_i - \mathbf{Z}_j) \ell_\beta(W_{(i,j)}, \tilde{\mathbf{Z}}_{i,j}),$$

where K is a kernel on $\mathbb{R}^p$, h a bandwidth, and $\tilde{\mathbf{Z}}_{i,j}$ a point belonging to a neighborhood of $\mathbf{Z}_i$ or $\mathbf{Z}_j$, for example $\mathbf{Z}_i$ or $\mathbf{Z}_j$ themselves, or $(\mathbf{Z}_i + \mathbf{Z}_j)/2$. Finally, we define a penalized approximate maximum likelihood estimator of β^* by

$$\hat{\beta}^{PAML} := \arg \max_{\beta \in \mathbb{R}^{p'}} L_n(\beta) - \lambda_n |\beta|_1. \tag{1.14}$$

Note that, contrary to the two-step estimators considered previously in Equation (1.11), $\hat{\beta}^{PAML}$ is not a solution of a convex program in general, and therefore may difficult to compute in practice. Nevertheless, it does not need any choice of design points as previously. If g is chosen as the equivalent of the probit or logit link function, then we prove that the optimization program (1.14) is in fact a convex program, that can be easily solved in polynomial time using a classical software for solving (penalized) weighted probit or logit regressions.

More generally, if we are given a dataset $\mathcal{D} := (\mathbf{X}_i, \mathbf{Z}_i)_{i=1,\dots,n}$, we can always construct the dataset of pairs, defined as $\tilde{\mathcal{D}} := (W_k, \tilde{\mathbf{Z}}_k, V_k)_{1 \leq k \leq n(n-1)/2}$, where each pair (i,j) with $i < j$ is indexed by an integer k. Each observation in this dataset is made up of three components. The first is $W_k := W_{i,j}$, i.e. the concordance/discordance of the pair (i,j) ; this is the variable that we want to predict. The explanatory variable of the pair (i,j) is defined by $\tilde{\mathbf{Z}}_k := \tilde{\mathbf{Z}}_{i,j}$. Finally we construct a weight variable $V_k := V_{i,j} = K_h(\mathbf{Z}_i - \mathbf{Z}_j)$, which measures how close $\mathbf{Z}_i$ and $\mathbf{Z}_j$ are, and therefore, it measures how relevant for the estimation of the conditional Kendall's tau the pair (i,j) is. Indeed, remember that in the definition of the conditional Kendall's tau $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$, the conditioning event is of the form $\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}$.

Assume we are given a classification algorithm that takes in input a dataset with an explained variable, a vector of explanatory variables and a weight for each observation. We remark that it corresponds exactly to the caracteristic of the dataset of pairs $\tilde{\mathcal{D}}$ constructed above. Therefore, we can directly apply any classification algorithm on the dataset $\tilde{\mathcal{D}}$. As a result we will obtained an estimated function $\hat{p}(z)$ of the probability of observing a concordant pair given the covariate $\mathbf{z}$. We can plug this estimate to obtain an estimator of the conditional Kendall's tau by $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} = 2\hat{p}(z) - 1$, using Equation (1.10).

Nearly all classification algorithms can be adapted following this principled, and we detail the case of the probit logit, decision trees, random forest, nearest neighbors and neural networks. One can remark that observations in the dataset of pairs $\tilde{\mathcal{D}}$ are not independent. For example, the pair $(1,2)$ and $(1,3)$ are not independent since they share the same first observation $(\mathbf{X}_1, \mathbf{Z}_1)$. Nevertheless, we show that this lack of independence is not too much harmful since most couple of pairs are independent in fact. The performance of all these estimators and their sensitivities to each component of the model is assessed in a simulation study. Finally, we apply all these estimators to a dataset of European stock indices, and compare the conditional dependence between them.

1.3 Other topics in inference

1.3.1 Estimation of a regular conditional functional by conditional U-statistic regression

Remember that Kendall's tau, defined in Equation (1.8), can be also rewritten as

$$\tau_{1,2} := \int \mathbb{1}\big\{(x_{2,1}-x_{1,1})(x_{2,2}-x_{1,2}) > 0\big\} - \mathbb{1}\big\{(x_{2,1}-x_{1,1})(x_{2,2}-x_{1,2}) < 0\big\} d\mathbb{P}_{\mathbf{X}}(x_{1,1}, x_{2,1}) d\mathbb{P}_{\mathbf{X}}(x_{1,1}, x_{2,1}).$$

This means Kendall's tau is what we will call a regular function $\theta(\mathbb{P}_{\mathbf{X}})$ of the law of $\mathbf{X}$. Similarly, the conditional Kendall's tau $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ can be rewritten as $\theta(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}})$ with the same conditional $\theta(\cdot)$. Many results detailed in Chapters 5 and 6 are not specific to the case of the (conditional) Kendall's tau, but can be generalized to any regular (conditional) functional. Such generalizations are studied in Chapter 8, with corresponding theoretical results.

Our framework will be the following. We observe n i.i.d. replications $(\mathbf{X}_i, \mathbf{Z}_i) \sim (\mathbf{X}, \mathbf{Z}),\ i = 1, \dots, n$. The random variable $\mathbf{X}$ belongs to a measurable space $(\mathcal{X}, \mathcal{A})$ while $\mathbf{Z} \in \mathbb{R}^p$. We will denote the joint law of $(\mathbf{X}, \mathbf{Z})$ by $\mathbb{P}_{\mathbf{X},\mathbf{Z}}$ and the conditional law of $\mathbf{X}|\mathbf{Z}$ by $\mathbb{P}_{\mathbf{X}|\mathbf{Z}}$. A regular conditional functional is defined as a functional of the form

$$\theta(\mathbf{z}_1, \dots, \mathbf{z}_k) = \theta\big(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_1}, \dots, \mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_k}\big) = \int g^*(\mathbf{x}_1, \dots, \mathbf{x}_k) d\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_1}(\mathbf{x}_1) \cdots d\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_k}(\mathbf{x}_k),$$

for a fixed function $g^* : \mathcal{X}^k \to \mathbb{R}$, where $k > 0$. For example, in the case of the conditional Kendall's tau, $\mathcal{X} = \mathbb{R}^2$, $k = 2$, and $g^*(\mathbf{x}_1, \mathbf{x}_2)$ is 1 if the pair $(\mathbf{x}_1, \mathbf{x}_2)$ is concordant and -1 if it is discordant.

Our goal is to estimate the function $(\mathbf{z}_1, \dots \mathbf{z}_k) \mapsto \theta(\mathbf{z}_1, \dots, \mathbf{z}_k)$. The first method consists in using a kernel estimator $\hat{\theta}$ of θ. Asymptotic properties of such an estimator were proved by Stute [132]. Under some regularity assumptions, we prove that, with probability greater than $1 - 2\exp\big(-[n/k]t^2h^{kp}/(C_1 + C_2t)\big) - 2\exp\big(-[n/k]t'^2h^{kp}/(C_6 + C_7t')\big)$, we have $\big|\hat{\theta}(\mathbf{z}_1, \dots, \mathbf{z}_k) - \theta(\mathbf{z}_1, \dots, \mathbf{z}_k)\big| \leq \big(1 + C_3h^{\alpha} + C_4t\big) \times \big(C_5h^{k+\alpha} + t'\big)$, for some explicit constants $C_1, \dots, C_7$.

Then we propose a regression-type model generalizing the parametric model that we proposed for the conditional Kendall's tau in Equation (1.10). It is defined as

$$\forall (\mathbf{z}_1, \dots, \mathbf{z}_k) \in \mathcal{Z}^k,\ \Lambda\big(\theta(\mathbf{z}_1, \dots, \mathbf{z}_k)\big) = \boldsymbol{\psi}(\mathbf{z}_1, \dots, \mathbf{z}_k)^T \beta^*, \tag{1.15}$$

for a given transformation $\Lambda : Range(g^*) \mapsto \mathbb{R}$, a basis $\boldsymbol{\psi}$ of size $r > 0$, and a true unknown parameter β^* to be estimated. This model cannot be directly estimated, as both sides of the equation are unobserved. Nevertheless, given a finite collection of points $\mathbf{z}'_1, \ldots, \mathbf{z}'_{n'} \in \mathcal{Z}^{n'}$ and a collection $\mathfrak{I}_{k,n'}$ of injective functions $\sigma : \{1, \ldots, k\} \to \{1, \ldots, n'\}$, we can use the kernel estimate $\hat{\theta}(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)})$ for each choice of $\sigma \in \mathfrak{I}_{k,n'}$. Then, the estimator $\hat{\beta}$ can be defined in a second step as the minimizer of the following l_1-penalized criteria

$$\hat{\beta}^{(g^*)} := \arg\min_{\beta \in \mathbb{R}^r} \left[\frac{(n'-k)!}{n'!} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \left(\Lambda\Big(\hat{\theta}\big(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\big)\Big) - \boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\big)^T \beta \right)^2 + \lambda |\beta|_1 \right],$$

where λ is a tuning parameter. We prove a nonasymptotic bound on $\hat{\beta}^{(g^*)}$ that holds with high probability. Then we consider two different asymptotic regimes : $n \to \infty$ with a fixed n' and (n, n') jointly tends to the infinity. In both regimes, we show that, under suitable regularity conditions, the estimator $\hat{\beta}^{(g^*)}$ is consistent and we derive its asymptotic law.

1.3.2 About confidence intervals for ratios of means

In Section 1.3.1, we have given non-asymptotic bounds for our kernel estimator, which is defined as a ratio of two sums. In fact, the same techniques can be used to obtain concentration inequalities for ratios of means. These inequalities are of the form, $\left|\overline{X}_n/\overline{Y}_n - \mathbb{E}[X]/\mathbb{E}[Y]\right| \leq \Psi(\alpha)$, with probability at least $1 - \alpha$, where Ψ is a function that depends on the regularity assumptions made on the distribution of X and Y, $\overline{X}_n := n^{-1} \sum_{i=1}^n X_{i,n}$, and we observe n i.i.d. replications $(X_{i,n}, Y_{i,n}) \sim (X, Y)$.

In Chapter 9, we prove two versions of these concentrations inequalities, when X and Y both admits second moments, or when the supports of X and Y are bounded. We give explicit expressions for the corresponding functions Ψ, which are of the form $(C + o(1))/\sqrt{n\alpha}$ (respectively $(C + o(1))\sqrt{\ln(1/\alpha)/n}$). This allows us to construct confidence intervals for ratios of means, that are valid for any fixed $n \in \mathcal{N}$.

The most popular method for constructing confidence intervals is based on an application of the delta method. In practice, statisticians tend to examine thinner and thinner effects as more and more data are available. This forms a sequence-of-models framework under which we prove a generalization of the delta method for ratios of means. Finally, we prove a lower bound and an impossibility result related to the nonasymptotic inference of ratios of means. More precisely, we show that there exists a minimum size for any nonasymptotic confidence interval and a minimum level under which no "reasonable test" can be defined.

Publications List

Journal publications

1. Derumigny, A., & Fermanian, J. D., About tests of the “simplifying” assumption for conditional copulas. *Dependence Modeling*, 5(1), 154-197, 2017.

2. Derumigny A., Improved bounds for Square-root Lasso and Square-root Slope. *Electronic Journal of Statistics*, 12(1), 741–766, 2018.

3. Derumigny, A., & Fermanian, J. D., A classification point-of-view about conditional Kendall’s tau. *Computational Statistics & Data Analysis*, 135, 70-94, 2019.

Preprints

1. Derumigny, A., & Fermanian, J. D., About Kendall’s regression. *ArXiv preprint*, arXiv:1802.07613, 2018.

2. Derumigny, A., & Fermanian, J. D., About kernel-based estimation of conditional Kendall’s tau: finite-distance bounds and asymptotic behavior. *ArXiv preprint*, arXiv:1810.06234, 2018.

3. Derumigny, A., Estimation of a regular conditional functional by conditional U-statistics regression. *Arxiv preprint*, arXiv:1903.10914, 2019.

4. Derumigny, A., Girard, L., & Guyonvarch Y., On the construction of confidence intervals for ratios of expectations. *Arxiv preprint*, arXiv:1904.07111, 2019.

Work in progress

1. Derumigny, A., Robust-to-outliers simultaneous inference and noise level estimation using a MOM approach, *in progress*, 2019.

Part I

Linear regression

Chapter 2

Improved bounds for Square-root Lasso and Square-root Slope

Abstract

Extending the results of Bellec, Lecué and Tsybakov [12] to the setting of sparse high-dimensional linear regression with unknown variance, we show that two estimators, the Square-Root Lasso and the Square-Root Slope can achieve the optimal minimax prediction rate, which is $(s/n)\log(p/s)$, up to some constant, under some mild conditions on the design matrix. Here, n is the sample size, p is the dimension and s is the sparsity parameter. We also prove optimality for the estimation error in the l_q-norm, with $q \in [1,2]$ for the Square-Root Lasso, and in the l_2 and sorted l_1 norms for the Square-Root Slope. Both estimators are adaptive to the unknown variance of the noise. The Square-Root Slope is also adaptive to the sparsity s of the true parameter. Next, we prove that any estimator depending on s which attains the minimax rate admits an adaptive to s version still attaining the same rate. We apply this result to the Square-root Lasso. Moreover, for both estimators, we obtain valid rates for a wide range of confidence levels, and improved concentration properties as in [12] where the case of known variance is treated. Our results are non-asymptotic.

Keywords: Sparse linear regression, minimax rates, high-dimensional statistics, adaptivity, square-root estimators.

Based on [35]: Derumigny A, Improved bounds for Square-root Lasso and Square-root Slope. *Electronic Journal of Statistics*, 12(1) :741–766, 2018.

2.1 Introduction

In a recent paper by Bellec, Lecué and Tsybakov [12], it is shown that there exist high-dimensional statistical methods realizable in polynomial time that achieve the minimax optimal rate $(s/n)\log(p/s)$ in the context of sparse linear regression. Here, n is the sample size, p is the dimension and s is the sparsity parameter. The result is achieved by the Lasso and Slope estimators, and the Slope estimator is adaptive to the unknown sparsity s. Bounds for more general estimators are proved by Bellec, Lecué and Tsybakov [13, 11]. These articles also establish bounds in deviation that hold for any confidence

level and for the risk in expectation. However, the estimators considered in [12, 13, 11] require the knowledge of the noise variance σ^2. To our knowledge, no polynomial-time methods, which would be at the same time optimal in a minimax sense and adaptive both to σ and s are available in the literature.

Estimators similar to the Lasso, but adaptive to σ are the Square-Root Lasso and the related Scaled Lasso, introduced by Sun and Zhang [134] and Belloni, Chernozhukov and Wang [14]. It has been shown to achieve the rate $(s/n)\log(p)$ in deviation with the value of the tuning parameter depending on the confidence level. A variant of this estimator is the Heteroscedastic Square-Root Lasso, which is studied in more general nonparametric and semiparametric setups by Belloni, Chernozhukov and Wang [15], but it also achieves the rate $(s/n)\log(p)$ and depends on the confidence level. We refer to the book by Giraud [67] for the link between the Lasso and the Square-Root Lasso and a short proof of oracle inequalities for the Square-root Lasso. In summary, there are two points to improve for the Square-root Lasso method:

(i) The available results on oracle inequalities are valid only for the estimators depending on the confidence level. Thus, one cannot have an oracle inequality for one given estimator at any confidence level except the one that was used to design it.

(ii) The obtained rate is $(s/n)\log(p)$ which is greater than the minimax rate $(s/n)\log(p/s)$.

The Slope, which is an acronym for Sorted L-One Penalized Estimation, is an estimator introduced by Bogdan et al. [20], that is close to the Lasso, but uses the sorted l_1 norm instead of the standard l_1 norm for penalization. Su and Candès [133] proved that, as opposed to the Lasso, the Slope estimator is asymptotically minimax, in the sense that it attains the rate $(s/n)\log(p/s)$ for two isotropic designs, that is either for $\mathbb{X}$ deterministic with $\frac{1}{n}\mathbb{X}^T\mathbb{X} = I_{p\times p}$ or when $\mathbb{X}$ is a matrix with i.i.d. standard normal entries. Moreover, their result has not only the optimal minimax rate, but also the exact optimal constant. General isotropic random designs are explored by Lecué and Mendelson [93]. For non-isotropic random designs and deterministic designs under conditions close to the Restricted Eigenvalue, the behavior of the Slope estimator is studied in [12]. The Slope estimator is adaptive only to s, and requires knowledge of σ, which is not available in practice. In order to have an estimator which is adaptive both to s and σ, we will use the Square-Root Slope, introduced by Stucky and van de Geer [131]. They give oracle inequalities for a large group of square-root estimators, including the new Square-Root Slope, but still following the scheme where (i) and (ii) cannot be avoided. The square-root estimators are also members of a more general family of penalized estimators defined by Owen [109, equations (8)-(9)] ; using their notation, these estimators correspond to the case where $\mathcal{H}_M$ is the squared loss and $\mathcal{B}_M$ is a norm (either the l_1 norm or the slope norm).

The paper is organized as follows. In Section 2.2, we provide the main definitions and notation. In Section 2.3, we show that the Square-Root Lasso is minimax optimal if s is known while being adaptive to σ under a mild condition on the design matrix (SRE). In Section 2.4, we show that any sequence of estimators can be made adaptive to the sparsity parameter s, while keeping the same rate up to some constant, with a computational cost increased by a factor of $\log(s_*)$ where s_* is an upper bound on the sparsity parameter s. As an application, the Square-root Lasso modified by this procedure is still optimal while being now adaptive to s (in addition of being already adaptive to σ). In Section 2.5, we show how to adapt any algorithm for computing the Slope estimator to the case of the Square-root Slope estimator. In Section 2.6, we study the Square-Root Slope estimator, and show that it is minimax optimal and adaptive both to s and σ, under a slightly stronger condition (WRE). The (SRE) and (WRE) conditions have already been studied by Bellec, Lecué and Tsybakov [12] and hold with high probability for a large

class of random matrices. Moreover, the inequalities we obtain for each estimator are valid for a wide range of confidence levels. Proofs are given in Section 2.7.

2.2 The framework

We use the notation $|\cdot|_q$ for the l_q norm, with $1 \leq q \leq \infty$, and $|\cdot|_0$ for the number of non-zero coordinates of a given vector. For any $v \in \mathbb{R}^p$, and any set of coordinates J, we denote by v_J the vector $(v_j \mathbb{1}\{i \in J\})_{i=1,\dots,p}$, where $\mathbb{1}$ is the indicator function. We also define the empirical norm of a vector $u = (u_1, \dots, u_n)$ as $||u||_n^2 := \frac{1}{n}\sum_{i=1}^n u_i^2$. For a vector $v \in \mathbb{R}^p$, we denote by $v_{(j)}$ the j-th largest component of v. As a particular case, $|v|_{(j)}$ is the j-th largest component of the vector $|v|$ whose components are the absolute values of the components of v. We use the notation $\langle \cdot, \cdot \rangle$ for the inner product with respect to the Euclidean norm and $(e_j)_{j=1,\dots,p}$ for the canonical basis in $\mathbb{R}^p$.

Let $Y \in \mathbb{R}^n$ be the vector of observations and let $\mathbb{X} \in \mathbb{R}^{n \times p}$ be the design matrix. We assume that the true model is the following

$$Y = \mathbb{X}\beta^* + \varepsilon. \tag{2.1}$$

Here $\beta^* \in \mathbb{R}^p$ is the unknown true parameter. We assume that ε is the random noise, with values in $\mathbb{R}^n$, distributed as $\mathcal{N}(0, \sigma^2 I_{n\times n})$, where $I_{n\times n}$ is the identity matrix. We denote by $\mathbb{P}_{\beta^*}$ the probability distribution of Y satisfying (2.1). In what follows, we define the set $B_0(s) := \{\beta^* \in \mathbb{R}^p : |\beta^*|_0 \leq s\}$. In the high-dimensional framework, we have typically in mind the case where s is small, p is large and possibly $p \gg n$.

We define two square-root type estimators of β^*: the Square-Root Lasso $\hat{\beta}^{SQL}$ and the Square-Root Slope $\hat{\beta}^{SQS}$ by the following relations

$$\hat{\beta}^{SQL} \in \arg\min_{\beta \in \mathbb{R}^p} \left(\frac{1}{\sqrt{n}} |Y - \mathbb{X}\beta|_2 + \lambda |\beta|_1 \right), \tag{2.2}$$

$$\hat{\beta}^{SQS} \in \arg\min_{\beta \in \mathbb{R}^p} \left(\frac{1}{\sqrt{n}} |Y - \mathbb{X}\beta|_2 + |\beta|_* \right), \tag{2.3}$$

where $\lambda > 0$ is a tuning parameter to be chosen, and the sorted l_1 norm, $|\cdot|_*$, is defined for all $u \in \mathbb{R}^p$ by $|u|_* = \sum_{j=1}^p \lambda_j |u|_{(j)}$, with tuning parameters $\lambda_1 \geq \cdots \geq \lambda_p > 0$.

2.3 Optimal rates for the Square-Root Lasso

In this section, we derive oracle inequalities with optimal rate for the Square-Root Lasso estimator. We will use the *Strong Restricted Eigenvalue* (SRE) condition, introduced in [12]. For $c_0 > 0$ and $s \in \{1, \dots, p\}$, it is defined as follows,

$SRE(s, c_0)$ **condition :** *The design matrix* $\mathbb{X}$ *satisfies*

$$\max_{j=1,\dots,p} ||\mathbb{X}e_j||_n \leq 1$$

and

$$\kappa(s) := \min_{\delta \in C_{SRE}(s,c_0): \delta \neq 0} \frac{||\mathbb{X}\delta||_n}{|\delta|_2} > 0, \tag{2.4}$$

where $C_{SRE}(s, c_0) := \{\delta \in \mathbb{R}^p : |\delta|_1 \leq (1 + c_0)\sqrt{s}|\delta|_2\}$ *is a cone in* $\mathbb{R}^p$.

The condition $\max_{j=1,\dots,p} ||\mathbb{X}e_j||_n \leq 1$ is standard and corresponds to a normalization. It is shown in [12, Proposition 8.1] that the SRE condition is equivalent to the Restricted Eigenvalue (RE) condition of [19] if that is considered in conjunction with such a normalization. By the same proposition, the RE condition is also equivalent to the s-sparse eigenvalue condition, which is satisfied with high probability for a large class of random matrices. It is the case, if for instance, $n \geq Cs\log(ep/s)$ and the rows of $\mathbb{X}$ satisfies the small ball condition, which is very mild, see, e.g. [12].

Note that the minimum in (2.4) is the same as the minimum of the function $\delta \mapsto ||\mathbb{X}\delta||_n$ on the set $C_{SRE}(s, c_0) \cap \{\delta \in \mathbb{R}^p : |\delta|_2 = 1\}$, which is a continuous function on a compact of $\mathbb{R}^p$, therefore this minimum is attained. When there is no ambiguity over the choice of s, we will just write κ instead of $\kappa(s)$.

Theorem 2.1. *Let $s \in \{1, \dots, p\}$ and assume that the $SRE(s, 5/3)$ condition holds. Choose the following tuning parameter*

$$\lambda = \gamma\sqrt{\frac{1}{n}\log\left(\frac{2p}{s}\right)}, \tag{2.5}$$

and assume that

$$\gamma \geq 16 + 4\sqrt{2} \quad \textit{and} \quad \frac{s}{n}\log\left(\frac{2p}{s}\right) \leq \frac{9\kappa^2}{256\gamma^2}. \tag{2.6}$$

Then, for every $\delta_0 \geq \exp(-n/4\gamma^2)$ and every $\beta^* \in \mathbb{R}^p$ such that $|\beta^*|_0 \leq s$, with $\mathbb{P}_{\beta^*}$-probability at least $1 - \delta_0 - (1 + e^2)e^{-n/24}$, we have

$$||\mathbb{X}(\hat{\beta}^{SQL} - \beta^*)||_n \leq \sigma \max\left(\frac{C_1}{\kappa^2}\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)},\ C_2\sqrt{\frac{\log(1/\delta_0)}{n}}\right), \tag{2.7}$$

$$|\hat{\beta}^{SQL} - \beta^*|_q \leq \sigma \max\left(\frac{C_3}{\kappa^2}s^{1/q}\sqrt{\frac{1}{n}\log\left(\frac{2p}{s}\right)},\ C_4 s^{1/q-1}\sqrt{\frac{\log^2(1/\delta_0)}{n\log(2p/s)}}\right), \tag{2.8}$$

where $1 \leq q \leq 2$, and $C_1 > 0$, $C_2 > 0$, $C_3 > 0$, $C_4 > 0$ are constants depending only on γ.

The values of the constants C_1, C_2, C_3 and C_4 in Theorem 2.1 can be found in the proof, in Section 2.7.2. Using the fact that $\kappa \leq 1$ and choosing $\delta_0 = (s/p)^s$, we get the following corollary of Theorem 2.1.

Corollary 2.2. *Under the assumptions of Theorem 2.1, with $\mathbb{P}_{\beta^*}$-probability at least $1 - (s/p)^s - (1 + e^2)e^{-n/24}$, we have*

$$||\mathbb{X}(\hat{\beta}^{SQL} - \beta^*)||_n \leq \frac{C_2}{\kappa^2}\sigma\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)},$$

$$|\hat{\beta}^{SQL} - \beta^*|_q \leq \frac{C_4}{\kappa^2}\sigma s^{1/q}\sqrt{\frac{1}{n}\log\left(\frac{2p}{s}\right)},$$

where $1 \leq q \leq 2$.

Theorem 2.1 and Corollary 2.2 give bounds that hold with high probability for both the prediction error and the estimation error in the l_q norm, for every q in $[1, 2]$. Note that the bounds are best when the tuning parameter is chosen as small as possible, i.e. with $\gamma = 16 + 4\sqrt{2}$. As shown in Section 7 of Bellec, Lecué and Tsybakov [12], the rates of estimation obtained in the latter corollary are optimal in a minimax sense on the set $B_0(s) := \{\beta^* \in \mathbb{R}^p : |\beta^*|_0 \leq s\}$. We obtain the same rate of convergence as [12] (see the paragraph after Corollary 4.3 in [12]) up to some multiplicative constant.

The rate is also the same as in Su and Candès [133], but the framework is quite different: we obtain a non-asymptotic bound in probability whereas they consider asymptotic bounds in expectation

(cf. Theorem 1.1 in [133]) and in probability (Theorem 1.2) but without giving an explicit expression of the probability that their bound is valid. Our result is non-asymptotic and valid when general enough conditions on $\mathbb{X}$ are satisfied whereas the result in [133] is asymptotic as $n \to \infty$, and valid for two isotropic designs, that is either for $\mathbb{X}$ deterministic with $\frac{1}{n}\mathbb{X}^T\mathbb{X} = I_{p\times p}$ or when $\mathbb{X}$ is a matrix with i.i.d. standard normal entries.

Similarly to [12], for each tuning parameter γ, there is a wide range of levels of confidence δ_0 under which the bounds of Theorem 2.1 are valid. However, [12] allows for an arbitrary small confidence level while in our case, there is a lower bound on the size of the confidence level under which the rate is obtained. Note that this bound can be made arbitrary small by choosing a sample size n large enough.

Note that the possible values chosen for the tuning parameter λ are independent of the underlying standard deviation σ, which is unknown in practice. This gives an advantage for the Square-Root Lasso over other methods such as the ordinary Lasso. Nevertheless, this estimator is not adaptive to the sparsity s, so that we need to know that $|\beta^*|_0 \leq s$ in order to be able to apply this result. In the following section, we suggest a procedure to make the Square-root Lasso adaptive to s while keeping its optimality and adaptivity to σ.

2.4 Adaptation to sparsity by a Lepski-type procedure

Let s_* be an integer in $\{2, \dots, p/e\}$. We want to show that the Square-Root Lasso can also achieve the minimax optimal bound, adaptively to the sparsity s on the interval $[1, s_*]$ (in addition of being already adaptive to σ). Following [12], we will use aggregation of at most $\log_2(s_*)$ Square-Root Lasso estimators with different tuning parameters to construct an adaptive estimator $\tilde{\beta}$ of β and at the same time an estimator $\tilde{s}$ of the sparsity s.

In the following, we use the notation

$$\kappa_* := \kappa(2s_*).$$

Note that $\kappa_* = \min_{s=1,\dots,2s_*} \kappa(s)$. Indeed, the function $\kappa(\cdot)$ is decreasing, because the minimization (2.4) is done on spaces that are growing with s, in the sense of the inclusion. We will assume that the condition $SRE(2s_*, 5/3)$ holds and that $(2s_*/n)\log\left(2p/(2s_*)\right) \leq 9\kappa_*^2/(256\gamma^2)$. The functions $b \mapsto (b/n)\log(2p/b)$ and $\kappa(\cdot)$ are respectively increasing (by Lemma 2.4) and decreasing, so this ensures that the second part of condition (2.6) is satisfied for any $s = 1, \dots, 2s_*$.

We can reformulate Corollary 2.2 as follows: for any $s = 1, \dots, 2s_*$ and any $\gamma \geq 16 + 4\sqrt{2}$

$$\sup_{\beta^* \in B_0(s)} \mathbb{P}_{\beta^*}\left(||\mathbb{X}(\hat{\beta}^{SQL}_{(s,\gamma)} - \beta^*)||_n \leq \frac{C_2(\gamma)}{\kappa_*^2}\sigma\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)} \right) \geq 1 - \left(\frac{s}{p}\right)^s - (1+e^2)e^{-n/24}, \tag{2.9}$$

denoting by $\hat{\beta}^{SQL}_{(s,\gamma)}$ the estimator (2.2) with the tuning parameter $\lambda_{(s,\gamma)}$ given by (2.5). Replacing s by $2s$ in equation (2.9), we get that for any $s = 1, \dots, s_*$ and any $\gamma \geq 16 + 4\sqrt{2}$,

$$\sup_{\beta^* \in B_0(2s)} \mathbb{P}_{\beta^*}\left(||\mathbb{X}(\hat{\beta}^{SQL}_{(2s,\gamma)} - \beta^*)||_n \leq \frac{C_2(\gamma)}{\kappa_*^2}\sigma\sqrt{\frac{2s}{n}\log\left(\frac{p}{2s}\right)} \right) \geq 1 - \left(\frac{2s}{p}\right)^{2s} - (1+e^2)e^{-n/24}. \tag{2.10}$$

Remark that $\lambda_{(s,\gamma)} = \gamma\sqrt{\frac{1}{n}\log\left(\frac{2p}{s}\right)} = \tilde{\gamma}\sqrt{\frac{1}{n}\log\left(\frac{2p}{s}\right) - \frac{\log(2)}{n}} = \lambda_{(2s,\tilde{\gamma})}$ for some $\tilde{\gamma} > \gamma$. As a consequence, $\hat{\beta}^{SQL}_{(s,\gamma)} = \hat{\beta}^{SQL}_{(2s,\tilde{\gamma})}$ and we can apply Equation (2.10), replacing γ by $\tilde{\gamma}$ and we get

$$\sup_{\beta^* \in B_0(2s)} \mathbb{P}_{\beta^*}\left(||\mathbb{X}(\hat{\beta}^{SQL}_{(s,\gamma)} - \beta^*)||_n \leq \frac{C_2(\tilde{\gamma})}{\kappa_*^2}\sigma\sqrt{\frac{2s}{n}\log\left(\frac{p}{2s}\right)}\right) \geq 1 - \left(\frac{2s}{p}\right)^{2s} - (1+e^2)e^{-n/24}. \tag{2.11}$$

Note that equations (2.9) and (2.11) are the same as equations (5.2) and (5.4) in Bellec, Lecué and Tsybakov [12], taking $C_0 := \max\big(C_2(\gamma), C_2(\tilde{\gamma})\big)/\kappa_*^2$, except that we have a supplementary term $-(1+e^2)e^{-n/24}$. Similarly, we deduce from Corollary 2.2 that

$$\sup_{\beta^* \in B_0(s)} \mathbb{P}_{\beta^*}\left(|\mathbb{X}(\hat{\beta}^{SQL}_{(s,\gamma)} - \beta^*)|_q \leq \frac{C_4(\gamma)}{\kappa_*^2}\sigma s^{1/q}\sqrt{\frac{s}{n}\log\left(\frac{2p}{s}\right)}\right) \geq 1 - \left(\frac{s}{p}\right)^{s} - (1+e^2)e^{-n/24}, \tag{2.12}$$

$$\sup_{\beta^* \in B_0(2s)} \mathbb{P}_{\beta^*}\left(|\hat{\beta}^{SQL}_{(s,\gamma)} - \beta^*|_q \leq \frac{C_4(\tilde{\gamma})}{\kappa_*^2}\sigma s^{1/q}\sqrt{\frac{2s}{n}\log\left(\frac{2p}{2s}\right)}\right) \geq 1 - \left(\frac{2s}{p}\right)^{2s} - (1+e^2)e^{-n/24}. \tag{2.13}$$

We describe now an algorithm to compute this adaptive estimator. The idea is to use an estimator $\tilde{s}$ of s which can be written as $\tilde{s} := 2^{\tilde{m}}$ for some positive data-dependent integer $\tilde{m}$. We will use the notation $M := \max\{m \in \mathbb{N} : 2^m \leq s_*\}$, so that the number of estimators we consider in the aggregation is M.

The suggested procedure is detailed in Algorithm 2 below, with the distance $d(\beta,\beta') = ||\mathbb{X}(\beta - \beta')||_n$ or $d(\beta,\beta') = |\beta - \beta'|_q$ for $q \in [1,2]$. It can be used for any family of estimators $(\hat{\beta}_{(s)})_{s=1,\dots s_*}$, and chooses the best one in terms of the distance $d(\cdot,\cdot)$, resulting in an aggregated estimator $\tilde{\beta}$. Note that the weight function $w(\cdot)$ used in the algorithm cannot depend on σ as in [12], i.e. to have the form $w(b) = C_0\sigma\sqrt{(b/n)\log(p/b)}$ (respectively $w(b) = C_0\sigma b^{1/q}\sqrt{(1/n)\log(p/b)}$), because we are looking for a procedure adaptive to σ. Therefore, we will remove σ from w and use an estimate $\hat{\sigma}$.

Algorithm 2: Algorithm for adaptivity.

Input: a distance $d(\cdot,\cdot)$ on $\mathbb{R}^p$

Input: a function $w(\cdot) : [1, s_*] \to \mathbb{R}_+$ satisfying Assumption 2.4.1

Input: a family of estimators $\big(\hat{\beta}_{(s)}\big)_{s=1,\dots,s_*}$

$M \leftarrow \lfloor \log_2(s_*) \rfloor$;

for $m \leftarrow 1$ **to** $M+1$ **do**

  compute the estimator $\hat{\beta}_{(2^m)}$;

end

compute $\hat{\sigma} \leftarrow ||Y - \mathbb{X}\hat{\beta}_{(2^{M+1})}||_n$;

compute the set $S_1 \leftarrow \Big\{m \in \{1,\dots,M\} : d\Big(\hat{\beta}_{(2^{k-1})}, \hat{\beta}_{(2^k)}\Big) \leq 4\hat{\sigma}C_0 w(2^k), \text{ for all } k \geq m\Big\}$;

if $S_1 \neq \emptyset$ **then** $\tilde{m} \leftarrow \min S_1$ **else** $\tilde{m} \leftarrow M$;

Output: $\tilde{s} \leftarrow 2^{\tilde{m}}$

Output: $\tilde{\beta} \leftarrow \hat{\beta}_{(\tilde{s})}$

Assumption 2.4.1. *The function $w(\cdot) : [1, s_*] \to \mathbb{R}_+$ satisfies the following conditions:*

1. *$w(\cdot)$ is increasing on $[1, s_*]$;*
2. *There exists a constant $C' > 0$ such that, for all $m = 1, \dots, M$, we have*

$$\sum_{k=1}^{m} w(2^k) \leq C' \cdot w(2^m) \ ;$$

3. *There exists a constant $C'' > 0$ such that, for all $b = 1, \dots, s_*$,*

$$w(2b) \leq C'' w(b).$$

Assumption 2.4.2. *The family of estimators $(\hat{\beta}_{(s)})_{s=1,\dots,s_*}$ satisfies*

$$\sup_{\beta^* \in B_0(2s)} \mathbb{P}_{\beta^*} \Big(\sigma/2 \leq \hat{\sigma} \leq \alpha\sigma \Big) \leq u_{n,p,M},$$

with a constant $\alpha > 0$, $\hat{\sigma} := ||Y - \mathbb{X}\hat{\beta}_{(2^{M+1})}||_n$, and $u_{n,p,M} > 0$.

Theorem 2.3. *Let $s_* \in \{2, \dots, p/e\}$ and let $(\hat{\beta}_{(s)})_{s=1,\dots,s_*}$ be a collection of estimators satisfying Assumption 2.4.2 such that, for any $s = 1, \dots, s_*$,*

$$\sup_{\beta^* \in B_0(s)} \mathbb{P}_{\beta^*} \left(d(\hat{\beta}_{(s)}\,,\, \beta^*) \leq C_0 \sigma w(s) \right) \geq 1 - \left(\frac{s}{p}\right)^s - u_n, \tag{2.14}$$

and

$$\sup_{\beta^* \in B_0(2s)} \mathbb{P}_{\beta^*} \left(d(\hat{\beta}_{(s)}\,,\, \beta^*) \leq C_0 \sigma w(2s) \right) \geq 1 - \left(\frac{2s}{p}\right)^{2s} - u_n, \tag{2.15}$$

for a constant $C_0 > 0$, a function $w(\cdot) : [1, s_*] \to \mathbb{R}_+$ satisfying Assumption 2.4.1, and $u_n > 0$.

Then, there exists a constant C_5, depending on $C_0, C', C'', C_2, \kappa$ and α such that, for all $\beta^* \in B_0(s)$, the aggregated estimator $\tilde{\beta}$ satisfies:

$$\begin{aligned} &\mathbb{P}_{\beta^*} \left(d(\tilde{\beta}, \beta^*) \leq C_5 \cdot \sigma w(s) \right) \\ &\qquad \geq 1 \ - \ 3(\log_2(s_*) + 1)^2 \left(\left(\frac{2s}{p}\right)^{2s} + u_n \right) - u_{n,p,M}. \end{aligned}$$

Furthermore,

$$\mathbb{P}_{\beta^*} \big(\tilde{s} \leq s \big) \geq 1 - 2(\log_2(s_*) + 1)^2 \left(\left(\frac{2s}{p}\right)^{2s} + u_n \right) - u_{n,p,M}.$$

This theorem is proved in Section 2.7.3.1. In particular, it implies that when $\hat{\beta}_{(s)} = \hat{\beta}^{SQL}_{(s,\gamma)}$, the aggregated estimator $\tilde{\beta}$ has the same rate on $B_0(s)$ as the estimators with known s. We detail it below. The following lemmas proved in Sections 2.7.3.2 and 2.7.3.3 assure that Theorem 2.3 can be applied to the family $\hat{\beta}_{(s)} = \hat{\beta}^{SQL}_{(s,\gamma)}$.

Lemma 2.4. *Assumption 2.4.1 is satisfied with the choices*

$$w(b) = \sqrt{(b/n)\log(p/b)} \text{ and } w(b) = b^{1/q}\sqrt{(1/n)\log(2p/b)}, \text{ for } q \in [1, 2].$$

Lemma 2.5. *Assume that the $SRE(2s_*, 5/3)$ condition holds and*

$$\gamma \geq 16 + 4\sqrt{2} \text{ and } \frac{2s_*}{n}\log\left(\frac{p}{s_*}\right) \leq \min\left(\frac{9\kappa_*^2}{256\gamma^2}, \frac{\kappa_*^4}{2C_2(\gamma)^2}\left(\frac{1}{\sqrt{2}} - \frac{1}{2}\right)^2\right),$$

where $\kappa_* := \kappa(2s_*)$. Then Assumption 2.4.2 is satisfied with the choice

$$(\hat{\beta}_{(s)})_{s=1,\dots,s_*} = (\hat{\beta}^{SQL}_{(s,\gamma)})_{s=1,\dots,s_*}, \quad \alpha = 2 + \frac{3\sqrt{2}C_2(\gamma)}{16\kappa\gamma},$$
$$\text{and } u_{n,p,M} = (2^{M+1}/p)^{2^{M+1}} - (1+e^2)e^{-n/24}.$$

Combining Equations (2.9), (2.11) with Theorem 2.3 and Lemmas 2.4 and 2.5, we obtain the following results for the case of the Square-root Lasso.

Corollary 2.6. *Under the same assumptions as in Lemma 2.5, using Algorithm 2, with $(\hat{\beta}_{(s)})_{s=1,\dots,s_*} = (\hat{\beta}^{SQL}_{(s,\gamma)})_{s=1,\dots,s_*}$, the distance $d(\beta,\beta') = ||\mathbb{X}(\beta-\beta')||_n$, and the weight $w(b) = \sqrt{(b/n)\log(p/b)}$, we have that, for all $\beta^* \in B_0(s)$, the aggregated estimator $\tilde{\beta}$ satisfies*

$$\mathbb{P}_{\beta^*}\left(||\mathbb{X}(\tilde{\beta}-\beta^*)||_n \leq C_5 \cdot \sigma\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)}\right)$$
$$\geq 1 - 3(\log_2(s_*)+1)^2\left(\left(\frac{2s}{p}\right)^{2s} + u_n\right) - u_{n,p,M},$$

and

$$\mathbb{P}_{\beta^*}\left(\tilde{s} \leq s\right) \geq 1 - 2(\log_2(s_*)+1)^2\left(\left(\frac{2s}{p}\right)^{2s} + u_n\right) - u_{n,p,M},$$

where $u_n = (1+e^2)e^{-n/24}$, $u_{n,p,M} = (2^{M+1}/p)^{2^{M+1}} - (1+e^2)e^{-n/24}$, and C_5 is a constant depending only on γ and κ_*.

Corollary 2.7. *Under the same assumptions as in Lemma 2.5, using Algorithm 2, with $(\hat{\beta}_{(s)})_{s=1,\dots,s_*} = (\hat{\beta}^{SQL}_{(s,\gamma)})_{s=1,\dots,s_*}$, the distance $d(\beta,\beta') = |\beta-\beta'|_q$, and the weight $w(b) = b^{1/q}\sqrt{(1/n)\log(2p/b)}$, for $q \in [1;2]$, we have that, for all $\beta^* \in B_0(s)$, the aggregated estimator $\tilde{\beta}$ satisfies*

$$\mathbb{P}_{\beta^*}\left(|\tilde{\beta}-\beta^*|_q \leq C_5 \cdot \sigma s^{1/q}\sqrt{\frac{1}{n}\log\left(\frac{p}{s}\right)}\right)$$
$$\geq 1 - 3(\log_2(s_*)+1)^2\left(\left(\frac{2s}{p}\right)^{2s} + u_n\right) - u_{n,p,M},$$

and

$$\mathbb{P}_{\beta^*}\left(\tilde{s} \leq s\right) \geq 1 - 2(\log_2(s_*)+1)^2\left(\left(\frac{2s}{p}\right)^{2s} + u_n\right) - u_{n,p,M},$$

where $u_n = (1+e^2)e^{-n/24}$, $u_{n,p,M} = (2^{M+1}/p)^{2^{M+1}} - (1+e^2)e^{-n/24}$, and C_5 is a constant depending only on γ and κ_*.

Thus, we have shown that the suggested aggregated procedure based on the Square-root Lasso is adaptive to s while still being adaptive to σ and minimax optimal. Note that the computational cost is multiplied by $O(\log(s_*))$.

2.5 Algorithms for computing the Square-root Slope

In this part, our goal is to provide algorithms for computing the square-root Slope estimator. A natural idea is revisiting the algorithms used for the square-root Lasso and for the Slope, then adapting or combining them.

Belloni, Chernozhukov and Wang [14, Section 4] have proposed to compute the Square-root Lasso estimator by reducing its definition to an equivalent problem, which can be solved by interior-point or first-order methods. The equivalent formulation as the Scaled Lasso, introduced by Sun and Zhang [134] allows one to view it as a joint minimization in (β, σ). Sun and Zhang [134] propose an iterative algorithm which alternates estimation of β using the ordinary Lasso and estimation of σ.

Zeng and Figueiredo [144] studied several algorithms related to estimation of the regression with the ordered weighted l_1-norm, which is the Slope penalization. Bogdan et al. [20] provide an algorithm for computing the Slope estimator using a proximal gradient.

As in the case of the Square-root Lasso, we still have for any β,

$$||Y - \mathbb{X}\beta||_n = \min_{\sigma>0} \left(\sigma + \frac{||Y - \mathbb{X}\beta||_n^2}{\sigma} \right), \tag{2.16}$$

where the minimum is attained for $\hat{\sigma} = ||Y - \mathbb{X}\beta||_n$. As a consequence,

$$\hat{\beta}^{SQS} \in \arg\min_{\beta \in \mathbb{R}^p} \left(||Y - \mathbb{X}\beta||_n + |\beta|_* \right)$$

is equivalent to take the estimator $\hat{\beta}$ in the joint minimization program

$$(\hat{\beta}, \hat{\sigma}) \in \underset{\beta \in \mathbb{R}^p,\, \sigma>0}{\arg\min} \left(\sigma + \frac{||Y - \mathbb{X}\beta||_n^2}{\sigma} + |\beta|_* \right).$$

Alternating minimization in β and in σ gives an iterative procedure for a "Scaled Slope" (see Algorithm 3).

Algorithm 3: Scaled Slope algorithm

Input: explained variable Y, design matrix $\mathbb{X}$;
Input: tuning parameters $\lambda_1 \leq \cdots \leq \lambda_p$;
choose some initialization value for $\hat{\sigma}$, for example the standard deviation of Y ;
repeat
  estimate $\hat{\beta}$ by the Slope algorithm with the parameters $\hat{\sigma} \cdot \lambda_1, \ldots, \hat{\sigma} \cdot \lambda_p$;
  estimate $\hat{\sigma}$ by $||Y - \mathbb{X}\hat{\beta}||_n$;
until *convergence*;
Output: a joint estimator $\left(\hat{\beta}, \hat{\sigma}\right)$;

2.6 Optimal rates for the Square-Root Slope

In this part, we will use another condition, the *Weighted Restricted Eigenvalue* condition, introduced in [12]. For $c_0 > 0$ and $s \in \{1, \ldots, p\}$, it is defined as follows,

$WRE(s, c_0)$ **condition :** *The design matrix* $\mathbb{X}$ *satisfies*

$$\max_{j=1,\ldots,p} ||\mathbb{X}e_j||_n \leq 1$$

and

$$\kappa' := \min_{\delta \in C_{WRE}(s,c_0):\delta \neq 0} \frac{||\mathbb{X}\delta||_n}{|\delta|_2} > 0, \tag{2.17}$$

where

$$C_{WRE}(s,c_0) := \left\{ \delta \in \mathbb{R}^p : |\delta|_* \leq (1+c_0)|\delta|_2 \sqrt{\sum_{j=1}^{s} \lambda_j^2} \right\}$$

is a cone in $\mathbb{R}^p$.

To obtain the following result, we assume that the Weighted Restricted Eigenvalue condition holds. This condition is shown to be only slightly more constraining than the usual Restricted Eigenvalue condition of [19], but is nevertheless satisfied with high probability for a large class of random matrices, see Bellec, Lecué and Tsybakov [12] for a discussion. Note that, in a similar way as in definition (2.4), the minimum is attained. Indeed, κ' is equal to the minimum of the function $\delta \mapsto ||\mathbb{X}\delta||_n$ on the set $C_{WRE}(s,c_0) \cap \{\delta \in \mathbb{R}^p : |\delta|_2 = 1\}$, which is a continuous function on a compact of $\mathbb{R}^p$.

Theorem 2.8. *Let* $s \in \{1,\dots,p\}$ *and assume that the* $WRE(s,20)$ *condition holds. Choose the following tuning parameters*

$$\lambda_j = \gamma' \sqrt{\frac{\log(2p/j)}{n}}, \text{ for } j = 1,\dots,p, \tag{2.18}$$

and assume that

$$\gamma' \geq 16 + 4\sqrt{2} \quad \text{and} \quad \frac{s}{n} \log\left(\frac{2ep}{s}\right) \leq \frac{\kappa'^2}{256\gamma'^2}. \tag{2.19}$$

Then, for every $\delta_0 \geq \exp(-n/4\gamma'^2)$ *and every* $\beta^* \in \mathbb{R}^p$ *such that* $|\beta^*|_0 \leq s$, *with* $\mathbb{P}_{\beta^*}$*-probability at least* $1 - \delta_0 - (1+e^2)e^{-n/24}$, *we have*

$$||\mathbb{X}(\hat{\beta}^{SQS} - \beta^*)||_n \leq \sigma \max\left(\frac{C_1'}{\kappa'} \sqrt{\frac{s}{n} \log\left(\frac{p}{s}\right)}, C_2' \sqrt{\frac{\log(1/\delta_0)}{n}} \right), \tag{2.20}$$

$$|\hat{\beta}^{SQS} - \beta^*|_* \leq \sigma \max\left(\frac{C_1'}{\kappa'^2} \frac{s}{n} \log\left(\frac{p}{s}\right), C_2' \frac{\log(1/\delta_0)}{n} \right), \tag{2.21}$$

$$|\hat{\beta}^{SQS} - \beta^*|_2 \leq \sigma \max\left(\frac{C_1'}{\kappa'^2} \sqrt{\frac{s}{n} \log\left(\frac{p}{s}\right)}, C_2' \sqrt{\frac{\log^2(1/\delta_0)}{sn\log(p/s)}} \right), \tag{2.22}$$

for constants $C_1' > 0$ *and* $C_2' > 0$ *depending only on* γ'.

The values of the constants C_1' and C_2' can be found in the proof, in Subsection 2.7.4. Note that the bounds are best when the tuning parameters is chosen as small as possible, i.e. using the choice $\gamma' = 16 + 4\sqrt{2}$. Using the fact that $\kappa' \leq 1$ and choosing $\delta_0 = (s/p)^s$, we get the following corollary.

Corollary 2.9. *Under the assumptions of Theorem 2.8, with* $\mathbb{P}_{\beta^*}$*-probability at least* $1 - (s/p)^s - (1 + e^2)e^{-n/24}$, *we have*

$$||\mathbb{X}(\hat{\beta}^{SQS} - \beta^*)||_n \leq \frac{C_1'}{\kappa'} \sigma \sqrt{\frac{s}{n} \log\left(\frac{p}{s}\right)},$$

$$|\hat{\beta}^{SQS} - \beta^*|_* \leq \frac{C_1'}{\kappa'^2} \sigma \frac{s}{n} \log\left(\frac{p}{s}\right),$$

$$|\hat{\beta}^{SQS} - \beta^*|_2 \leq \frac{C_1'}{\kappa'^2} \sigma \sqrt{\frac{s}{n} \log\left(\frac{p}{s}\right)},$$

These results show that the Square-Root Slope estimator, with a given choice of parameters, attains the optimal rate of convergence in the prediction norm $||\cdot||_n$ and in the estimation norm $|\cdot|_2$. We also provide a bound on the sorted l_1 norm $|\cdot|_*$ of the estimation error. One can note that the choice of λ_i that allows us to obtain optimal bounds does not depend on the level of confidence δ_0, but only influence the size of the range of valid δ_0. This improves upon the oracle result of Stucky and van de Geer [131], in which the parameter does depend on the level of confidence and the rate does not scale in the optimal way, i.e., as $\sqrt{(s/n)\log(p/s)}$. Moreover, we can see that our estimator is independent of the underlying standard deviation σ and of the sparsity s, even if the rates depend on them. Note that, up to some multiplicative constant, we obtain the same rates as for the Slope in Bellec, Lecué and Tsybakov [12]. In Su and Candès [133], the Slope estimator is proved to attain the sharp constant in the asymptotic framework where σ is known and for specific $\mathbb{X}$; whereas here we obtain only the minimax rates, but in a non-asymptotic framework, and under general assumptions on the design matrix $\mathbb{X}$.

For this estimator, we did not provide a bound for the l_1 norm, for the same reasons as in [12]. Indeed, the coefficients λ_j of the components of β are different in the sorted norm. As a consequence, we do not provide inequalities for l_q norms when $q < 2$, that are obtained by interpolation between the l_1 and l_2 norms.

2.7 Proofs

2.7.1 Preliminary lemmas

Let $\beta^* \in \mathbb{R}^p$, $\mathcal{S} \subset \{1, \dots, p\}$ with cardinality s and denote by $\mathcal{S}^C$ the complement of $\mathcal{S}$. For $i \in \{1, \dots, p\}$, let β_i^* be the i-th component of β^* and assume that for every $i \in \mathcal{S}^C$, $\beta_i^* = 0$.

Lemma 2.10. *We have*

$$\left|(\hat{\beta}^{SQL} - \beta^*)_{\mathcal{S}^C}\right|_1 \leq \left|(\hat{\beta}^{SQL} - \beta^*)_{\mathcal{S}}\right|_1 + \frac{1}{\lambda\sqrt{n}|\varepsilon|_2}\left\langle \mathbb{X}^T\varepsilon\ ,\ \hat{\beta}^{SQL} - \beta^*\right\rangle.$$

The proof follows from the arguments in Giraud [67, pages 110-111], and it is therefore omitted.

Lemma 2.11. *Let $u \in \mathbb{R}^p$ be defined by $u := \hat{\beta}^{SQS} - \beta^*$. We have*

$$\sum_{j=s+1}^{p} \lambda_j |u|_{(j)} \leq \sum_{j=1}^{s} \lambda_j |u|_{(j)} + \frac{1}{\sqrt{n}|\varepsilon|_2}\left\langle \mathbb{X}^T\varepsilon\ ,\ u\right\rangle.$$

Proof : We combine the arguments from Giraud [67, pages 110-111], and from the proof of Lemma A.1 in [12]. First, we remark that the sorted l_1 norm can be written as follows, for any $v \in \mathbb{R}^p$,

$$|v|_* = \max_{\phi} \sum_{j=1}^{p} \lambda_j \left|v_{\phi(j)}\right|,$$

where the maximum is taken over all permutations $\phi = (\phi(1), \dots, \phi(p))$ of $\{1, \dots, p\}$.

By definition, $\hat{\beta}^{SQS}$ is a minimizer of (2.3), so we have

$$|Y - \mathbb{X}\hat{\beta}^{SQS}|_2 - |Y - \mathbb{X}\beta^*|_2 \leq \sqrt{n}\left(|\beta^*|_* - |\hat{\beta}^{SQS}|_*\right).$$

Let ϕ be any permutation of $\{1,\dots,p\}$ such that

$$|\beta^*|_* = \sum_{j=1}^{s} \lambda_j |\beta^*_{\phi(j)}| \text{ and } |u_{\phi(s+1)}| \geq |u_{\phi(s+2)}| \geq \cdots \geq |u_{\phi(p)}|. \tag{2.23}$$

We have

$$\begin{aligned}
|\beta^*|_* - |\hat{\beta}^{SQS}|_* &\leq \sum_{j=1}^{s} \lambda_j \left(|\beta^*_{\phi(j)}| - |\hat{\beta}^{SQS}_{\phi(j)}| \right) - \sum_{j=s+1}^{p} \lambda_j |\hat{\beta}^{SQS}_{\phi(j)}| \\
&\leq \sum_{j=1}^{s} \lambda_j |u_{\phi(j)}| - \sum_{j=s+1}^{p} \lambda_j |\hat{\beta}^{SQS}_{\phi(j)}| \\
&= \sum_{j=1}^{s} \lambda_j |u_{\phi(j)}| - \sum_{j=s+1}^{p} \lambda_j |u_{\phi(j)}|.
\end{aligned}$$

Since the sequence λ_j is non-increasing, we have $\sum_{j=1}^{s} \lambda_j |u_{\phi(j)}| \leq \sum_{j=1}^{s} \lambda_j |u|_{(j)}$. The permutation ϕ satisfies (2.23), therefore, $\sum_{j=s+1}^{p} \lambda_j |u|_{(j)} \leq \sum_{j=s+1}^{p} \lambda_j |u_{\phi(j)}|$. From the previous inequalities, we get that

$$|Y - \mathbb{X}\hat{\beta}^{SQS}|_2 - |Y - \mathbb{X}\beta^*|_2 \leq \sqrt{n} \left(\sum_{j=1}^{s} \lambda_j |u|_{(j)} - \sum_{j=s+1}^{p} \lambda_j |u|_{(j)} \right). \tag{2.24}$$

By convexity of the mapping $\beta \mapsto ||Y - X\beta||_2$, we have

$$\begin{aligned}
&|Y - \mathbb{X}\hat{\beta}^{SQS}|_2 - |Y - \mathbb{X}\beta^*|_2 \\
&\qquad \geq -\left\langle \frac{\mathbb{X}^T\varepsilon}{|\varepsilon|_2} \,,\, \hat{\beta}^{SQS} - \beta^* \right\rangle = -\frac{1}{|\varepsilon|_2} \left\langle \mathbb{X}^T\varepsilon \,,\, \hat{\beta}^{SQS} - \beta^* \right\rangle.
\end{aligned} \tag{2.25}$$

Combining (2.24) and (2.25), we get

$$-\frac{1}{|\varepsilon|_2} \left\langle \mathbb{X}^T\varepsilon \,,\, \hat{\beta}^{SQS} - \beta^* \right\rangle \leq \sqrt{n} \left(\sum_{j=1}^{s} \lambda_j |u|_{(j)} - \sum_{j=s+1}^{p} \lambda_j |u|_{(j)} \right),$$

which concludes the proof.

□

Lemma 2.12. *We have*

$$|\mathbb{X}(\hat{\beta}^{SQL} - \beta^*)|_2^2 \leq \left\langle \mathbb{X}^T\varepsilon \,,\, \hat{\beta}^{SQL} - \beta^* \right\rangle + \lambda\sqrt{n} |Y - \mathbb{X}\hat{\beta}^{SQL}|_2 |\hat{\beta}^{SQL} - \beta^*|_1.$$

Lemma 2.13. *We have*

$$|\mathbb{X}(\hat{\beta}^{SQS} - \beta^*)|_2^2 \leq \left\langle \mathbb{X}^T\varepsilon \,,\, \hat{\beta}^{SQS} - \beta^* \right\rangle + \sqrt{n} |Y - \mathbb{X}\hat{\beta}^{SQS}|_2 |\hat{\beta}^{SQS} - \beta^*|_*.$$

Proof : We will give a general proof of Lemmas 2.12 and 2.13 in the case of an estimator defined by

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \left(\frac{1}{\sqrt{n}} |Y - \mathbb{X}\beta|_2 + ||\beta|| \right), \tag{2.26}$$

where $||\cdot||$ is a norm on $\mathbb{R}^p$. Lemmas 2.12 and 2.13 are obtained as special cases corresponding to $||\cdot|| = \lambda|\cdot|_1$ and $||\cdot|| = |\cdot|_*$. Denote by $||\cdot||_{dual}$ the norm dual to $||\cdot||$.

Since $\hat{\beta}$ is optimal, we know that $\mathbb{X}^T(Y - \mathbb{X}\hat{\beta})/(\sqrt{n}|Y - \mathbb{X}\hat{\beta}|_2)$ belongs to the subdifferential of the function $||\cdot||$ evaluated at $\hat{\beta}$. Thus, there exists $v \in \mathbb{R}^p$ such that $||v||_{dual} \leq 1$ and

$$\frac{\mathbb{X}^T(Y - \mathbb{X}\hat{\beta})}{\sqrt{n}|Y - \mathbb{X}\hat{\beta}|_2} + v = 0.$$

Thus, we have

$$|\mathbb{X}(\hat{\beta} - \beta^*)|_2^2 = \left\langle \mathbb{X}^T\varepsilon \ , \ \hat{\beta} - \beta^* \right\rangle + \sqrt{n}|Y - \mathbb{X}\hat{\beta}^*|_2\langle v \ , \ \hat{\beta} - \beta^*\rangle.$$

The conclusion results from the inequality

$$\langle v \ , \ \hat{\beta} - \beta^*\rangle \leq ||v||_{dual}||\hat{\beta} - \beta^*|| \leq ||\hat{\beta} - \beta^*||.$$

□

Lemma 2.14. *We have*

$$\gamma'\sqrt{(s/n)\log(2p/s)} \leq \sqrt{\sum_{j=1}^{s}\lambda_j^2} \leq \gamma'\sqrt{(s/n)\log(2ep/s)}.$$

Proof : From Stirling's formula, we deduce that $s\log(s/e) \leq \log(s!) \leq s\log(s)$. Therefore

$$s\log(2p/s) \leq \sum_{j=1}^{s}\log(2p/j) = \log(2p) - \log(s!) \leq s\log(2ep/s).$$

The conclusion follows from the definition of the λ_j in (2.18).

□

The following simple property is proved in Giraud [67, page 112]. For convenience, it is stated here as a lemma.

Lemma 2.15. *With $\mathbb{P}_{\beta^*}$-probability at least $1 - (1 + e^2)e^{-n/24}$, we have*

$$\frac{\sigma}{\sqrt{2}} \leq \frac{|\varepsilon|_2}{\sqrt{n}} \leq 2\sigma.$$

We will also use the following theorem from Bellec, Lecué and Tsybakov [12, Theorem 4.1].

Lemma 2.16. *Let $0 < \delta_0 < 1$ and let $\mathbb{X}$ in $\mathbb{R}^{n\times p}$ be a matrix such that*

$$\max_{j=1,\dots,p}||\mathbb{X}e_j||_n \leq 1.$$

For any $u = (u_1, \dots u_p)$ in $\mathbb{R}^p$, we define :

$$G(u) := (4 + \sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n,$$

$$H(u) := (4 + \sqrt{2})\sum_{j=1}^{p}|u|_{(j)}\sigma\sqrt{\frac{\log(2p/j)}{n}},$$

$$F(u) := (4 + \sqrt{2})\sigma\sqrt{\frac{\log(2p/s)}{n}}\left(\sqrt{s}|u|_2 + \sum_{j=s+1}^{p}|u|_{(j)}\right).$$

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$, then the random event

$$\left\{ \frac{1}{n}\varepsilon^T \mathbb{X}u \leq \max\Big(H(u), G(u)\Big), \forall u \in \mathbb{R}^p \right\},$$

is of probability at least $1 - \delta_0/2$.

Moreover, by the Cauchy-Schwarz inequality, we have $H(u) \leq F(u)$, for all u in $\mathbb{R}^p$.

2.7.2 Proof of Theorem 2.1

Lemma 2.16 allows one to control the random variable $\varepsilon^T \mathbb{X}u$ that appears in Lemmas 2.10 and 2.12 with $u := \hat{\beta}^{SQL} - \beta^*$. Our calculations will take place on an event of probability at least $1-\delta_0-(1+e^2)e^{-n/24}$, where both Lemmas 2.15 and 2.16 can be used. Applying Lemma 2.16, we will distinguish between the two cases : $G(u) \leq F(u)$ and $F(u) < G(u)$.

First case : $G(u) \leq F(u)$.

Then we have

$$(4+\sqrt{2})\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n \leq (4+\sqrt{2})\sqrt{\frac{\log(2p/s)}{n}}\left(\sqrt{s}|u|_2 + \sum_{j=s+1}^{p}|u|_{(j)}\right).$$

We will show first that u is in the SRE cone, so that we can use the SRE assumption. From Lemma 2.10, we have

$$\begin{aligned}
|u_{\mathcal{S}^c}|_1 &\leq |u_{\mathcal{S}}|_1 + \frac{1}{\lambda\sqrt{n}|\varepsilon|_2}\left\langle \mathbb{X}^T\varepsilon\ ,\ \hat{\beta}^{SQL} - \beta^* \right\rangle \\
&\leq |u_{\mathcal{S}}|_1 + \frac{1}{\sqrt{n}\lambda|\varepsilon|_2} n\sigma(4+\sqrt{2})\sqrt{\frac{\log(2p/s)}{n}}\left(\sqrt{s}|u|_2 + \sum_{j=s+1}^{p}|u|_{(j)}\right) \\
&\leq |u_{\mathcal{S}}|_1 + \frac{1}{4}\left(\sqrt{s}|u|_2 + |u_{\mathcal{S}^c}|_1\right),
\end{aligned}$$

where in the last inequality, we have used Lemma 2.15 and assumption (2.6). We deduce that

$$\frac{3}{4}|u|_1 \leq \frac{7}{4}|u_{\mathcal{S}}|_1 + \frac{1}{4}\sqrt{s}|u|_2 \leq \frac{7}{4}\sqrt{s}|u|_2 + \frac{1}{4}\sqrt{s}|u|_2 = 2\sqrt{s}|u|_2.$$

Therefore, we have

$$|u|_1 \leq \frac{8}{3}\sqrt{s}|u|_2, \tag{2.27}$$

and thus, the following inequality holds $|u|_1 \leq (1+c_0)\sqrt{s}|u|_2$, with $c_0 = 5/3$, allowing us to use the $SRE(s, 5/3)$ assumption.

From Lemmas 2.12 and 2.16, and using that, in view of the $SRE(s, 5/3)$ condition, $||\mathbb{X}u||_n \geq \kappa|u|_2$, we deduce that

$$\begin{aligned}
||\mathbb{X}u||_n^2 &\leq (4+\sqrt{2})\sigma\sqrt{\frac{\log(2p/s)}{n}}\left(\sqrt{s}|u|_2 + \sum_{j=s+1}^{p}|u|_{(j)}\right) \\
&\quad + \left(\frac{|\varepsilon|_2}{\sqrt{n}} + ||\mathbb{X}u||_n\right)\frac{8}{3}\lambda\sqrt{s}|u|_2 \\
&\leq (4+\sqrt{2})\frac{11}{3}\sigma\sqrt{s\frac{\log(2p/s)}{n}}\frac{||\mathbb{X}u||_n}{\kappa} + (2\sigma + ||\mathbb{X}u||_n)\frac{8}{3}\lambda\sqrt{s}\frac{||\mathbb{X}u||_n}{\kappa}.
\end{aligned}$$

Thus,

$$||\mathbb{X}u||_n \leq (4+\sqrt{2})\frac{11}{3}\sigma\sqrt{s\frac{\log(2p/s)}{n}}\frac{1}{\kappa} + (2\sigma + ||\mathbb{X}u||_n)\frac{8}{3}\lambda\sqrt{s}\frac{1}{\kappa}.$$

Under assumptions (2.5) and (2.6), we have

$$\frac{8\lambda\sqrt{s}}{3\kappa} = \frac{8\gamma}{3\kappa}\sqrt{\frac{s}{n}\log\left(\frac{2p}{s}\right)} \leq \frac{1}{2}.$$

Thus, we have

$$\begin{aligned}||\mathbb{X}u||_n &\leq 2\left(\frac{44+11\sqrt{2}}{3\kappa}\sigma\sqrt{\frac{s}{n}\log\left(\frac{2p}{s}\right)} + \frac{16\sigma\lambda\sqrt{s}}{3\kappa}\right)\\ &\leq \frac{88+22\sqrt{2}+32\gamma}{3\kappa}\sigma\sqrt{\frac{s}{n}\log\left(\frac{2p}{s}\right)}. \end{aligned} \tag{2.28}$$

We have proved in (2.27) that $|u|_1 \leq (1+c_0)\sqrt{s}|u|_2$, with $c_0 = 5/3$, so we get that $|u|_2 \leq ||\mathbb{X}u||_n/\kappa$. Therefore, we can deduce the following inequalities

$$|u|_2 \leq \frac{88+22\sqrt{2}+32\gamma}{3\kappa^2}\sigma\sqrt{\frac{s}{n}\log\left(\frac{2p}{s}\right)}, \tag{2.29}$$

$$|u|_1 \leq \frac{704+176\sqrt{2}+256\gamma}{9\kappa^2}\sigma s\sqrt{\frac{1}{n}\log\left(\frac{2p}{s}\right)}. \tag{2.30}$$

Second case : $F(u) \leq G(u)$.

Then we have

$$(4+\sqrt{2})\sqrt{\frac{\log(2p/s)}{n}}\left(\sqrt{s}|u|_2 + \sum_{j=s+1}^{p}|u|_{(j)}\right) \leq (4+\sqrt{2})\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n.$$

Thus

$$|u|_1 \leq \sqrt{s}|u|_2 + \sum_{j=s+1}^{p}|u|_{(j)} \leq \sqrt{\frac{\log(1/\delta_0)}{\log(2p/s)}}||\mathbb{X}u||_n.$$

From Lemmas 2.12 and 2.16, we find

$$\begin{aligned}||\mathbb{X}u||_n^2 &\leq (4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n + \lambda\left(\frac{|\varepsilon|_2}{\sqrt{n}} + ||\mathbb{X}u||_n\right)|u|_1\\ &\leq (4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n + \lambda\left(2\sigma + ||\mathbb{X}u||_n\right)\sqrt{\frac{\log(1/\delta_0)}{\log(2p/s)}}||\mathbb{X}u||_n.\end{aligned}$$

Thus,

$$||\mathbb{X}u||_n \leq (4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}} + \lambda\left(2\sigma + ||\mathbb{X}u||_n\right)\sqrt{\frac{\log(1/\delta_0)}{\log(2p/s)}}.$$

We have chosen $\lambda = \gamma\sqrt{\frac{1}{n}\log\left(\frac{2p}{s}\right)}$, therefore we have

$$||\mathbb{X}u||_n \leq \sigma\sqrt{\frac{\log(1/\delta_0)}{n}}(4+\sqrt{2}+2\gamma) + ||\mathbb{X}u||_n\gamma\sqrt{\frac{\log(1/\delta_0)}{n}}.$$

By assumption, $\exp(-n/4\gamma^2) \leq \delta_0$, thus we have

$$||\mathbb{X}u||_n \leq \sigma\sqrt{\frac{\log(1/\delta_0)}{n}}(8+2\sqrt{2}+4\gamma). \tag{2.31}$$

As a consequence, we have

$$|u|_1 \leq \sqrt{\frac{\log(1/\delta_0)}{\log(2p/s)}}||\mathbb{X}u||_n \leq \sigma\sqrt{\frac{\log^2(1/\delta_0)}{n\log(2p/s)}}(8+2\sqrt{2}+4\gamma). \tag{2.32}$$

We have also $\sqrt{s}|u|_2 \leq \sqrt{\frac{\log(1/\delta_0)}{\log(2p/s)}}||\mathbb{X}u||_n$, thus

$$|u|_2 \leq \sigma\sqrt{\frac{\log^2(1/\delta_0)}{sn\log(2p/s)}}(8+2\sqrt{2}+4\gamma). \tag{2.33}$$

As a conclusion, we can prove the result (2.7) by combining the inequalities (2.28) and (2.31). The general bound for $|u|_q$, with $1 \leq q \leq 2$ is a consequence of the norm interpolation inequality $|u|_q \leq |u|_1^{2/q-1}|u|_2^{2-2/q}$ which proves (2.8).

□

2.7.3 Proofs of the adaptive procedure

2.7.3.1 Proof of Theorem 2.3

We choose $s \in [1, s_*]$ and assume that $\beta^* \in B_0(s)$. Define $\mathbb{P} := \mathbb{P}_{\beta^*}$ and $m_0 := \lfloor \log_2(s) \rfloor + 1$.

For any $a > 0$, we have

$$\mathbb{P}\big(d(\tilde{\beta}, \beta^*) \geq a\big) \leq \mathbb{P}\big(d(\tilde{\beta}, \beta^*) \geq a, \tilde{m} \leq m_0\big) + \mathbb{P}(\tilde{m} \geq m_0 + 1). \tag{2.34}$$

On the event $\{\tilde{m} \leq m_0\}$, we have the decomposition

$$d(\tilde{\beta}, \beta^*) \leq \sum_{k=\tilde{m}+1}^{m_0} d\left(\hat{\beta}_{(2^{k-1})}, \hat{\beta}_{(2^k)}\right) + d\big(\hat{\beta}_{(2^{m_0})}, \beta^*\big). \tag{2.35}$$

Using Assumption 2.4.1, we get that,

$$\sum_{k=\tilde{m}+1}^{m_0} d\left(\hat{\beta}_{(2^{k-1})}, \hat{\beta}_{(2^k)}\right) \leq \sum_{k=\tilde{m}+1}^{m_0} 4\hat{\sigma}C_0 w(2^k) \tag{2.36}$$

$$\leq 4\hat{\sigma}C_0C'w(2^{m_0}) \leq 4\hat{\sigma}C_0C'C''w(s). \tag{2.37}$$

We have $2^{m_0} \leq 2s$, therefore applying Assumption (2.15), we have with $\mathbb{P}_{\beta^*}$-probability at least $1 - (2s/p)^{2s} - u_n$,

$$d(\hat{\beta}_{(2^{m_0})}, \beta^*) \leq \frac{C_2(\tilde{\gamma})}{\kappa^2}\sigma w(2s) \leq \frac{C_2(\tilde{\gamma})C''}{\kappa^2}\sigma w(s). \tag{2.38}$$

Combining equations (2.35), (2.37), (2.38) and Assumption 2.4.2, we get with $\mathbb{P}_{\beta^*}$-probability at least $1 - (2s/p)^{2s} - u_n - u_{n,p,M}$,

$$d(\tilde{\beta}, \beta^*) \leq \left(4\sigma C_0C'C''\alpha + \frac{C_2(\tilde{\gamma})C''}{\kappa^2}\right)\sigma w(s). \tag{2.39}$$

We now bound the probability $\mathbb{P}(\tilde{m} \geq m_0 + 1)$.

$$\begin{aligned}
\mathbb{P}(\tilde{m} \geq m_0 + 1) &\leq \sum_{m=m_0+1}^{M} \mathbb{P}(\tilde{m} = m_0 + 1) \\
&\leq \sum_{m=m_0+1}^{M} \sum_{k=m}^{M} \mathbb{P}\left(d\left(\hat{\beta}_{(2^{k-1})}, \hat{\beta}_{(2^k)}\right) > 4\hat{\sigma} C_0 w(2^k) \right) \\
&\leq \sum_{m=m_0+1}^{M} \sum_{k=m}^{M} \mathbb{P}\left(d\left(\hat{\beta}_{(2^{k-1})}, \beta^*\right) > 2\hat{\sigma} C_0 w(2^k) \right) \\
&\quad + \mathbb{P}\left(d\left(\hat{\beta}_{(2^k)}, \beta^*\right) > 2\hat{\sigma} C_0 w(2^k) \right) \\
&\leq 2 \sum_{m=m_0+1}^{M} \sum_{k=m-1}^{M} \mathbb{P}\left(d\left(\hat{\beta}_{(2^{k-1})}, \beta^*\right) > 2\hat{\sigma} C_0 w(2^k) \right) \\
&\leq 2 \sum_{m=m_0+1}^{M} \sum_{k=m-1}^{M} \mathbb{P}\left(d\left(\hat{\beta}_{(2^{k-1})}, \beta^*\right) > 2\hat{\sigma} C_0 w(2^k), \hat{\sigma} \geq \frac{\sigma}{2} \right) + \mathbb{P}\left(\hat{\sigma} < \frac{\sigma}{2} \right).
\end{aligned}$$

Combining the previous equation with Assumption 2.4.2, and then with Assumption (2.15), we get

$$\begin{aligned}
&\mathbb{P}(\tilde{m} \geq m_0 + 1) \\
&\quad \leq 2 \sum_{m=m_0+1}^{M} \sum_{k=m-1}^{M} \mathbb{P}\left(d\left(\hat{\beta}_{(2^{k-1})}, \beta^*\right) > \sigma C_0 w(2^k) \right) - u_{n,p,M} \\
&\quad \leq 2M^2 \left(\left(\frac{2s}{p} \right)^{2s} + u_n \right) - u_{n,p,M} \\
&\quad \leq 2(\log_2(s_*) + 1)^2 \left(\left(\frac{2s}{p} \right)^{2s} + u_n \right) - u_{n,p,M}.
\end{aligned}$$

As a consequence, we deduce the bound on $\tilde{s}$. Combining the last equation with equations (2.34) and (2.39), we finally get that

$$\begin{aligned}
\mathbb{P}\left(d(\tilde{\beta}, \beta^*) \geq \left(4\sigma C_0 C' C'' \alpha + \frac{C_2(\tilde{\gamma}) C''}{\kappa^2} \right) \sigma w(s) \right) \\
\leq 3(\log_2(s_*) + 1)^2 \left(\left(\frac{2s}{p} \right)^{2s} + u_n \right) - 2u_{n,p,M}.
\end{aligned}$$

□

2.7.3.2 Proof of Lemma 2.4

Now, we consider the general case of the function $w(b) = b^{1/q}\sqrt{(1/n)\log(ap/b)}$, with q a fixed number of the interval $[1, 2]$. The first case will correspond to $a = 1$ and $q = 2$ and the second case will correspond to $a = 2$ with any choice of q.

We want to that the first part of Assumption 2.4.1 is satisfied, i.e., w is increasing on the interval $[1, s_*]$. Let $b \in [1, s_*]$. We have

$$\begin{aligned}
w'(b) &= \frac{1}{q} b^{(1/q)-1} \sqrt{\frac{1}{n} \log\left(\frac{ap}{b}\right)} + b^{(1/q)} \frac{-\frac{1}{nb}}{2\sqrt{\frac{1}{n}\log\left(\frac{ap}{b}\right)}} \\
&= \frac{b^{(1/q)-1} n^{-1/2} \left((2/q) \log\left(\frac{ap}{b}\right) - 1 \right)}{2\sqrt{\log\left(\frac{ap}{b}\right)}},
\end{aligned}$$

which is positive when $(2/q)\log\left(\frac{ap}{b}\right) - 1 \geq 0$, that is, when $b \leq ape^{-q/2}$.

We have $b \leq s_* \leq p/e = ape^{-q/2}$ when $a = 1$ and $q = 2$. When $a = 2$ and $q \in [1, 2]$, $p/e \leq 2pe^{-1} \leq ape^{-q/2}$. In the two cases we consider, we have proved that $w'(\cdot) \geq 0$ on the interval $[1, s_*]$, thus the function w is increasing on this interval. This proves that the first part of Assumption 2.4.1 is satisfied.

Let m be an integer in the interval $[1, M]$.

$$\begin{aligned}
\sum_{k=1}^{m} w(2^k) &= \sum_{k=1}^{m} 2^{k/q}\sqrt{\frac{1}{n}\log\left(\frac{ap}{2^k}\right)} = \sum_{k=0}^{m-1} 2^{(m-k)/q}\sqrt{\frac{1}{n}\log\left(\frac{ap}{2^{m-k}}\right)} \\
&= \frac{2^{m/q}}{\sqrt{n}}\sum_{k=0}^{m-1}\frac{1}{2^{k/q}}\sqrt{\left(\log\left(\frac{ap}{2^m}\right) + k\log(2)\right)} \\
&\leq \frac{2^{m/q}}{\sqrt{n}}\left(\sum_{k=0}^{m-1}\frac{1}{2^{k/q}}\sqrt{\log\left(\frac{ap}{2^m}\right)} + \sum_{k=0}^{m-1}\frac{\sqrt{k}}{2^{k/q}}\sqrt{\log(2)}\right) \\
&\leq \frac{2^{m/q}}{\sqrt{n}}\left(\sqrt{\log\left(\frac{ap}{2^m}\right)}\frac{1}{1-2^{-1/q}} + \sum_{k=0}^{m-1}\frac{4}{2^{k/2q}}\sqrt{\log(2)}\right) \\
&\leq 2^{m/q}\sqrt{\frac{1}{n}\log\left(\frac{ap}{2^m}\right)}\left(\frac{1}{1-2^{-1/q}} + \frac{4\sqrt{\log(2)}}{1-2^{-1/(2q)}}\right),
\end{aligned}$$

which proves that the second part is satisfied.

Let b be an integer of $[1, s_*]$. We have $w(2b) = (2b)^{1/q}\sqrt{(1/n)\log(2p/(2b))} \leq 2^{1/q}w(b)$, which proves that the third part is satisfied.

□

2.7.3.3 Proof of Lemma 2.5

We have $\beta^* \in B_0(s) \subset B_0(2^{M+1})$, therefore, we can apply Corollary 2.2 and Lemma 2.15, we have with $\mathbb{P}_{\beta^*}$-probability at least $1 - (2^{M+1}/p)^{2^{M+1}} - (1+e^2)e^{-n/24}$,

$$\begin{aligned}
\hat{\sigma} &\leq ||\varepsilon||_n + \left|\left|\mathbb{X}(\hat{\beta}_{(2^{M+1})} - \beta^*)\right|\right|_n \\
&\leq 2\sigma + \frac{C_2(\gamma)}{\kappa_*^2}\sigma\sqrt{\frac{2^{M+1}}{n}\log\left(\frac{p}{2^{M+1}}\right)} \\
&\leq \sigma\left(2 + \frac{C_2(\gamma)}{\kappa_*^2}\sqrt{\frac{2s}{n}\log\left(\frac{2p}{s}\right)}\right) \leq \sigma\left(2 + \frac{3\sqrt{2}C_2(\gamma)}{16\kappa_*\gamma}\right),
\end{aligned}$$

and

$$\begin{aligned}
\hat{\sigma} &\geq ||\varepsilon||_n - \left|\left|\mathbb{X}(\hat{\beta}_{(2^{M+1})} - \beta^*)\right|\right|_n \\
&\geq \frac{\sigma}{\sqrt{2}} - \frac{C_2(\gamma)}{\kappa_*^2}\sigma\sqrt{\frac{2^{M+1}}{n}\log\left(\frac{p}{2^{M+1}}\right)} \\
&\geq \sigma\left(\frac{1}{\sqrt{2}} - \frac{\sqrt{2}C_2(\gamma)}{\kappa_*^2}\sqrt{\frac{s}{n}\log\left(\frac{2p}{s}\right)}\right) \\
&\geq \sigma\left(\frac{1}{\sqrt{2}} - \frac{\sqrt{2}C_2(\gamma)}{\kappa_*^2}\sqrt{\frac{2s_*}{n}\log\left(\frac{p}{s_*}\right)}\right) \\
&\geq \sigma\left(\frac{1}{\sqrt{2}} - \sqrt{\left(\frac{1}{\sqrt{2}} - \frac{1}{2}\right)^2}\right) \geq \frac{\sigma}{2}.
\end{aligned}$$

□

2.7.4 Proof of Theorem 2.8

We act as in Section 2.7.2, with suitable modifications. We place ourselves in the event where both Lemmas 2.15 and 2.16 are valid, and set now $u := \hat{\beta}^{SQS} - \beta^*$. Applying Lemma 2.16, we will distinguish between the two cases : $G(u) \leq H(u) + \sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2}$ and $H(u) + \sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2} < G(u)$.

First case : $G(u) \leq H(u) + \sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2}$.

Applying Lemma 2.11, Lemma 2.16 and then Lemma 2.15, we have

$$\begin{aligned}
|u|_* &= \sum_{j=1}^p \lambda_j |u|_{(j)} \\
&\leq 2\sum_{j=1}^s \lambda_j |u|_{(j)} + \frac{1}{\sqrt{n}|\varepsilon|_2}\left\langle \mathbb{X}^T\varepsilon \ , \ \hat{\beta}^{SQS} - \beta^* \right\rangle \\
&\leq 2\sqrt{\sum_{j=1}^s \lambda_j^2}|u|_2 + \frac{n}{\sqrt{n}|\varepsilon|_2}\left((4+\sqrt{2})\frac{\sigma}{\gamma'}|u|_* + \sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2}\right) \\
&\leq 4\sqrt{\sum_{j=1}^s \lambda_j^2}|u|_2 + \frac{8+2\sqrt{2}}{\gamma'}|u|_*,
\end{aligned}$$

and we get

$$|u|_* \leq \frac{4|u|_2}{1 - \dfrac{8+2\sqrt{2}}{\gamma'}}\sqrt{\sum_{j=1}^s \lambda_j^2},$$

Using assumption (2.19), we have $\gamma' \geq 16 + 4\sqrt{2}$, therefore $|u|_* \leq 8|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2}$. As a consequence, we get $u \in C_{WRE}(s, c_0)$ with $c_0 := 8$. Invoking Lemmas 2.13, 2.14, 2.16 and using the $WRE(s, c_0)$ condition, we get

$$\begin{aligned}
||\mathbb{X}u||_n^2 &\leq \frac{1}{n}\left\langle \mathbb{X}^T\varepsilon \ , \ u \right\rangle + \frac{1}{\sqrt{n}}|Y - \mathbb{X}\hat{\beta}|_2|u|_* \\
&\leq (4+\sqrt{2})\frac{\sigma}{\gamma'}|u|_* + \sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2} + (2\sigma + ||\mathbb{X}u||_n)|u|_* \\
&\leq \left((32+8\sqrt{2})\frac{\sigma}{\gamma'} + 17\sigma + 8||\mathbb{X}u||_n \right)|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2} \\
&\leq \left((32+8\sqrt{2})\frac{\sigma}{\gamma'} + 17\sigma + 8||\mathbb{X}u||_n \right)\frac{||\mathbb{X}u||_n}{\kappa'}\gamma'\sqrt{(s/n)\log(2ep/s)}.
\end{aligned}$$

Thus,

$$||\mathbb{X}u||_n \leq \frac{\sigma}{\kappa'}\sqrt{\frac{s}{n}\log\left(\frac{2ep}{s}\right)}\frac{32+8\sqrt{2}+17\gamma'}{1 - \dfrac{8\gamma'}{\kappa'}\sqrt{\dfrac{s}{n}\log\left(\dfrac{2ep}{s}\right)}}.$$

Applying condition (2.19), we obtain

$$||\mathbb{X}u||_n \leq (64 + 16\sqrt{2} + 34\gamma')\frac{\sigma}{\kappa'}\sqrt{\frac{s}{n}\log\left(\frac{2ep}{s}\right)}. \tag{2.40}$$

This and the WRE condition imply

$$|u|_2 \leq (64 + 16\sqrt{2} + 34\gamma')\frac{\sigma}{\kappa'^2}\sqrt{\frac{s}{n}\log\left(\frac{2ep}{s}\right)}. \tag{2.41}$$

Therefore, using the inequality $|u|_* \leq 8|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2}$, we get from Lemma 2.14

$$|u|_* \leq 8(64 + 16\sqrt{2} + 34\gamma')\gamma'\frac{\sigma}{\kappa'^2}\frac{s}{n}\log\left(\frac{2ep}{s}\right). \tag{2.42}$$

Second case : $H(u) + \sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2} \leq G(u)$.

Then we have

$$(4+\sqrt{2})\frac{\sigma}{\gamma'}|u|_* + \sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2} \leq (4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n.$$

Therefore we have

$$|u|_* \leq \gamma'\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n, \text{ and } |u|_2\sqrt{\sum_{j=1}^s \lambda_j^2} \leq (4+\sqrt{2})\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n. \tag{2.43}$$

Invoking Lemmas 2.13 and 2.16, and using (2.43), we get

$$\begin{aligned}||\mathbb{X}u||_n^2 &\leq (4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n + \sigma|u|_2\sqrt{\sum_{j=1}^s \lambda_j^2} + (2\sigma + ||\mathbb{X}u||_n)|u|_* \\ &\leq (4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n + \sigma(4+\sqrt{2})\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n \\ &\quad + (2\sigma + ||\mathbb{X}u||_n)\gamma'\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n.\end{aligned}$$

which yields

$$||\mathbb{X}u||_n \leq (8 + 2\sqrt{2} + 2\gamma')\sigma\sqrt{\frac{\log(1/\delta_0)}{n}} + ||\mathbb{X}u||_n\gamma'\sqrt{\frac{\log(1/\delta_0)}{n}},$$

We have chosen $\exp(-n/4\gamma'^2) \leq \delta_0$, which implies that

$$||\mathbb{X}u||_n \leq (16 + 4\sqrt{2} + 4\gamma')\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}. \tag{2.44}$$

We can deduce from (2.43) that

$$|u|_* \leq (16 + 4\sqrt{2} + 4\gamma')\sigma\gamma'\frac{\log(1/\delta_0)}{n}, \tag{2.45}$$

and combining the second part of (2.43) with Lemma 2.14, we get

$$\begin{aligned}|u|_2\gamma'\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)} &\leq (4+\sqrt{2})\sqrt{\frac{\log(1/\delta_0)}{n}}||\mathbb{X}u||_n \\ &\leq (4+\sqrt{2})(16 + 4\sqrt{2} + 4\gamma')\sigma\frac{\log(1/\delta_0)}{n}.\end{aligned}$$

Finally, we get that

$$|u|_2 \leq \frac{(4+\sqrt{2})(16 + 4\sqrt{2} + 4\gamma')}{\gamma'}\sigma\sqrt{\frac{\log^2(1/\delta_0)}{sn\log(p/s)}}. \tag{2.46}$$

□

Acknowledgement

This work is supported by the Labex Ecodec under the grant ANR-11-LABEX-0047 from the French Agence Nationale de la Recherche. The author thanks Professor Alexandre Tsybakov for helpful comments and discussions. The author acknowledge the Associate Editor and two anonymous reviewers for their comments which lead to significant improvements of this paper.

Chapter 3

Robust-to-outliers simultaneous inference and noise level estimation using a MOM approach

Abstract

In this article, we extend the results of the previous chapter to a more robust setting, where outliers are present in the data. We present a family of penalized Median-of-Means (MOM) estimators to estimate conditional mean models, and as a special case, high-dimensional linear regression models. Our procedures allows simultaneous inference of conditional mean and noise level using a joint convex-concave optimization procedure, that can be computed easily. We give simultaneous estimation bounds for the conditional mean and the noise level, as well as for the risk of the estimator. In the high-dimensional linear regression, we show that our estimator is minimax optimal in estimation with the l_p norm while being robust to outliers and adaptive to the noise variance. Bounds for the estimated standard deviation are also given.

Keywords: Median-of-means, robustness, adaptivity, minimax optimal rates.

Based on [37]: Derumigny, A., Robust-to-outliers simultaneous inference and noise level estimation using a MOM approach, *in progress*, 2019.

3.1 Introduction

One of the most simple problem in statistics is the estimation of an univariate mean. Of course, it is always possible to use the empirical mean, but it may not be an ideal choice. Indeed, the empirical mean is not robust, meaning that only one outlier in the dataset can push the mean towards infinity. One of the procedures to guarantee optimality even in the presence of outliers is the Median-of-Means (MOM) framework [43]. Its principle is very simple: divide the sample in K blocks, compute the empirical mean on each block, and return the median of these means. Note that this procedure is naturally robust to the presence of up to $K/2$ outliers. In the multivariate case, it is more difficult to construct robust and optimal estimators, see [98, 72].

Recently MOM procedures have received a great amount of attention for estimating conditional means, see [90, 91, 104]. Furthermore, Median-of-Means (MOM) type approaches have been developed for a lot of frameworks such as classification [92], empirical risk minimization [30], sub-sampling and hyper-parameter tuning [88], and reproducing kernel Hilbert space embedding [96].

In the recent paper [91], it was proved that minimaximization of a MOM-type criteria can be used to construct robust minimax optimal estimators in the linear high-dimensional framework. Nevertheless, these estimators require the knowledge of the standard deviation of the noise, which is unknown in practice. It is proposed in [91] to use the square-root Lasso or the square-root Slope to estimate the noise level at a first step. If there is even one outlier in the data, this noise level estimator will not be robust which makes the whole procedure not robust as a consequence.

In Section 3.2, we propose a joint estimator for the parameter and the noise level in the high dimensional framework under sparsity. We propose an algorithm to compute it using penalized optimization techniques. We show that, under some conditions, this estimator can achieve optimal rates of estimation. This is proved by an application of Section 3.3 where we give a more general result of the estimation of a conditional mean in a given function class F under some technical assumptions. The rest of the chapter is devoted to the proofs of the main results.

3.2 Results in the high-dimensional linear regression framework

Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$, with $d \geq 1$ be two random variables such that

$$Y = X^T \beta^* + \zeta, , \tag{3.1}$$

where ζ is a real random variable satisfying $\mathbb{E}|\zeta| < +\infty$. We assume that $\mathbb{E}[\zeta] = 0$ and define $\sigma^{*2} := \mathbb{E}[\zeta^2]$ unknown, but greater than a lower bound $\sigma_{\min}$ and $\kappa^* := \mathbb{E}[\zeta^4]/\sigma^{*4}$. We do not necessarily assume that ζ is independent of X, but we will assume that the link between ζ and X is not too strong, see Assumption 3.2.1.

Our goal is to estimate (β^*, σ^*) in Model 3.1, assuming that β^* is sparse, in the sense that $|\beta^*|_0 \leq s$, where s is a given integer smaller than d and $|\beta^*|_0$ is the number of coefficients of β^*. Finally, we assume that we observe $n > 0$ pairs of random variables $(X_i, Y_i)_{i=1,\dots,n} \in \mathbb{R}^{n \times (d+1)}$, where the observations are divided into the two following groups:

- the informative group $\mathcal{I} \subset \{1, \dots, n\}$, where the pairs $(X_i, Y_i), i \in \mathcal{I}$ are independent and follow the same distribution as (X, Y) ;
- the outlier group $\mathcal{O} \subset \{1, \dots, n\}$: for every $i \in \mathcal{O}$, (X_i, Y_i) are outliers, meaning that they may not follow Model (3.1) ; they may be deterministic and even adversarial in the sense that they may depend on the informative data $(X_i, Y_i)_{i \in \mathcal{I}}$ defined above and on the choice of the estimator.

Obviously, we have $\mathcal{I} \cup \mathcal{O} = \{1, \dots, n\}$ and $\mathcal{I} \cap \mathcal{O} = \emptyset$ meaning that, for each pair (X_i, Y_i), the corresponding index $i \in \{1, \dots, n\}$ belongs to one of the two groups. Of course, it is unknown to the statistician whether i belongs to $\mathcal{I}$ or $\mathcal{O}$. Otherwise, we would just remove the outlier group. To construct a robust estimator, we will use the Median-of-Means framework. Our estimator is defined as follows:

1. Given a dataset $\mathcal{D} = \{(X_i, Y_i), i = 1, \dots, n\}$ of size $n > 0$, we can divide it into K blocks $\mathcal{D}_1, \dots \mathcal{D}_K$ of size n/K (assumed to be an integer) corresponding to a partition $\{1, \dots, n\} = B_1 \sqcup \dots \sqcup B_K$, where K is a tuning parameter to be chosen later ;

2. For each $k = 1, \dots, K$, the criteria of (β, σ) against $(\tilde{\beta}, \chi)$ on the block B_k is defined by

$$P_{B_k}\left(\frac{l_\beta}{\sigma} + \sigma - \frac{l_{\tilde{\beta}}}{\chi} - \chi\right) := \frac{1}{|B_k|}\sum_{i \in B_k}\left(\frac{(Y_i - X_i^T\beta)^2}{\sigma} + \sigma - \frac{(Y_i - X_i^T\tilde{\beta})^2}{\chi} - \chi\right),$$

for every $\beta, \sigma, \tilde{\beta}, \chi \in \mathbb{R}^p \times \mathbb{R}_+ \times \mathbb{R}^p \times \mathbb{R}_+$, where $|B_k|$ denotes the cardinal of B_k ;

3. The global MOM criteria of (β, σ) against $(\tilde{\beta}, \chi)$ is defined by

$$MOM_K\left(\frac{l_\beta}{\sigma} + \sigma - \frac{l_{\tilde{\beta}}}{\chi} - \chi\right) := Median\left\{P_{B_k}\left(\frac{l_\beta}{\sigma} + \sigma - \frac{l_{\tilde{\beta}}}{\chi} - \chi\right),\, k = 1, \dots, K\right\};$$

4. Finally, the MOM-K estimator of (β^*, σ^*) is defined by

$$(\hat{\beta}, \hat{\sigma}) := \underset{\beta \in \mathbb{R}^d,\, \sigma > \sigma_{\min}}{\arg\min} \max_{\tilde{\beta} \in \mathbb{R}^d,\, \chi > \sigma_{\min}} \left[MOM_K\left(\frac{l_\beta}{\sigma} + \sigma - \frac{l_{\tilde{\beta}}}{\chi} - \chi\right) + \mu\big(|\beta|_1 - |\tilde{\beta}|_1\big)\right], \tag{3.2}$$

where μ is a tuning parameter to be chosen.

An algorithm to compute the joint estimator $(\hat{\beta}, \hat{\sigma})$ is proposed in Algorithm 4. The main idea is to alternate minimization steps in (β, σ) and maximization steps in $(\tilde{\beta}, \rho)$. For this reason, we call (3.2) a minmax-MOM estimator. The minimization and maximization steps should only use data from the central block, i.e. the block B_k realizing the median in step 3 above. Indeed, the MOM_K criteria is not affected by local variations of $(\beta, \sigma, \tilde{\beta}, \chi)$ in the other blocks $B_{k'}$ when $k' \neq k$.

As the minimization and maximization steps are similar, we detail only the former. We separate the joint minimization step in (β, σ) in two steps, one in β, and one in σ. When σ and $(\tilde{\beta}, \chi)$ are fixed, the partial minimization in β of the criterion (3.2) is locally equivalent to the minimization of the program $\min_\beta \sum_{i \in B_k}(Y_i - X_i^T\beta)^2 + \mu\sigma|\beta|_1$. This is the classical Lasso program on the dataset $\mathcal{D}_k$, with a tuning parameter defined as $\lambda := \mu\sigma$. Therefore, any algorithm to compute the usual Lasso estimator (or even one step of the optimization program of the Lasso) may be used to update β. In Algorithm 4 and as an example, we choose to do a subgradient step. It is equal to

$$2\mathbb{X}_k^T(\mathbb{Y}_k - \mathbb{X}_k\beta^{(t-1)}) - \sigma^{(t-1)}sign(\beta^{(t-1)}),$$

where $\sigma^{(t-1)}$ is the estimated standard deviation at the previous step, $sign$ is the component-wide sign, with the convention $sign(0) = 0$, $\mathbb{Y}_k$ is the vector $(Y_i)_{i \in B_k}$ and $\mathbb{X}_k$ is the matrix whose lines are the vectors X_i^T, for $i \in B_k$. Note that other choices of updates for β are possible, such as the use of proximal gradient descent, alternating direction method of multipliers (ADMM), and cyclic coordinate descent, as for non-adaptive minmax-MOM estimators (see respectively Algorithms 2, 3 and 4 in [91]).

When β and $(\tilde{\beta}, \chi)$ are fixed, the partial minimization in σ of the criterion (3.2) is locally equivalent to the minimization of the program $\min_\sigma a/\sigma + \sigma$, where $a = (1/|B_k|)\sum_{i \in B_k}(Y_i - X_i^T\beta)^2$. This optimization program admits a closed-form solution $\sigma = \sqrt{a}$, which is the update step of σ that we will use. If we observe $\sigma < \sigma_{\min}$, then we update $\sigma := \sigma_{\min}$, but this should not occur too much often in practice. Indeed, if a given σ is very small, it means that for one of the blocks, there exists a parameter β for which $a = (1/|B_k|)\sum_{i \in B_k}(Y_i - X_i^T\beta)^2$. This would mean that at least for one of the blocks, the noise has a very small intensity, and should rather invite the applied statistician to set a lower $\sigma_{\min}$. There exists other possibilities of updating σ. For instance, if σ is too much unstable, then we could use instead the update step $\sigma^{(t)} \leftarrow (\sigma^{(t-1)} + \sqrt{a})/2$ or other autoregressive filters. This would slow the convergence of σ, but would introduce more stability.

The maximization in $(\tilde{\beta}, \chi)$ follows exactly the same procedure as the minimization in (β, σ): we do a first update step on $\tilde{\beta}$ and a second update step of the parameter σ. Following the conclusions of [91], we use a random choice of blocks at each step. Such a choice is useful in the sense that it should help to avoid saddle-points. Indeed, when the blocks are fixed, only the central block is used in the optimization step. This could mean that other blocks don't influence the optimization process. By randomizing the blocks, we are sure to change at each step which dataset is used. This allows to "explore" better the space of parameters, and improves strongly the performance of minmax-MOM algorithms.

Algorithm 4: An alternating sub-gradient algorithm to compute the adaptive minmax MOM-LASSO estimator of (β^*, σ^*) using random blocks

Input: a dataset $\mathcal{D} = (X_i, Y_i)_{i=1,\dots,n}$;

Input: tuning parameters $\mu, \sigma_{\min} > 0$, $K \in \{1, \dots, n\}$, two step size sequences $(\eta_t)_{t\in\mathbb{N}}$ and $(\tilde{\eta}_t)_{t\in\mathbb{N}}$

Input: initial point $(\beta^{(0)}, \sigma^{(0)}, \tilde{\beta}^{(0)}, \chi^{(0)}) \in \mathbb{R}^d \times \mathbb{R}_+ \times \mathbb{R}^d \times \mathbb{R}_+$, and stopping criteria $\epsilon_0 > 0$

Initialize $t \leftarrow 0$;

repeat

Update $t \leftarrow t + 1$;

`/* Update of` β `and` σ `*/`

Partition $\{1, \dots, n\}$ into K blocks $B_1, \dots B_K$ at random ;

Find $k \in \{1, \dots, K\}$ such that

$$MOM_K\left(\frac{l_{\beta^{(t-1)}}}{\sigma^{(t-1)}} + \sigma^{(t-1)} - \frac{l_{\tilde{\beta}^{(t-1)}}}{\chi^{(t-1)}} - \chi^{(t-1)}\right) = P_{B_k}\left(\frac{l_{\beta^{(t-1)}}}{\sigma^{(t-1)}} + \sigma^{(t-1)} - \frac{l_{\tilde{\beta}^{(t-1)}}}{\chi^{(t-1)}} - \chi^{(t-1)}\right) ;$$

Update $\beta^{(t)} \leftarrow \beta^{(t-1)} + 2\eta_t \mathbb{X}_k^T(\mathbb{Y}_k - \mathbb{X}_k\beta^{(t-1)}) - \mu\sigma^{(t-1)}\eta_t sign(\beta^{(t-1)})$;

Update $\sigma^{(t)} \leftarrow |\mathbb{Y}_k - \mathbb{X}_k\beta^{(t)}|_2/|B_k|^{1/2}$;

`/* Update of` $\tilde{\beta}$ `and` χ `*/`

Partition $\{1, \dots, n\}$ into K blocks $B_1, \dots B_K$ at random ;

Find $k \in \{1, \dots, K\}$ such that

$$MOM_K\left(\frac{l_{\beta^{(t)}}}{\sigma^{(t)}} + \sigma^{(t)} - \frac{l_{\tilde{\beta}^{(t-1)}}}{\chi^{(t-1)}} - \chi^{(t-1)}\right) = P_{B_k}\left(\frac{l_{\beta^{(t)}}}{\sigma^{(t)}} + \sigma^{(t)} - \frac{l_{\tilde{\beta}^{(t-1)}}}{\chi^{(t-1)}} - \chi^{(t-1)}\right) ;$$

Update $\tilde{\beta}^{(t)} \leftarrow \tilde{\beta}^{(t-1)} + 2\tilde{\eta}_t \mathbb{X}_k^T(\mathbb{Y}_k - \mathbb{X}_k\tilde{\beta}^{(t-1)}) - \mu\chi^{(t-1)}\tilde{\eta}_t sign(\tilde{\beta}^{(t-1)})$;

Update $\chi^{(t)} \leftarrow |\mathbb{Y}_k - \mathbb{X}_k\tilde{\beta}^{(t)}|_2/|B_k|^{1/2}$;

until $|\beta^{(t-1)} - \beta^{(t)}|_2 < \epsilon_0$, $|\tilde{\beta}^{(t-1)} - \tilde{\beta}^{(t)}|_2 < \epsilon_0$, $|\sigma^{(t-1)} - \sigma^{(t)}| < \epsilon_0$ *and* $|\chi^{(t-1)} - \chi^{(t)}| < \epsilon_0$;

Output: a joint estimator $\left(\hat{\beta}, \hat{\sigma}\right) := \left(\beta^{(t)}, \sigma^{(t)}\right)$;

To prove statistical properties of $(\hat{\beta}, \hat{\sigma})$, we will need the following set of assumptions.

Assumption 3.2.1. *Denote by $(e_j)_{j=1,\dots d}$ the canonical basis of $\mathbb{R}^d$. We assume that there exist some finite constants $C_1, C_2, C_3, \theta_0, \theta_m$ such that*

1. $|\mathcal{I}| \geq n/2$ *and* $|\mathcal{O}| \leq C_1 s\log(ed/s)$,
2. X *is isotropic and* $\forall t \in \mathbb{R}^d$, $1 \leq p \leq C_2\log(ed)$, $1 \leq j \leq d$, $||X^T e_j||_{L_p} \leq C_3\sqrt{p}\,||X^T e_j||_{L_2}$,
3. $\forall t \in \mathbb{R}^d$, $||X^T t||_{L_2} \leq \theta_0||X^T t||_{L_1}$, *and* $Var\left[\zeta X^T t\right] \leq \theta_m||X^T t||_{L_2}$,

where for $p > 0$ and for any function $f : \mathbb{R}^p \to \mathbb{R}$, $||f(X)||_{L_p}$ is defined, if it exists, as $||f(X)||_{L_p(P_X)} = \left(\int |f(x)|^p dP_X(x)\right)^{1/p}$, P_X denotes the law of the random vector X, and X^T denotes the transpose of the random vector X.

The first assumption affirms that the number of informative data should be at least half of the sample size. Conversely, if more than half of the data have been corrupted, than there is little hope of recovering information about the informative data. Second, the outlier should not be too much numerous compared to the number of observations times the minimax optimal rates. This is similar to the assumption needed in [91]. The second assumption is necessary to apply [103, Theorem 1.6] to bound Gaussian mean widths (that are related to the Rademacher complexities that we will define in Section 3.3) by local Gaussian mean widths. These local Gaussian mean widths will be bounded by the minimax rates using [93, Lemma 5.3]. These two steps will ensure that our estimator of β^* attains the optimal minimax rates. The third assumption is a special case of small ball property, which is often needed to establish bounds in learning without using concentration inequalities, see [102, 83]. Finally, the last inequality is satisfied when ζ and X are independent. We relax this hypothesis of independence, but still assume that the dependence between ζ and X is not too strong.

The following theorem is proved in Section 3.7.2, and show that the estimator $(\hat{\beta}, \hat{\sigma})$ attains the minimax optimal rate adaptively in σ.

Theorem 3.1. *Assume that β^* is s-sparse for a given $s \leq d$, and that $n \gtrsim s\log(ed/s)$. Under Assumption 3.2.1 and if there exist c, c' such that Conditions (3.5)-(3.13) are satisfied for $\mu_0 = \kappa^{*1/4}\sigma^*$, there exist constants $c_1, c_2, c_3, c_4, c_5, c_6 > 0$ independent of $s, d, n, \sigma^*, \kappa$ such that, choosing the regularization parameter as*

$$\mu := c_1\sqrt{\frac{1}{n}\log\left(\frac{ed}{s}\right)},$$

we have

$$\begin{aligned}\mathbb{P}\Bigg(\forall p \in [1,2],\ |\hat{\beta} - \beta^*|_p \leq c_2||\zeta||_{L^4}s^{1/p}\sqrt{\frac{1}{n}\log\left(\frac{ed}{s}\right)} \text{ and } |\hat{\sigma} - \sigma^*| \leq \sigma^*(K/n)^{1/2}(\kappa^* - 1)^{1/2}/10\Bigg)\\ \geq 1 - c_3\exp(-K/c_4),\end{aligned}$$

for every integer K such that $c_5\max\left(|\mathcal{O}|, ||\zeta||_{L^4}s\log(ed/s)\right) \leq K \leq c_6 n$.

Note that these constants are fixed but might be difficult to compute in practice. To obtain precise values for these constants, for instance c_1, one would need to quantify the constants c_2 in [103, Theorem 1.6] for $q_0 = 4$ and $L = 2$ and C in [93, Lemma 5.3], using their notation.

3.3 A general framework

In this section, we generalize the results of the previous section. We now assume that the explanatory variable X has values in a measurable space $\mathcal{X}$, and we denote by F a class of measurable functions from $\mathcal{X}$ to $\mathbb{R}$ included in the space $L_2(P_X)$, where P_X is the law of X. This means that for every $f \in F$, $||f(X)||_{L_2}^2 := \int f(x)^2 dP_X(x)$ is supposed to be finite. The random variable Y is still with values in $\mathbb{R}$, and we assume that

$$Y = f^*(X) + \zeta, \tag{3.3}$$

for a given unknown function f^* in the class F. The assumptions on ζ are the same as in the previous section: it has finite moments at least of the order 4, and an unknown square deviation σ^* greater than a lower bound $\sigma_{\min} > 0$. This lower bound will be needed in the proofs, see Section 3.6.9.1.

Note that, if $\mathcal{X} = \mathbb{R}^d$ for a fixed $d > 0$ and F is the class of sparse linear functionals, i.e. if for a given $s \leq d$, we have $F = \{f_\beta := (x \in \mathbb{R}^d \mapsto \beta^T x \in \mathbb{R}),\ \beta \in \mathbb{R}^d,\ |\beta|_0 \leq s\}$, we recognize indeed the framework of the previous section.

Finally, our goal is to estimate (f^*, σ^*) given a dataset of random pairs $\mathcal{D} = (X_i, Y_i), i = 1, \ldots, n$, where the observations are split in two groups : the informative group $\mathcal{I}$, for which (X_i, Y_i) are independent and distributed as (X, Y) ; and the outlier group $\mathcal{O}$, made up of potentially adversarial variables. Our estimators is a generalization of (3.2), defined by :

1. Given a dataset $\mathcal{D} = \{(X_i, Y_i), i = 1, \ldots, n\}$ of size $n > 0$, we can divide it into K blocks $\mathcal{D}_1, \ldots \mathcal{D}_K$ of size n/K (assumed to be an integer) corresponding to a partition $\{1, \ldots, n\} = B_1 \sqcup \cdots \sqcup B_K$;

2. For each $k = 1, \ldots, K$, the criteria of (f, σ) against $(\tilde{f}, \chi)$ on the block B_k is defined by

$$P_{B_k}\left(\frac{l_f}{\sigma} + \sigma - \frac{l_{\tilde{f}}}{\chi} - \chi\right) := \frac{1}{|B_k|}\sum_{i \in B_k}\left(\frac{(Y_i - f(X_i))^2}{\sigma} + \sigma - \frac{(Y_i - f(X_i))^2}{\chi} - \chi\right),$$

   for every $f, \sigma, \tilde{f}, \chi \in F \times \mathbb{R}_+ \times F \times \mathbb{R}_+$, where $|B_k|$ denotes the cardinal of B_k ;

3. The global MOM criteria of (f, σ) against $(\tilde{f}, \chi)$ is defined by

$$MOM_K\left(\frac{l_f}{\sigma} + \sigma - \frac{l_{\tilde{f}}}{\chi} - \chi\right) := Median\left\{P_{B_k}\left(\frac{l_f}{\sigma} + \sigma - \frac{l_{\tilde{f}}}{\chi} - \chi\right),\ k = 1, \ldots, K\right\};$$

4. Finally, the MOM-K estimator of (β^*, σ^*) is defined by

$$\begin{aligned}(\hat{f}, \hat{\sigma}) &:= \underset{f \in F,\, \sigma > \sigma_{\min}}{\arg\min} \max_{g \in F,\, \chi > \sigma_{\min}} \left[MOM_K\left(\frac{l_f}{\sigma} + \sigma - \frac{l_g}{\chi} - \chi\right) + \mu(||f|| - ||g||)\right] \qquad (3.4)\\ &= \underset{f \in F,\, \sigma > \sigma_{\min}}{\arg\min} \max_{g \in F,\, \chi > \sigma_{\min}} T_{K,\mu}(g, \chi, f, \sigma) = \underset{f \in F,\, \sigma > \sigma_{\min}}{\arg\min}\ \mathcal{C}_{K,\mu}(f, \sigma),\end{aligned}$$

   where μ is a tuning parameter to be chosen, $||\cdot||$ is a norm on the space of functions generated by F,

$$T_{K,\mu}(g, \chi, f, \sigma) := MOM_K\left(\frac{l_f}{\sigma} + \sigma - \frac{l_g}{\chi} - \chi\right) + \mu(||f|| - ||g||)$$

   and $\mathcal{C}_{K,\mu}(f, \sigma) := \max_{g \in F,\, \chi > \sigma_{\min}} T_{K,\mu}(g, \chi, f, \sigma)$.

Note that this estimator $(\hat{f}, \hat{\sigma})$ depends only on two tuning parameters K and μ. To prove its properties, we will need the following assumption, which is a generalization of Assumption 3.2.1.

Assumption 3.3.1. *There exists θ_0, θ_m such that, for all $i \in \mathcal{I}$ and $f \in F$*

1. $Var\big[\zeta(f - f^*)(X)\big] \leq \theta_m^2 ||f - f^*||_{L_2}^2$,

2. $||f - f^*||_{L_2} \leq \theta_0 ||f - f^*||_{L_1}$.

The first part of this assumption specifies that the dependence between the noise ζ and transformations $f - f^*$ of X is not too strong compared to the norm of these transformations. It is satisfied as a special case when ζ and X are independent. The second assumption is related to the small ball assumption and controls the link between the L_1 and the L_2 norms of $f - f^*$.

Definition 3.2. *Let F be a class of functions $\mathcal{X} \to \mathbb{R}$. Let E be the vector space generated by F and $||\cdot||$ a norm on E, which will be used for regularization.*

1. *The subdifferential of* $||\cdot||$ *at any* $f \in F$ *is denoted by* $(\partial||\cdot||)_f := \{z^* \in E^* : ||f+h|| \geq ||f|| + z^*(h), \forall h \in E\}$*, where* $(E^*, ||\cdot||_*)$ *is the dual normed space of* $(E, ||\cdot||)$*.*

2. *For any* $\rho > 0$*, we set*

$$\begin{aligned} H_\rho &:= \{f \in F : ||f - f^*|| = \rho \text{ and } ||f - f^*||_{L_2} \leq r(\rho)\}, \\ \Gamma_{f^*} &:= \bigcup_{f \in f^* + (\rho/20)B} \big(\partial||\cdot||\big)_f, \\ \Delta(\rho) &:= \inf_{f \in H_\rho} \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*), \end{aligned}$$

3. *We will use the so-called sparsity inequality* $\Delta(\rho) \geq 4\rho/5$*, and we define* ρ^* *as the smallest* $\rho > 0$ *satisfying it.*

4. *The risk of a function* $f \in F$ *is defined by* $R(f) := \mathbb{E}\big[(Y - f(X))^2\big]^{1/2}$*.*

Definition 3.3. *Let* ϵ_i *be independent random variables uniformly distributed on* $\{-1, 1\}$*, and independent from the dataset* $\mathcal{D}$*. For all* $f \in F$*,* $r > 0$ *and* $\rho > 0$*, we denote* $B_{reg}(f, \rho, r) := \{g \in F : ||g - f||_{L_2} \leq r, ||g - f|| \leq \rho\}$*. For every* $\gamma_Q, \gamma_M > 0$*, we define two so-called Rademacher complexities by*

$$r_Q(\rho, \gamma_Q) = \inf\left\{r > 0 : \forall J \subset \mathcal{I}, |J| \geq \frac{N}{2}, \mathbb{E}\sup_{f \in B_{reg}(f^*, \rho, r)} \Big|\sum_{i \in J} \epsilon_i (f - f^*)(X_i)\Big| \leq \gamma_Q |J| r\right\},$$

$$r_M(\rho, \gamma_M) = \inf\left\{r > 0 : \forall J \subset \mathcal{I}, |J| \geq \frac{N}{2}, \mathbb{E}\sup_{f \in B_{reg}(f^*, \rho, r)} \Big|\sum_{i \in J} \epsilon_i \xi_i (f - f^*)(X_i)\Big| \leq \gamma_M |J| r^2\right\},$$

and let $r = r(\,\cdot\,, \gamma_M, \gamma_Q)$ *be a continuous non-decreasing function* $\mathbb{R}_+ \to \mathbb{R}_+$ *depending on* γ_Q, γ_M *such that for every* $\rho > 0$*,* $r(\rho, \gamma_Q, \gamma_M) > r_Q(\rho, \gamma_Q)$ *and* $r(\rho, \gamma_Q, \gamma_M) > r_M(\rho, \gamma_M)$*.*

Theorem 3.4. *Let Assumption 3.3.1 hold and assume that*

$$n \geq K(\kappa^* - 1)^{1/2}/400, \tag{3.5}$$

$$r^2(2\rho_K) \leq \min\left(2\tilde{\gamma}_1 \sigma^{*2}\,,\, 2c\theta_0^2 \sigma^{*2}\right), \tag{3.6}$$

$$(2 + \tilde{\gamma}_1)\tilde{\gamma}_1 \sigma^* \leq \frac{c'}{40\mu_0 c\theta_0} r^2(2\rho_K), \tag{3.7}$$

where $\tilde{\gamma}_1 := (K/n)^{1/2}(\kappa^* - 1)^{1/2}/10$*, with the choice* $\gamma_Q = 1/(720\theta_0)$*, and* $\gamma_M = \epsilon/360$*. Let us define* K^* *as the smallest integer satisfying* $K^* \geq n\epsilon r^2(\rho^*)/(284\theta_m^2)$ *and* $\epsilon := 1/(c\theta_0^2)$ *for a constant* $c > 0$*. Let* K *be an integer in* $\big[\max(K^*, 8|\mathcal{O}|)\,,\, n/(96(\theta_0\theta_m)^2)\big]$*. We define implicitly* ρ_K *as the solution of* $r^2(\rho_K) = (384\theta_m^2/\epsilon^2) \cdot (K/n)$*, where* $r(\cdot)$ *is defined in [91, Definition 5]. Let* $c', \mu_0 > 0$ *such that*

$$\mu_0 \leq 64\sigma^*(1 + \tilde{\gamma}_1)/5, \tag{3.8}$$

$$\mu_0 \leq 4c'\sigma_{\min}/5, \tag{3.9}$$

$$\mu_0 \geq \frac{16c'}{c - 32}\sigma^*\big(1 + \tilde{\gamma}_1/2\big) \tag{3.10}$$

$$\sigma^* \leq \big(1 + (17c'/20\mu_0)\big)\sigma_{\min} \tag{3.11}$$

$$c \geq 32\sigma^*(1 - \tilde{\gamma}_1)\left(\frac{2}{\sigma^* - r(\rho_K)} + \frac{2}{\sigma^* - \sigma^*\tilde{\gamma}_1} + \frac{4c'}{\mu_0}\right) \tag{3.12}$$

$$c' \geq \frac{10\mu_0}{\sigma_{\min}} - \frac{5\mu_0}{\sigma^*}. \tag{3.13}$$

Choosing the regularization parameter as $\mu := (c'\epsilon/\mu_0)r^2(\rho_K)/\rho_K$, *we have*

$$||\hat{f} - f^*|| \leq 2\rho_K, \tag{3.14}$$

$$||\hat{f} - f^*||_{L_2} \leq r(2\rho_K), \tag{3.15}$$

$$|\hat{\sigma} - \sigma^*| \leq \sigma^*\tilde{\gamma}_1, \tag{3.16}$$

$$\begin{aligned} R(\hat{f}) \leq R(f^*) + 2\Big(1 + \epsilon + \tilde{\gamma}_1\Big)r^2(2\rho_K) + \sigma^*(1+\tilde{\gamma}_1)(\sigma^*\tilde{\gamma}_1 + 2\mu\rho_K) \\ + (\sigma^* + r(2\rho_K))\left[2\big(1+\tilde{\gamma}_1\big)\sigma^*\tilde{\gamma}_1 + \left(\frac{2\epsilon}{\sigma^* - r(\kappa\rho_K)} + \frac{c'\epsilon\kappa}{\mu_0}\right)r^2(\rho_K)\right]. \end{aligned} \tag{3.17}$$

with probability $1 - c_1\exp(-K/c_2)$, *where* c_1 *and* c_2 *are universal constants.*

This theorem allows us to bound in a general framework statistical errors on an event that holds with high probability. Indeed, on this event of probability $1 - c_1\exp(-K/c_2)$, we control the distance of the estimators $\hat{f}$ to its true value f^* in the regularization norm $||\cdot||$ and in the $||\cdot||_{L_2}$ norm. We also control the distance between the estimated standard deviation $\hat{\sigma}$ and the true standard deviation σ^*. Finally, we control the risk of $\hat{f}$, meaning that it is smaller than the risk associated with the true function f^*, plus a residual term.

Note that, even if many parameters appear in the statement of Theorem 3.4 above, the estimator still depends only on two parameters, that are μ and K. If μ can be chosen as in the statement of this theorem, as well as other parameters, the theorem apply, and gives bounds on our estimators with high probability. If the condition of Theorem 3.4 could not be satisfied with any choice of c', c, μ_0 for a given μ, then the joint estimator $(\hat{f}, \hat{\sigma})$ can still be computed, but without any theoretical guarantee on its performance.

3.4 Technical lemmas

The following lemma is adapted from from Lecué and Lerasle [91, Equations (16), (18), and (19)].

Lemma 3.5. *There exists an event* $\Omega_1(K)$ *of probability bigger than* $1 - 4\exp(-K/4320)$ *such that, for all* $\rho \in \{\kappa\rho_K : \kappa \in \{1,2\}\}$, *and all* $f \in F$ *such that* $||f - f^*|| \leq \rho$, *we have*

1. *If* $||f - f^*||_{L_2} \geq r_Q(\rho, \gamma_Q)$, *then* $Q_{1/4,K}\big((f-f^*)^2\big) \geq Q_{1/8,K}\big((f-f^*)^2\big) \geq \frac{1}{(4\theta_0)^2}||f - f^*||_{L_2}^2$,
2. $Q_{3/4,K}\big(2\zeta(f-f^*)\big) \leq Q_{7/8,K}\big(2\zeta(f-f^*)\big) \leq \alpha_1$
3. $P[-2\zeta(f-f^*)] \leq \min\big(Q_{1/8,K}[-2\zeta(f-f^*)]\,,\, Q_{1/4,K}[-2\zeta(f-f^*)]\big) + \alpha_1$

where $\alpha_1 := 2\epsilon \cdot \max\left(r_M^2(\rho,\gamma_M)\,,\, \frac{720\theta_m^2}{\epsilon^2}\frac{K}{n}\,,\, ||f-f^*||_{L_2}\right)$ *and* $\epsilon := 1/(c\theta_0^2)$.

Proof of Lemma 3.5: The results are proved by following the same steps as in the proof of Equations (16), (18), and (19) in Lecué and Lerasle [91], choosing $\eta = 1/8$, $\gamma = 15/16$, $\alpha = x = 1/45$, $\gamma_Q = 1/(720\theta_0)$, and $\gamma_M = \epsilon/360$.

□

The following lemma is proved in [91, Lemma 4]. We reproduce it here for convenience.

Lemma 3.6. *Let* $\rho \geq 0$, $\Gamma_{f^*}(\rho) := \bigcup_{f \in f^* + (\rho/20)B} \left(\partial||\cdot||\right)_f$. *For all* $g \in F$,

$$||f^*|| - ||g|| \leq \frac{\rho}{10} - \sup_{z^* \in \Gamma_{f^*}(\rho)} z^*(g - f^*).$$

Lemma 3.7. *Let* $\rho > 0$, *for all* $f \in F$ *such that* $||f - f^*|| \leq \rho$ *and* $|\sigma - \sigma^*| \leq \alpha_\sigma$, *we have*

$$P[-2\zeta(f - f^*)] \leq (\sigma^* + \alpha_\sigma) T_{K,\lambda}(f^*, \sigma^*, \hat{f}, \hat{\sigma}) + \alpha_1 + (\sigma^{*2} + \gamma_1)\frac{\alpha_\sigma}{\sigma^{*2}} + (\sigma^* + \alpha_\sigma)(\alpha_\sigma + \mu\rho).$$

where $\alpha_1 := 2\epsilon \cdot \max\left(r_M^2(\rho, \gamma_M)\,,\, \frac{720\theta_m^2}{\epsilon^2}\frac{K}{n}\,,\, ||f - f^*||_{L_2}\right)$.

Proof: We apply part 3 of Lemma 3.5 and get

$$\begin{aligned}
P[-2\zeta(f - f^*)] &\leq Q_{1/4,K}[-2\zeta(f - f^*)] + \alpha_1 \\
&\leq Q_{1/4,K}[(f - f^*)^2 - 2\zeta(f - f^*)] + \alpha_1 \\
&\leq \sigma Q_{1/4,K}\left[\frac{1}{\sigma}(l_f - l_{f^*})\right] + \alpha_1 \\
&\leq \sigma Q_{1/4,K}\left[\frac{l_f}{\sigma} + (\sigma - \sigma^*) - \frac{l_{f^*}}{\sigma^*} + l_{f^*}\left(\frac{1}{\sigma^*} - \frac{1}{\sigma}\right)\right] + \alpha_1 + \sigma\alpha_\sigma \\
&\leq \sigma Q_{1/2,K}\left[\frac{l_f}{\sigma} + (\sigma - \sigma^*) - \frac{l_{f^*}}{\sigma^*}\right] - Q_{1/4,K}\left[-l_{f^*}\left(\frac{1}{\sigma^*} - \frac{1}{\sigma}\right)\right] + \sigma\mu(||f|| - ||f^*||) + \alpha_1 + \sigma(\alpha_\sigma + \mu\rho) \\
&\leq \sigma T_{K,\lambda}(f^*, \sigma^*, \hat{f}, \hat{\sigma}) - Q_{1/4,K}[-\zeta^2]\left(\frac{1}{\sigma^*} - \frac{1}{\sigma}\right) + \alpha_1 + \sigma(\alpha_\sigma + \mu\rho).
\end{aligned}$$

We apply now Lemma 3.9 and the mean value theorem to the function $x \mapsto 1/x$, and we get

$$\begin{aligned}
P[-2\zeta(f - f^*)] &\leq \sigma T_{K,\lambda}(f^*, \sigma^*, \hat{f}, \hat{\sigma}) + \alpha_1 + (\sigma^{*2} + \gamma_1)\frac{\alpha_\sigma}{\sigma^{*2}} + \sigma(\alpha_\sigma + \mu\rho) \\
&\leq (\sigma^* + \alpha_\sigma) T_{K,\lambda}(f^*, \sigma^*, \hat{f}, \hat{\sigma}) + \alpha_1 + (\sigma^{*2} + \gamma_1)\frac{\alpha_\sigma}{\sigma^{*2}} + (\sigma^* + \alpha_\sigma)(\alpha_\sigma + \mu\rho).
\end{aligned}$$

□

In the following, we will use the notation $\mathcal{K} := \{k \in \{1, \ldots, K\} : B_k \subset \mathcal{I}\}$.

Lemma 3.8. *Let* $\gamma, \gamma_1, \eta, x > 0$ *such that* $\gamma(1 - KVar[Z]/(n\gamma_1^2) - x) \geq 1 - \eta$. *Let* $K \in [|\mathcal{O}|/(1-\gamma), n]$. *Let* Z *be a real-valued random variable. There exists an event* $\Omega(Z, K)$ *with probability greater than* $1 - \exp(-\gamma x^2 K/2)$ *such that, on the event* $\Omega(Z, K)$

$$\left|\{k \in [K] : |P_{B_k}(Z) - \mathbb{E}Z| \leq \gamma_1\}\right| \geq K(1 - \eta).$$

Proof : We have

$$\begin{aligned}
&|\{k \in [K] : |P_{B_k}(Z) - \mathbb{E}Z| \leq \gamma_1\}| \geq \sum_{k \in \mathcal{K}} \mathbb{1}\{|P_{B_k}(Z) - \mathbb{E}Z| \leq \gamma_1\} \\
&= |\mathcal{K}| - \sum_{k \in \mathcal{K}} \mathbb{P}\{|P_{B_k}(Z) - \mathbb{E}Z| \geq \gamma_1\} - \sum_{k \in \mathcal{K}} \Big(\mathbb{1}\{|P_{B_k}(Z) - \mathbb{E}Z| \geq \gamma_1\} - \mathbb{P}\{|P_{B_k}(Z) - \mathbb{E}Z| \geq \gamma_1\}\Big).
\end{aligned}$$

We bound the first term using Chebychev's inequality

$$\sum_{k \in \mathcal{K}} \mathbb{P}\{|P_{B_k}(Z) - \mathbb{E}Z| \geq \gamma_1\} \leq |\mathcal{K}|\frac{Var[P_{B_k}(Z) - \mathbb{E}Z]}{\gamma_1^2} = |\mathcal{K}|\frac{Var[Z]}{|B_k|\gamma_1^2} = |\mathcal{K}|\frac{KVar[Z]}{n\gamma_1^2}.$$

We bound the other term using Hoeffding's inequality

$$\sum_{k\in\mathcal{K}} \Big(\mathbb{1}\{|P_{B_k}(Z) - \mathbb{E}Z| \geq \gamma_1\} - \mathbb{P}\{|P_{B_k}(Z) - \mathbb{E}Z| \geq \gamma_1\} \Big) \leq x|\mathcal{K}|,$$

on an event $\Omega(Z,K)$ of probability greater than $1-\exp(-x^2|\mathcal{K}|/2)$. Combining the previous inequalities, we get that on $\Omega(Z,K)$,

$$|\{k\in\mathcal{K} : |P_{B_k}(Z) - \mathbb{E}Z| \leq \gamma_1\}| \geq |\mathcal{K}| \left(1 - \frac{KVar[Z]}{n\gamma_1^2} - x\right) \geq K\gamma \left(1 - \frac{KVar[Z]}{n\gamma_1^2} - x\right).$$

□

Lemma 3.9 (Bounding the quantiles of ζ^2)**.** *Assume that $K \leq n\alpha_3$. On an event $\Omega_2(K)$ of probability greater than $1-\exp(-5K/13824)$, we have $Q_{1/4,K}[\zeta^2] \leq \sigma^{*2} + \gamma_1$ and $-\sigma^{*2} - \gamma_1 \leq Q_{1/4,K}[-\zeta^2] \leq -\sigma^{*2} + \gamma_1$, where $\gamma_1 = \alpha_3^{1/2}\sigma^{*2}(\kappa^*-1)^{1/2}/5$. The same inequalities are also valid with $1/4$ replaced by $1/8$, on the same event. As a consequence, this is valid for the choice $\alpha_3 := K/n$.*

Proof of Lemma 3.9: We apply Lemma 3.8 with $Z := \zeta^2$, $Var[Z] = \mathbb{E}[\zeta^4] - \mathbb{E}[\zeta^2]^2 = \sigma^{*4}(\kappa^*-1)$, $\eta = 1/8$, $\gamma = 15/16$, $x = 1/36$, $\gamma_1 = 6\big(\alpha_3 Var[Z]\big)^{1/2}$, so that $\gamma(1 - KVar[Z]/(n\gamma_1^2) - x) \geq 1-\eta$ with probability $1-\exp(-\gamma x^2 K/2) = 1-\exp(-5K/13824)$. Therefore, on the same event, all $Q_{1/8,K}[\zeta^2]$, $Q_{1/4,K}[\zeta^2]$, $Q_{3/4,K}[\zeta^2]$ and $Q_{7/8,K}[\zeta^2]$ belongs to the interval $\big[\mathbb{E}\zeta^2 - \gamma_1\,,\, \mathbb{E}\zeta^2 + \gamma_1\big]$. We get also that $Q_{1/8,K}[-\zeta^2], Q_{1/4,K}[-\zeta^2] \in \big[-\mathbb{E}\zeta^2 + \gamma_1\,,\, -\mathbb{E}\zeta^2 + \gamma_1\big]$.

□

Lemma 3.10. *Let $\kappa \in \{1,2\}$ and $\alpha_{2,\kappa} = (K/n)^{1/2}\sigma^*(\kappa^*-1)^{1/2}/10$. Under the assumptions of Theorem 3.4, we have*

1. $\gamma_1 \leq 2\sigma^*\alpha_{2,\kappa} + \alpha_{2,\kappa}^2$, *therefore* $\sqrt{\sigma^{*2}+\gamma_1} \leq \sigma^* + \alpha_{2,\kappa}$;
2. $\frac{2\epsilon - (4\theta_0)^{-2}}{\sigma^* + \alpha_{2,\kappa}} + \frac{c'\epsilon}{\mu_0} \leq 0$;
3. ${\alpha_{2,\kappa}}^2 \leq \gamma_1 + 2\alpha_{2,\kappa}\sigma^* - r^2(\kappa\rho_K)$, *therefore* $\sigma^* - \alpha_{2,\kappa} \leq \sqrt{\sigma^{*2} + \gamma_1 - r^2(\kappa\rho_K)}$;
4. $\frac{2\epsilon r^2(\rho_K) - \gamma_1}{\sigma^{*2}} \leq 4 + \frac{\gamma_1}{\sigma^{*2}}$;
5. $\frac{2\epsilon - (4\theta_0)^{-2}}{\sigma^* + \alpha_{2,1}} + \frac{11c'\epsilon}{10\mu_0} \leq \frac{2\epsilon}{\sigma^* - r(\rho_K)} + \frac{c'\epsilon}{\mu_0}$

where $\gamma_1 := (K/n)^{1/2}\sigma^{*2}(\kappa^*-1)^{1/2}/5$

Proof of Lemma 3.10: 1. By construction, we have $\gamma_1 = 2\sigma^*\alpha_{2,\kappa}$. Therefore, the claimed inequality is satisfied.

2. Because we have assumed that Equation (3.10) holds, we have

$$\mu_0 \geq \frac{16c'}{c-32}\sigma^*\big(1 + (K/n)^{1/2}(\kappa^*-1)^{1/2}/10\big).$$

This is equivalent to

$$\mu_0\left(\frac{2}{c} - \frac{1}{16}\right) + \frac{c'}{c}\sigma^*\big(1 + (K/n)^{1/2}(\kappa^*-1)^{1/2}/10\big) \leq 0,$$

which can be rewritten as

$$\frac{2/(c\theta_0^2) - 1/(4\theta_0)^2}{\sigma^*\big(1 + (K/n)^{1/2}(\kappa^*-1)^{1/2}/10\big)} + \frac{c'/(c\theta_0)}{\mu_0} \leq 0.$$

3. By Equation (3.5), we derive that $\alpha_{2,\kappa} \leq 2\sigma^*$, which means that ${\alpha_{2,\kappa}}^2 \leq 2\alpha_{2,\kappa}\sigma^*$. Using Equation (3.6), we get $r^2(\kappa\rho_K) \leq 2\alpha_{2,\kappa}\sigma^* = \gamma_1$. Combining the two equations, we get ${\alpha_{2,\kappa}}^2 + r^2(\kappa\rho_K) \leq 2\alpha_{2,\kappa}\sigma^* + \gamma_1$.

4. Using Equation (3.6), we get $r^2(\rho_K) \leq 2c\theta^2{\sigma^*}^2$. Therefore $2\epsilon r^2(\rho_K)/{\sigma^*}^2 \leq 4$, which implies the claimed result.

5. We have

$$\frac{2\epsilon - (4\theta_0)^{-2}}{\sigma^* + \alpha_{2,1}} + \frac{11c'\epsilon}{10\mu_0} \leq \frac{c'\epsilon}{10\mu_0} \leq \frac{2\epsilon}{\sigma^* - r(\rho_K)} + \frac{c'\epsilon}{\mu_0},$$

where the first inequality is a consequence of part 4 of this lemma.

□

3.5 Control of the supremum of $T_{K,\mu}(g, \chi, f^*, \sigma^*)$ on each $F_i^{(\kappa)}$

3.5.1 Preliminaries

In this section, we will assume to be on the event $\Omega(K) := \Omega_1(K) \cap \Omega_2(K)$, where $\Omega_1(K)$ is defined in Lemma 3.5 and $\Omega_2(K)$ is defined in Lemma 3.9. For $\kappa \in \{1, 2\}$, and any fixed $\alpha_{2,\kappa} > 0$, let us define

$$\begin{aligned}
F_1^{(\kappa)} &:= \{(g,\chi) \in F \times \mathbb{R}_+^* : ||g - f^*|| \leq \kappa\rho_K, ||g - f^*||_{L_2} \leq r(\kappa\rho_K) \text{ and } |\sigma^* - \chi| \leq \alpha_{2,\kappa}\} \\
F_2^{(\kappa)} &:= \{(g,\chi) \in F \times \mathbb{R}_+^* : ||g - f^*|| \leq \kappa\rho_K, ||g - f^*||_{L_2} > r(\kappa\rho_K) \text{ and } |\sigma^* - \chi| \leq \alpha_{2,\kappa}\} \\
F_3^{(\kappa)} &:= \{(g,\chi) \in F \times \mathbb{R}_+^* : ||g - f^*|| > \kappa\rho_K \text{ and } |\sigma^* - \chi| \leq \alpha_{2,\kappa}\} \\
F_4^{(\kappa)} &:= \{(g,\chi) \in F \times \mathbb{R}_+^* : ||g - f^*|| \leq \kappa\rho_K, ||g - f^*||_{L_2} \leq r(\kappa\rho_K) \text{ and } \chi > \sigma^* + \alpha_{2,\kappa}\} \\
F_5^{(\kappa)} &:= \{(g,\chi) \in F \times \mathbb{R}_+^* : ||g - f^*|| \leq \kappa\rho_K, ||g - f^*||_{L_2} > r(\kappa\rho_K) \text{ and } \chi > \sigma^* + \alpha_{2,\kappa}\} \\
F_6^{(\kappa)} &:= \{(g,\chi) \in F \times \mathbb{R}_+^* : ||g - f^*|| > \kappa\rho_K \text{ and } \chi > \sigma^* + \alpha_{2,\kappa}\} \\
F_7^{(\kappa)} &:= \{(g,\chi) \in F \times \mathbb{R}_+^* : ||g - f^*|| \leq \kappa\rho_K, ||g - f^*||_{L_2} \leq r(\kappa\rho_K) \text{ and } \chi < \sigma^* - \alpha_{2,\kappa}\} \\
F_8^{(\kappa)} &:= \{(g,\chi) \in F \times \mathbb{R}_+^* : ||g - f^*|| \leq \kappa\rho_K, ||g - f^*||_{L_2} > r(\kappa\rho_K) \text{ and } \chi < \sigma^* - \alpha_{2,\kappa}\} \\
F_9^{(\kappa)} &:= \{(g,\chi) \in F \times \mathbb{R}_+^* : ||g - f^*|| > \kappa\rho_K \text{ and } \chi < \sigma^* - \alpha_{2,\kappa}\}.
\end{aligned}$$

Lemma 3.11. *On the event $\Omega(K)$, it holds for all $\kappa \in \{1, 2\}$ that*

$$\sup_{(g,\chi)\in F_1^{(\kappa)}} T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq \left(2 + \frac{\gamma_1}{{\sigma^*}^2}\right)\alpha_{2,\kappa} + \left(\frac{2\epsilon}{\sigma^* - r(\kappa\rho_K)} + \frac{c'\epsilon\kappa}{\mu_0}\right) r^2(\rho_K) := B_{1,\kappa}$$

$$\sup_{(g,\chi)\in F_2^{(\kappa)}} T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq \left(2 + \frac{\gamma_1}{{\sigma^*}^2}\right)\alpha_{2,\kappa} + \left(\frac{2\epsilon - (4\theta_0)^{-2}}{\sigma^* - \alpha_{2,\kappa}} + \frac{c'\epsilon\kappa}{\mu_0}\right) r^2(\rho_K) := B_{2,\kappa}$$

$$\begin{aligned}
\sup_{(g,\chi)\in F_3^{(\kappa)}} T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq \max\Bigg(&\left(\left(2 + \frac{\gamma_1}{{\sigma^*}^2}\right)\alpha_{2,\kappa} + \kappa\epsilon\Big(\frac{2}{\sigma^* - \alpha_{2,\kappa}} - \frac{7c'}{10\mu_0}\Big) r^2(\rho_K)\right), \\
&\left(2 + \frac{\gamma_1}{{\sigma^*}^2}\right)\alpha_{2,\kappa} + \kappa\left(\frac{2\epsilon - (4\theta_0)^{-2}}{\sigma^* + \alpha_{2,\kappa}} + \frac{11c'\epsilon}{10\mu_0}\right) r^2(\rho_K)\Bigg) := B_{3,\kappa}
\end{aligned}$$

$$\sup_{(g,\chi)\in F_4^{(\kappa)}} T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq \frac{\gamma_1}{{\sigma^*}^2}\alpha_{2,\kappa} + \left(\frac{2\epsilon}{\sigma^* + \alpha_{2,\kappa}} + \frac{c'\epsilon\kappa}{\mu_0}\right) r^2(\rho_K) := B_{4,\kappa}$$

$$\sup_{(g,\chi)\in F_5^{(\kappa)}} T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq \frac{\gamma_1}{{\sigma^*}^2}\alpha_{2,\kappa} + \frac{c'\epsilon\kappa}{\mu_0} r^2(\rho_K) := B_{5,\kappa}$$

$$\sup_{(g,\chi)\in F_6^{(\kappa)}} T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq \max\left(\frac{\gamma_1}{{\sigma^*}^2}\alpha_{2\kappa} + \kappa\epsilon\Big(\frac{2}{\sigma^* + \alpha_{2,\kappa}} - \frac{7c'}{10\mu_0}\Big) r^2(\rho_K)\right),$$

$$\left(\frac{\gamma_1}{\sigma^{*2}}-1\right)\alpha_{2,\kappa}+\kappa\left(\frac{2\epsilon-(4\theta_0)^{-2})}{\sigma+\alpha_{2,\kappa}}+\frac{11c'\epsilon}{10\mu_0}\right)r^2(\rho_K)\Bigg) := B_{6,\kappa}$$

$$\sup_{(g,\chi)\in F_7^{(\kappa)}} T_{K,\mu}(g,\chi,f^*,\sigma^*) \le \left(\frac{2\epsilon r^2(\kappa\rho_K)-\gamma_1}{\sigma^{*2}}-2\right)\alpha_{2,\kappa}+\left(\frac{2\epsilon}{\sigma^*}+\frac{c'\epsilon\kappa}{\mu_0}\right)r^2(\rho_K) := B_{7,\kappa}$$

$$\sup_{(g,\chi)\in F_8^{(\kappa)}} T_{K,\mu}(g,\chi,f^*,\sigma^*) \le \left(2+\frac{\gamma_1}{\sigma^{*2}}\right)\alpha_{2,\kappa}+\left(\frac{(2\epsilon-(4\theta_0)^{-2})}{\sigma^*-\alpha_{2,\kappa}}+\frac{c'\epsilon\kappa}{\mu_0}\right)r^2(\rho_K) := B_{8,\kappa}$$

$$\sup_{(g,\chi)\in F_9^{(\kappa)}} T_{K,\mu}(g,\chi,f^*,\sigma^*) \le \max\Bigg(\frac{\gamma_1}{\sigma^{*2}}\alpha_{2,\kappa}+\left(\frac{2\epsilon}{\sigma_{\min}}-\frac{7c'\epsilon}{10\mu_0}\right)\kappa r^2(\rho_K),$$

$$\left(2+\frac{\gamma_1}{\sigma^{*2}}\right)\alpha_{2,\kappa}+\kappa\left(\frac{2\epsilon-(4\theta_0)^{-2})}{\sigma-\alpha_{2,\kappa}}+\frac{11c'\epsilon}{10\mu_0}\right)r^2(\rho_K)\Bigg) := B_{9,\kappa}.$$

Each bound is respectively proved in each of the subsections of Section 3.6. The following lemma gives a comparison between all these bounds.

Lemma 3.12. *We have* $B_{1,1}=\max_{i=1,\dots,9}B_{i,1}$ *and* $-B_{1,1}>\max_{i=2,\dots,9}B_{i,2}$.

3.5.2 Proof of the first assertion of Lemma 3.12

In this section, we show that $B_{1,1}$ is bigger than the other $B_{i,1}$, $i=2,\dots,9$.

Case $i=2$**:** using $\alpha_{2,1}\le r(1\times\rho_K)$, we get

$$\frac{2\epsilon-(4\theta_0)^2}{\sigma^*-\alpha_{2,1}}\le\frac{2\epsilon}{\sigma^*-\alpha_{2,1}}\le\frac{2\epsilon}{\sigma^*-r(1\times\rho_K)},$$

therefore $B_{2,1}\le B_{1,1}$.

Case $i=3$**:**

$$\epsilon\Big(\frac{2}{\sigma^*-\alpha_{2,1}}-\frac{7c'}{10\mu_0}\Big)\le\frac{2\epsilon}{\sigma^*-r(1\times\rho_K)}-\frac{7c'\epsilon}{10\mu_0}\le\frac{2\epsilon}{\sigma^*-r(1\times\rho_K)}+\frac{c'\epsilon\times1}{\mu_0},$$

and, applying part 5 of Lemma 3.10, $\frac{2\epsilon-(4\theta_0)^{-2}}{\sigma^*+\alpha_{2,1}}+\frac{11c'\epsilon}{10\mu_0}\le\frac{2\epsilon}{\sigma^*-r(1\times\rho_K)}+\frac{c'\epsilon}{\mu_0}$, therefore $B_{3,1}\le B_{1,1}$.

Cases $i=4,5,8$**:** Each of the two terms in the definition of $B_{i,1}$ is smaller than the corresponding term in the definition of $B_{1,1}$, therefore $B_{i,1}\le B_{1,1}$.

Case $i=6$**:** Same as for $i=3$.

Case $i=7$**:** The result in this case follows because $2+\frac{\gamma_1}{\sigma^{*2}}\ge\frac{2\epsilon r^2(\rho_K)-\gamma_1}{\sigma^{*2}}-2$, by part 4 of Lemma 3.10.

Case $i=9$**:** $B_{9,1}\le B_{1,1}$ if

$$\frac{2\epsilon}{\sigma_{\min}}-\frac{7c'\epsilon}{10\mu_0}\le\frac{2\epsilon}{\sigma^*-r(1\times\rho_K)}+\frac{c'\epsilon\times1}{\mu_0},$$

which is equivalent to

$$\frac{1}{\sigma_{\min}}\le\frac{1}{\sigma^*-r(\rho_K)}+\frac{17c'}{20\mu_0},$$

i.e. $\sigma^*-r(\kappa\rho_K)\le\big(1+(17c'/20\mu_0)\big)\sigma_{\min}$, which is true as we have assumed Equation (3.11).

□

3.5.3 Proof of the second assertion of Lemma 3.12

In this section, we show that $-B_{1,1}$ is bigger than $B_{i,2}$, for all $i = 2, \dots, 9$. We only give a sketch of the proof here since it results from elementary computations. Because of Equation (3.7) we can control all the terms of the form $\alpha_{2,\kappa}$ by terms of the form $r^2(\rho_K)$. Therefore, we only compare such terms. All inequalities are similar and consequences of Equation (3.12), using the definition $\epsilon = 1/(c\theta_0^2)$. For the term $i = 9$, the inequality is a consequence of Equation (3.13).

□

3.6 Proof of Lemma 3.11

Let us remark that

$$\begin{aligned}
T_{K,\mu}(f,\sigma,f^*,\sigma^*) &:= MOM_K\Big(\frac{l_{f^*}}{\sigma^*} + \sigma^* - \frac{l_f}{\sigma} - \sigma\Big) + \mu(||f^*|| - ||f||) \\
&= MOM_K\Big(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\sigma}\Big) + (\sigma^* - \sigma) + \frac{1}{\sigma}(l_{f^*} - l_f)\Big) + \mu(||f^*|| - ||f||) \\
&= MOM_K\Big(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\sigma}\Big) + (\sigma^* - \sigma) + \frac{1}{\sigma}\Big(2\zeta(f - f^*) - (f - f^*)^2\Big)\Big) + \mu(||f^*|| - ||f||).
\end{aligned}$$

This decomposition will be a key component of the proofs below.

3.6.1 Bound on $F_1^{(\kappa)}$

Let $g \in F_1^{(\kappa)}$. Using the inequality $(g - f^*)^2 \geq 0$ and the triangular inequality, we get

$$\begin{aligned}
T_{K,\mu}(g,\chi,f^*,\sigma^*) &= MOM_K\Big(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + (\sigma^* - \chi) + \frac{1}{\chi}\Big(2\zeta(g - f^*) - (g - f^*)^2\Big)\Big) + \mu(||f^*|| - ||g||) \\
&\leq \alpha_{2,\kappa} + MOM_K\Big(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + \frac{2}{\chi}\zeta(g - f^*)\Big)\Big) + \mu||f^* - f|| \\
&\leq \alpha_{2,\kappa} + \frac{2}{\chi}Q_{3/4,K}\big[\zeta(g - f^*)\big] - Q_{1/4,K}[\zeta^2]\Big(\frac{1}{\chi} - \frac{1}{\sigma^*}\Big) + \mu\kappa\rho_K \\
&\leq \alpha_{2,\kappa} + \frac{1}{\sigma^* - r(\kappa\rho_K)}Q_{3/4,K}\big[2\zeta(g - f^*)\big] + (\sigma^{*2} + \gamma_1)\frac{\alpha_{2,\kappa}}{\sigma^{*2}} + \mu\kappa\rho_K.
\end{aligned}$$

where in the last line, we used Lemma 3.9. By the part 2 of Lemma 3.5, using the fact that $\alpha_1 \leq 2\epsilon r^2(\kappa\rho_K)$, and plugging our choice of μ, we get

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq \Big(1 + \frac{\sigma^{*2} + \gamma_1}{\sigma^{*2}}\Big)\alpha_{2,\kappa} + \Big(\frac{2\epsilon}{\sigma^* - r(\kappa\rho_K)} + \frac{c'\epsilon\kappa}{\mu_0}\Big)r^2(\rho_K).$$

□

3.6.2 Bound on $F_2^{(\kappa)}$

Let $(g,\chi) \in F_2^{(\kappa)}$. We have

$$\begin{aligned}
T_{K,\mu}(g,\chi,f^*,\sigma^*) &= MOM_K\Big(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + (\sigma^* - \chi) + \frac{1}{\chi}\Big(2\zeta(g - f^*) - (g - f^*)^2\Big)\Big) + \mu(||f^*|| - ||g||) \\
&\leq (\sigma^* - \chi) + Q_{3/4,K}\Big(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + \frac{1}{\chi}\Big(2\zeta(g - f^*)\Big)\Big) - \frac{1}{\chi}Q_{1/4}[(g - f^*)^2] + \mu(||f^* - g||) \\
&\leq \alpha_{2,\kappa} + Q_{3/4,K}\Big(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + \frac{1}{\chi}\Big(2\zeta(g - f^*)\Big)\Big) - \frac{1}{\chi}Q_{1/4}[(g - f^*)^2] + \mu\kappa\rho_K
\end{aligned}$$

$$\leq \alpha_{2,\kappa} + \frac{1}{\chi} Q_{7/8,K}\Big[2\zeta(g-f^*)\Big] - Q_{1/8}[-\zeta^2]\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) - \frac{1}{\chi} Q_{1/4}[(g-f^*)^2] + \mu\kappa\rho_K.$$

On the right hand-side of the last equation, we will bound the second term using part 1 of Lemma 3.5, the third term using Lemma 3.9 and the fourth term using part 2 of Lemma 3.5. Finally, we replace μ and ρ_K by their values so that we get

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq \alpha_{2,\kappa} + \frac{\alpha_1 - ||f^*-g||^2_{L^2_P}(4\theta_0)^{-2}}{\chi} + \frac{\alpha_{2,\kappa}(\sigma^{*2}+\gamma_1)}{\sigma^{*2}} + \frac{c'\epsilon\kappa}{\mu_0} r^2(\rho_K).$$

We have $2\epsilon < (4\theta_0)^{-2}$, therefore $\alpha_1 - ||f^*-g||^2_{L^2_P}(4\theta_0)^{-2} \leq (2\epsilon - (4\theta_0)^{-2}) r^2(\rho_K)$ and we can deduce that

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq \left(1 + \frac{\sigma^{*2}+\gamma_1}{\sigma^{*2}}\right)\alpha_{2,\kappa} + \left(\frac{2\epsilon - (4\theta_0)^{-2}}{\sigma^* - \alpha_{2,\kappa}} + \frac{c'\epsilon\kappa}{\mu_0}\right) r^2(\rho_K).$$

□

3.6.3 Bound on $F_3^{(\kappa)}$

Let $(g,\chi) \in F_3^{(\kappa)}$. We have

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) &= MOM_K\left(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + (\sigma^* - \chi) + \frac{1}{\chi}\Big(2\zeta(g-f^*) - (g-f^*)^2\Big)\right) + \mu(||f^*|| - ||g||) \\ &\leq (\sigma^* - \chi) + MOM_K\left(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + \frac{1}{\chi}\Big(2\zeta(g-f^*) - (g-f^*)^2\Big)\right) - \mu \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g-f^*) + \frac{\mu\rho_K}{10}, \end{aligned} \tag{3.18}$$

where the last line results from the application of Lemma 3.6 with $\rho = \rho_K$. We follow now the proof of Lemma 5 in [91]. Let us define $f := f^* + \rho_K(g-f^*)/||g-f^*||$. By convexity of F, we get $f \in F$. Let $\Upsilon := ||g-f^*||/\rho_K$. Noting that $||f-f^*|| = \rho_K$, and $g - f^* = \Upsilon(f-f^*)$, we have

$$\begin{aligned} &MOM_K\left(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + \frac{1}{\chi}\Big(2\zeta(g-f^*) - (g-f^*)^2\Big)\right) - \mu \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(g-f^*) \\ &\quad = MOM_K\left(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + \frac{1}{\chi}\Big(2\Upsilon\zeta(f-f^*) - \Upsilon^2(f-f^*)^2\Big)\right) - \mu\Upsilon \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f-f^*) \\ &\quad \leq MOM_K\left(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + \frac{1}{\chi}\Big(2\Upsilon\zeta(f-f^*) - \Upsilon(f-f^*)^2\Big)\right) - \mu\Upsilon \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f-f^*), \end{aligned}$$

where in the last line, we used the fact that $\Upsilon \geq 1$. Therefore, combining the previous equation with Equation (3.18), we get

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq (\sigma^* - \chi) &+ MOM_K\left(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + \frac{\Upsilon}{\chi}\Big(2\zeta(f-f^*) - (f-f^*)^2\Big)\right) \\ &- \mu\Upsilon \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f-f^*) + \frac{\mu\kappa\rho_K}{10}, \end{aligned} \tag{3.19}$$

3.6.3.1 Case $||f-f^*||_{L^2_P} \leq r(\rho_K)$

Remembering that $||f-f^*|| = \rho_K$, we can deduce that $f \in H_{\rho_K}$ Using the definition of K^*, the fact that $K \geq K^*$, we get that $\rho_K \geq \rho$, and therefore ρ_K follows the sparsity equation, from which we derive $\sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f-f^*) \geq \Delta(\rho_K) \geq 4\rho_K/5$. Using our choice of μ, we get

$$-\mu \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f-f^*) \leq -\frac{4c'\epsilon}{5\mu_0} r^2(\rho_K).$$

Combining this with Equation (3.19) and the fact that $(f-f^*)^2 \geq 0$, we get that

$$\begin{aligned}
T_{K,\mu}(g,\chi,f^*,\sigma^*) &\leq (\sigma^*-\chi) + MOM_K\Big(l_{f^*}\Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big) + \frac{\Upsilon}{\chi}\Big(2\zeta(f-f^*)\Big)\Big) - \Upsilon\frac{4c'\epsilon}{5\mu_0}r^2(\rho_K) + \frac{c'\epsilon\kappa}{10\mu_0}r^2(\rho_K) \\
&\leq \alpha_{2,\kappa} + \frac{\Upsilon}{\chi}Q_{3/4,K}\big[2\zeta(f-f^*)\big] - \Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big)Q_{1/4,K}\big[-\zeta^2\big] - \Upsilon\frac{4c'\epsilon}{5\mu_0}r^2(\rho_K) + \frac{c'\epsilon\kappa}{10\mu_0}r^2(\rho_K) \\
&\leq \Big(1+\frac{\sigma^{*2}+\gamma_1}{\sigma^{*2}}\Big)\alpha_{2,\kappa} + \Upsilon\Big(\frac{1}{\chi}Q_{3/4,K}\big[2\zeta(f-f^*)\big] - \frac{4c'\epsilon}{5\mu_0}r^2(\rho_K)\Big) + \frac{c'\epsilon\kappa}{10\mu_0}r^2(\rho_K).
\end{aligned}$$

Applying part 2 of Lemma 3.5, we get $Q_{3/4,K}\big[2\zeta(f-f^*)\big] \leq \alpha_1 \leq 2\epsilon r^2(\rho_K)$. Therefore,

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq \Big(2+\frac{\gamma_1}{\sigma^{*2}}\Big)\alpha_{2,\kappa} + \Upsilon\Big(\frac{2\epsilon}{\sigma^*-\alpha_{2,\kappa}} - \frac{4c'\epsilon}{5\mu_0}\Big)r^2(\rho_K) + \frac{c'\epsilon\kappa}{10\mu_0}r^2(\rho_K).$$

Using the inequalities $\mu_0 \leq 4c'(\sigma^*-\alpha_{2,\kappa})/5$, and $\Upsilon > \kappa$, we finally get

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq \Big(2+\frac{\gamma_1}{\sigma^{*2}}\Big)\alpha_{2,\kappa} + \kappa\epsilon\Big(\frac{2}{\sigma^*-\alpha_{2,\kappa}} - \frac{7c'}{10\mu_0}\Big)r^2(\rho_K).$$

3.6.3.2 Case $||f-f^*||_{L_P^2} > r(\rho_K)$

We have $||f-f^*|| = \rho_K$, therefore, it follows from parts 1 and 2 of Lemma 3.5,

$$\begin{aligned}
&MOM_K\Big(l_{f^*}\Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big) + \frac{\Upsilon}{\chi}\Big(2\zeta(f-f^*) - (f-f^*)^2\Big)\Big) \\
&\qquad \leq \frac{\Upsilon}{\chi}\Big(Q_{7/8,K}\big[2\zeta(f-f^*)\big] - Q_{1/4,K}\big[(f-f^*)^2\big]\Big) - \Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big)Q_{1/8,K}\big[-\zeta^2\big] \\
&\qquad \leq \frac{\Upsilon}{\chi}\Big(\alpha_1 - 4\theta_0^{-2}||f-f^*||_{L_2}^2\Big) + \frac{\sigma^{*2}+\gamma_1}{\sigma^{*2}}\alpha_{2,\kappa} \\
&\qquad \leq \frac{\Upsilon}{\chi}\big(2\epsilon - (4\theta_0)^{-2}\big)||f-f^*||_{L_2}^2 + \frac{\sigma^{*2}+\gamma_1}{\sigma^{*2}}\alpha_{2,\kappa} \\
&\qquad \leq \frac{\kappa}{\sigma^*+\alpha_{2,\kappa}}\big(2\epsilon - (4\theta_0)^{-2}\big)r^2(\rho_K) + \frac{\sigma^{*2}+\gamma_1}{\sigma^{*2}}\alpha_{2,\kappa},
\end{aligned}$$

because $2\epsilon - (4\theta_0)^{-2} \leq 0$. Plugging this back in Equation (3.19), we get

$$\begin{aligned}
T_{K,\mu}(g,\chi,f^*,\sigma^*) &\leq \Big(2+\frac{\gamma_1}{\sigma^{*2}}\Big)\alpha_{2,\kappa} + \Upsilon\left(\frac{2\epsilon-(4\theta_0)^{-2}}{\sigma^*+\alpha_{2,\kappa}}r^2(\rho_K) + \mu\rho_K\right) + \frac{\mu\kappa\rho_K}{10} \\
&\leq \Big(2+\frac{\gamma_1}{\sigma^{*2}}\Big)\alpha_{2,\kappa} + \Upsilon\left(\frac{2\epsilon-(4\theta_0)^{-2}}{\sigma^*+\alpha_{2,\kappa}} + 16\frac{\epsilon}{\mu_0}\right)r^2(\rho_K) + \frac{c'\epsilon\kappa}{10\mu_0}r^2(\rho_K) \\
&\leq \Big(2+\frac{\gamma_1}{\sigma^{*2}}\Big)\alpha_{2,\kappa} + \kappa\left(\frac{2\epsilon-(4\theta_0)^{-2}}{\sigma^*+\alpha_{2,\kappa}} + \frac{11c'\epsilon}{10\mu_0}\right)r^2(\rho_K).
\end{aligned}$$

□

3.6.4 Bound on $F_4^{(\kappa)}$

Let $(g,\chi) \in F_4^{(\kappa)}$. Recall that, in this case, $\chi > \sigma^*+\alpha_{2,\kappa}$. We have

$$\begin{aligned}
T_{K,\mu}(g,\chi,f^*,\sigma^*) &= MOM_K\Big(l_{f^*}\Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big) + (\sigma^*-\chi) + \frac{1}{\chi}\Big(2\zeta(g-f^*) - (g-f^*)^2\Big)\Big) + \mu(||f^*|| - ||g||) \\
&\leq (\sigma^*-\chi) + MOM_K\Big(\frac{l_{f^*}}{\sigma^*} + \frac{1}{\chi}\Big(2\zeta(g-f^*)\Big)\Big) + \mu||f^*-g|| \\
&\leq (\sigma^*-\chi) + \frac{1}{\chi}Q_{3/4,K}\Big[2\zeta(g-f^*)\Big] + \Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big)Q_{1/4,K}[-\zeta^2] + \mu\kappa\rho_K.
\end{aligned}$$

We apply part 2 of Lemma 3.5 and Lemma 3.9, and, using our choice of μ and ρ_K, we deduce that

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq (\sigma^* - \chi) + \Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big)\big(\sigma^{*2} + \gamma_1\big) + \left(\frac{2\epsilon}{\sigma^* + \alpha_{2,\kappa}} + \frac{c'\epsilon\kappa}{\mu_0}\right) r^2(\rho_K).$$

The function $\chi \mapsto -\chi - a/\chi$ is decreasing for $\chi \geq \sqrt{a}$. Applying the first part of Lemma 3.10, we have $\sqrt{\sigma^{*2} + \gamma_1} \leq \sigma^* + \alpha_{2,\kappa}$. Therefore, the maximum of the former function is attained for $\chi = \sigma^* + \alpha_{2,\kappa}$, which yields

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) &\leq -\alpha_{2,\kappa} + \Big(\frac{1}{\sigma^*} - \frac{1}{\sigma^* + \alpha_{2,\kappa}}\Big)\big(\sigma^{*2} + \gamma_1\big) + \left(\frac{1}{\sigma^* + \alpha_{2,\kappa}} + \frac{c'\epsilon\kappa}{\mu_0}\right) r^2(\rho_K) \\ &\leq \left(\frac{\sigma^{*2} + \gamma_1}{\sigma^{*2}} - 1\right)\alpha_{2,\kappa} + \left(\frac{2\epsilon}{\sigma^* + \alpha_{2,\kappa}} + \frac{c'\epsilon\kappa}{\mu_0}\right) r^2(\rho_K). \end{aligned}$$

□

3.6.5 Bound on $F_5^{(\kappa)}$

Let $(g,\chi) \in F_5^{(\kappa)}$. Recall that, in this case, $\chi > \sigma^* + \alpha_{2,\kappa}$. We have

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) &= MOM_K\left(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + (\sigma^* - \chi) + \frac{1}{\chi}\Big(2\zeta(g - f^*) - (g - f^*)^2\Big)\right) + \mu(||f^*|| - ||g||) \\ &\leq (\sigma^* - \chi) + \frac{1}{\chi}\Big(Q_{7/8,K}\big[2\zeta(g - f^*)\big] - Q_{1/4,K}\big[(g - f^*)^2\big]\Big) - \Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big)Q_{1/8,K}\big[-\zeta_2\big] + \mu\kappa\rho_K. \end{aligned}$$

We bound $Q_{7/8,K}\big[2\zeta(g - f^*)\big]$ using part 2 of Lemma 3.5 ; $Q_{1/4,K}\big[(g - f^*)^2\big]$ using part 1 of Lemma 3.5 ; and $Q_{1/8,K}\big[-\zeta_2\big]$ by Lemma 3.9.

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq (\sigma^* - \chi) + \frac{\alpha_1 - ||f^* - g||^2_{L^2_P}(4\theta_0)^{-2}}{\chi} + \Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big)\big(\sigma^{*2} + \gamma_1\big) + \mu\kappa\rho_K.$$

We have $2\epsilon < (4\theta_0)^{-2}$, therefore $\alpha_1 - ||f^* - g||^2_{L^2_P}(4\theta_0)^{-2} \leq 0$ and, plugging in our choice of μ and ρ_K, we can deduce that

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq (\sigma^* - \chi) + \Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big)\big(\sigma^{*2} + \gamma_1\big) + \frac{c'\epsilon\kappa}{\mu_0} r^2(\rho_K)$$

The function $\chi \mapsto -\chi - a/\chi$ is decreasing for $\chi \geq \sqrt{a}$. Applying the first part of Lemma 3.10, we have $\sqrt{\sigma^{*2} + \gamma_1} \leq \sigma^* + \alpha_{2,\kappa}$. Therefore, the maximum of the former function is attained for $\chi = \sigma^* + \alpha_{2,\kappa}$, which yields

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) &\leq -\alpha_{2,\kappa} + \Big(\frac{1}{\sigma^*} - \frac{1}{\sigma^* + \alpha_{2,\kappa}}\Big)\big(\sigma^{*2} + \gamma_1\big) + \frac{c'\epsilon\kappa}{\mu_0} r^2(\rho_K) \\ &\leq \left(\frac{\sigma^{*2} + \gamma_1}{\sigma^{*2}} - 1\right)\alpha_{2,\kappa} + \frac{c'\epsilon\kappa}{\mu_0} r^2(\rho_K). \end{aligned}$$

3.6.6 Bound on $F_6^{(\kappa)}$

Let $(g,\chi) \in F_6^{(\kappa)}$. Recall that, in this case, $\chi > \sigma^* + \alpha_{2,\kappa}$. Following the beginning of the proof in Section 3.6.3, we have as in Equation (3.19)

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq (\sigma^* - \chi) &+ MOM_K\left(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + \frac{\Upsilon}{\chi}\Big(2\zeta(f - f^*) - (f - f^*)^2\Big)\right) \\ &- \mu\Upsilon \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) + \frac{\mu\kappa\rho_K}{10}, \end{aligned} \tag{3.20}$$

3.6.6.1 Case $||f - f^*||_{L^2_P} \leq r(\rho_K)$

Remembering that $||f - f^*|| = \rho_K$, we can deduce that $f \in H_{\rho_K}$ Using the definition of K^*, the fact that $K \geq K^*$, we get that $\rho_K \geq \rho$, and therefore ρ_K follows the sparsity equation, from which we derive $\sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) \geq \Delta(\rho_K) \geq 4\rho_K/5$. Using our choice of μ, we get

$$-\mu \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) \leq -\frac{4c'\epsilon}{5\mu_0} r^2(\rho_K).$$

Combining this with Equation (3.20) and the fact that $(f - f^*)^2 \geq 0$, we get that

$$\begin{aligned}
T_{K,\mu}(g, \chi, f^*, \sigma^*) &\leq (\sigma^* - \chi) + MOM_K\Big(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + \frac{\Upsilon}{\chi}\Big(2\zeta(f - f^*)\Big)\Big) - \Upsilon\frac{4c'\epsilon}{5\mu_0} r^2(\rho_K) + \frac{c'\epsilon\kappa}{10\mu_0} r^2(\rho_K) \\
&\leq (\sigma^* - \chi) + \frac{\Upsilon}{\chi} Q_{3/4,K}\big[2\zeta(f - f^*)\big] - \Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) Q_{1/4,K}\big[-\zeta^2\big] - \Upsilon\frac{4c'\epsilon}{5\mu_0} r^2(\rho_K) + \frac{c'\epsilon\kappa}{10\mu_0} r^2(\rho_K) \\
&\leq (\sigma^* - \chi) + \Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big)(\sigma^{*2} + \gamma_1) + \Upsilon\Big(\frac{1}{\chi} Q_{3/4,K}\big[2\zeta(f - f^*)\big] - \frac{4c'\epsilon}{5\mu_0} r^2(\rho_K)\Big) + \frac{c'\epsilon\kappa}{10\mu_0} r^2(\rho_K).
\end{aligned}$$

Applying part 2 of Lemma 3.5, we get $Q_{3/4,K}\big[2\zeta(f - f^*)\big] \leq \alpha_1 \leq 2\epsilon r^2(\rho_K)$. Therefore,

$$T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq (\sigma^* - \chi) + \Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big)(\sigma^{*2} + \gamma_1) + \Upsilon\Big(\frac{2\epsilon}{\sigma^* + \alpha_{2,\kappa}} - \frac{4c'\epsilon}{5\mu_0}\Big) r^2(\rho_K) + \frac{c'\epsilon\kappa}{10\mu_0} r^2(\rho_K).$$

Using the inequalities $\mu_0 \leq 4c'(\sigma^* + \alpha_{2,\kappa})/5$, and $\Upsilon > \kappa$, we finally get

$$T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq (\sigma^* - \chi) + \Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big)(\sigma^{*2} + \gamma_1) + \kappa\epsilon\Big(\frac{2}{\sigma^* + \alpha_{2,\kappa}} - \frac{7c'}{10\mu_0}\Big) r^2(\rho_K). \tag{3.21}$$

The function $\chi \mapsto -\chi - a/\chi$ is decreasing for $\chi \geq \sqrt{a}$. Applying the first part of Lemma 3.10, we have $\sqrt{\sigma^{*2} + \gamma_1} \leq \sigma^* + \alpha_{2,\kappa}$. Therefore, the maximum of the former function is attained for $\chi = \sigma^* + \alpha_{2,\kappa}$, which, combined with Equation (3.21), yields

$$\begin{aligned}
T_{K,\mu}(g, \chi, f^*, \sigma^*) &\leq -\alpha_{2\kappa} + \frac{\alpha_{2,\kappa}}{\sigma^{*2}}(\sigma^{*2} + \gamma_1) + \kappa\epsilon\Big(\frac{2}{\sigma^* + \alpha_{2,\kappa}} - \frac{7c'}{10\mu_0}\Big) r^2(\rho_K) \\
&\leq \frac{\gamma_1}{\sigma^{*2}}\alpha_{2\kappa} + \kappa\epsilon\Big(\frac{2}{\sigma^* + \alpha_{2,\kappa}} - \frac{7c'}{10\mu_0}\Big) r^2(\rho_K).
\end{aligned}$$

3.6.6.2 Case $||f - f^*||_{L^2_P} > r(\rho_K)$

We have $||f - f^*|| = \rho_K$, therefore, it follows from parts 1 and 2 of Lemma 3.5,

$$\begin{aligned}
&MOM_K\Big(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + \frac{\Upsilon}{\chi}\Big(2\zeta(f - f^*) - (f - f^*)^2\Big)\Big) \\
&\qquad \leq \frac{\Upsilon}{\chi}\Big(Q_{7/8,K}\big[2\zeta(f - f^*)\big] - Q_{1/4,K}\big[(f - f^*)^2\big]\Big) - \Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) Q_{1/8,K}\big[-\zeta^2\big] \\
&\qquad \leq \frac{\Upsilon}{\chi}\Big(\alpha_1 - (4\theta_0)^{-2}||f - f^*||^2_{L_2}\Big) + \Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big)(\sigma^{*2} + \gamma_1). \\
&\qquad \leq \frac{\Upsilon}{\chi}\big(2\epsilon - (4\theta_0)^{-2}\big)||f - f^*||^2_{L_2} + \Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big)(\sigma^{*2} + \gamma_1).
\end{aligned}$$

Plugging this back in Equation 3.20, we get

$$\begin{aligned}
T_{K,\mu}(g, \chi, f^*, \sigma^*) \leq (\sigma^* - \chi) + \frac{\Upsilon}{\chi}\big(2\epsilon - (4\theta_0)^{-2}\big)||f - f^*||^2_{L_2} + \Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big)(\sigma^{*2} + \gamma_1) \\
- \mu\Upsilon \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) + \frac{\mu\kappa\rho_K}{10}.
\end{aligned}$$

As $2\epsilon - (4\theta_0)^{-2} \leq 0$, the function $\chi \mapsto -\chi - a/\chi$ is decreasing for $\chi \geq \sqrt{a}$, and $\sigma^* + \alpha_{2,\kappa} \geq \sqrt{a}$, where here, $a = \sigma^{*2} + \gamma_1 + (4\theta_0)^{-2} - 2\epsilon$, we get

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) &\leq -\alpha_{2,\kappa} + \frac{\Upsilon}{\sigma + \alpha_{2,\kappa}}\big(2\epsilon - (4\theta_0)^{-2}\big)||f - f^*||_{L_2}^2 + \frac{\sigma^{*2} + \gamma_1}{\sigma^{*2}}\alpha_{2,\kappa} \\ &\quad - \mu\Upsilon \sup_{z^* \in \Gamma_{f^*}(\rho_K)} z^*(f - f^*) + \frac{\mu\kappa\rho_K}{10} \\ &\leq \frac{\gamma_1}{\sigma^{*2}}\alpha_{2,\kappa} + \Upsilon\left(\frac{\big(2\epsilon - (4\theta_0)^{-2}\big)r^2(\rho_K)}{\sigma^* + \alpha_{2,\kappa}} + \mu\rho_K\right) + \frac{\mu\kappa\rho_K}{10} \\ &\leq \frac{\gamma_1}{\sigma^{*2}}\alpha_{2,\kappa} + \Upsilon\left(\frac{2\epsilon - (4\theta_0)^{-2}}{\sigma^* + \alpha_{2,\kappa}} + \frac{c'\epsilon}{\mu_0}\right) r^2(\rho_K) + \frac{c'\kappa\epsilon r^2(\rho_K)}{10\mu_0}. \end{aligned}$$

Using $\Upsilon \geq \kappa \geq 0$, and $\left(\frac{2\epsilon - (4\theta_0)^{-2}}{\sigma^* + \alpha_{2,\kappa}} + \frac{c'\epsilon}{\mu_0}\right) \leq 0$, by the second part of Lemma 3.10, we get

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) &\leq \frac{\gamma_1}{\sigma^{*2}}\alpha_{2,\kappa} + \kappa\left(\frac{2\epsilon - (4\theta_0)^{-2}}{\sigma^* + \alpha_{2,\kappa}} + \frac{c'\epsilon}{\mu_0}\right) r^2(\rho_K) + \frac{c'\kappa\epsilon r^2(\rho_K)}{10\mu_0} \\ &\leq \frac{\gamma_1}{\sigma^{*2}}\alpha_{2,\kappa} + \kappa\left(\frac{2\epsilon - (4\theta_0)^{-2}}{\sigma^* + \alpha_{2,\kappa}} + \frac{11c'\epsilon}{10\mu_0}\right) r^2(\rho_K). \end{aligned}$$

□

3.6.7 Bound on $F_7^{(\kappa)}$

Let $(g,\chi) \in F_7^{(\kappa)}$. Recall that, in this case, $\chi < \sigma^* - \alpha_{2,\kappa}$. We have

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) &= MOM_K\left(l_{f^*}\Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big) + (\sigma^* - \chi) + \frac{1}{\chi}\Big(2\zeta(g - f^*) - (g - f^*)^2\Big)\right) + \mu(||f^*|| - ||g||) \\ &\leq \sigma^* - \chi + MOM_K\left(-l_{f^*}\Big(\frac{1}{\chi} - \frac{1}{\sigma^*}\Big) + \frac{1}{\chi}\Big(2\zeta(g - f^*)\Big)\right) + \mu(||f^* - g||) \\ &\leq \sigma^* - \chi + \frac{1}{\chi}Q_{3/4,K}\Big[2\zeta(g - f^*)\Big] + \Big(\frac{1}{\chi} - \frac{1}{\sigma^*}\Big)Q_{1/4,K}\big[-\zeta^2\big] + \mu\kappa\rho_K. \end{aligned}$$

We apply part 2 of Lemma 3.5, using the fact that $\alpha_1 \leq 2\epsilon r^2(\rho_K)$, and Lemma 3.9 to get

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) &\leq \sigma^* - \chi + \frac{2\epsilon r^2(\rho_K)}{\chi} + \Big(\frac{1}{\sigma^*} - \frac{1}{\chi}\Big)(\sigma^{*2} + \gamma_1) + \mu\kappa\rho_K \\ &\leq \sigma^* - \chi + \frac{\sigma^{*2} + \gamma_1}{\sigma^*} + \frac{2\epsilon r^2(\kappa\rho_K) - \sigma^{*2} - \gamma_1}{\chi} + \mu\kappa\rho_K. \end{aligned}$$

We have $2\epsilon r^2(\rho_K) < \sigma^{*2} + \gamma_1$, so that $r^2(\kappa\rho_K) - \sigma^{*2} - \gamma_1 < 0$. The function $\chi \mapsto -\chi - a/\chi$ is increasing for $0 < \chi \leq \sqrt{a}$. Applying the third part of Lemma 3.10, we have $\sigma^* - \alpha_{2,\kappa} \leq \sqrt{\sigma^{*2} + \gamma_1 - r^2(\kappa\rho_K)}$. Therefore, the highest value is attained when $\chi = \sigma^* - \alpha_{2,\kappa}$, and we get

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) &\leq -\alpha_{2,\kappa} + \frac{\sigma^{*2} + \gamma_1}{\sigma^*} + \frac{2\epsilon r^2(\kappa\rho_K) - \sigma^{*2} - \gamma_1}{\sigma^* - \alpha_{2,\kappa}} + \mu\kappa\rho_K \\ &= -\alpha_{2,\kappa} + \frac{\sigma^{*2} + \gamma_1}{\sigma^*} + \frac{2\epsilon r^2(\kappa\rho_K) - \sigma^{*2} - \gamma_1}{\sigma^*} - \frac{2\epsilon r^2(\kappa\rho_K) - \sigma^{*2} - \gamma_1}{\sigma^*} + \frac{2\epsilon r^2(\kappa\rho_K) - \sigma^{*2} - \gamma_1}{\sigma^* - \alpha_{2,\kappa}} + \mu\kappa\rho_K \\ &\leq -\alpha_{2,\kappa} + \frac{2\epsilon r^2(\kappa\rho_K)}{\sigma^*} + \alpha_{2,\kappa}\frac{2\epsilon r^2(\kappa\rho_K) - \sigma^{*2} - \gamma_1}{\sigma^{*2}} + \mu\kappa\rho_K, \end{aligned}$$

Replacing μ and ρ_K by their values, we get

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq \left(\frac{2\epsilon r^2(\kappa\rho_K) - \sigma^{*2} - \gamma_1}{\sigma^{*2}} - 1\right)\alpha_{2,\kappa} + \left(\frac{2\epsilon}{\sigma^*} + \frac{c'\epsilon\kappa}{\mu_0}\right) r^2(\rho_K),$$

□

3.6.8 Bound on $F_8^{(\kappa)}$

Let $(g,\chi) \in F_8^{(\kappa)}$. Recall that, in this case, $\chi < \sigma^* - \alpha_{2,\kappa}$. We have

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) = MOM_K\Big(l_{f^*}\Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big) + (\sigma^*-\chi) + \frac{1}{\chi}\Big(2\zeta(g-f^*) - (g-f^*)^2\Big)\Big) + \mu(||f^*|| - ||g||)$$
$$\leq (\sigma^*-\chi) + \frac{1}{\chi}Q_{7/8,K}\Big[2\zeta(g-f^*)\Big] - \frac{1}{\chi}Q_{1/4,K}\Big[(g-f^*)^2\Big] + \left(\frac{1}{\chi}-\frac{1}{\sigma^*}\right)Q_{1/8,K}[-\zeta^2] + \mu\kappa\rho_K.$$

We bound the second term using part 2 of Lemma 3.5, the third using part 1 of Lemma 3.5 and the fourth using Lemma 3.9. Finally, we replace in the last term μ and ρ_K by their values, so that we have

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq (\sigma^*-\chi) + \frac{\alpha_1 - ||f^*-g||^2_{L^2_P}(4\theta_0)^{-2}}{\chi} + \left(\frac{1}{\chi}-\frac{1}{\sigma^*}\right)\big(-\sigma^{*2}-\gamma_1\big) + \frac{c'\epsilon\kappa}{\mu_0}r^2(\rho_K)$$

We have $2\epsilon < (4\theta_0)^{-2}$, therefore $\alpha_1 - ||f^*-g||^2_{L^2_P}(4\theta_0)^{-2} \leq (2\epsilon - (4\theta_0)^{-2})r^2(\rho_K)$ and we can deduce that

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq (\sigma^*-\chi) + \left(\frac{1}{\sigma^*}-\frac{1}{\chi}\right)\big(\sigma^{*2}+\gamma_1\big) + \left(\frac{(2\epsilon-(4\theta_0)^{-2})}{\sigma^*-\alpha_{2,\kappa}} + \frac{c'\epsilon\kappa}{\mu_0}\right)r^2(\rho_K).$$

The function $\chi \mapsto -\chi - a/\chi$ is increasing for $0 < \chi < \sqrt{a}$. We have $\sigma^* - \alpha_{2,\kappa} < \sqrt{\sigma^{*2}+\gamma_1}$. Therefore, the highest value is attained when $\chi = \sigma^* - \alpha_{2,\kappa}$, and we get

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq \left(1 + \frac{\sigma^{*2}+\gamma_1}{\sigma^{*2}}\right)\alpha_{2,\kappa} + \left(\frac{(2\epsilon-(4\theta_0)^{-2})}{\sigma^*-\alpha_{2,\kappa}} + \frac{c'\epsilon\kappa}{\mu_0}\right)r^2(\rho_K).$$

□

3.6.9 Bound on $F_9^{(\kappa)}$

Let $(g,\chi) \in F_9^{(\kappa)}$. Recall that in this case, $\chi < \sigma^* - \alpha_{2,\kappa}$ Following the beginning of the proof in Section 3.6.3, we have as in Equation 3.19

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq (\sigma^*-\chi) + MOM_K\Big(l_{f^*}\Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big) + \frac{\Upsilon}{\chi}\Big(2\zeta(f-f^*) - (f-f^*)^2\Big)\Big)$$
$$- \mu\Upsilon \sup_{z^*\in\Gamma_{f^*}(\rho_K)} z^*(f-f^*) + \frac{\mu\kappa\rho_K}{10}. \tag{3.22}$$

3.6.9.1 Case $||f-f^*||_{L^2_P} \leq r(\rho_K)$

Remembering that $||f-f^*|| = \rho_K$, we can deduce that $f \in H_{\rho_K}$. Using the definition of K^*, the fact that $K \geq K^*$, we get that $\rho_K \geq \rho$, and therefore ρ_K follows the sparsity equation, from which we derive $\sup_{z^*\in\Gamma_{f^*}(\rho_K)} z^*(f-f^*) \geq \Delta(\rho_K) \geq 4\rho_K/5$. Using our choice of μ, we get

$$-\mu \sup_{z^*\in\Gamma_{f^*}(\rho_K)} z^*(f-f^*) \leq -\frac{4c'\epsilon}{5\mu_0}r^2(\rho_K).$$

Combining this with Equation (3.20) and the fact that $(f-f^*)^2 \geq 0$, we get that

$$T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq (\sigma^*-\chi) + MOM_K\Big(l_{f^*}\Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big) + \frac{\Upsilon}{\chi}\Big(2\zeta(f-f^*)\Big)\Big) - \Upsilon\frac{4c'\epsilon}{5\mu_0}r^2(\rho_K) + \frac{c'\epsilon\kappa}{10\mu_0}r^2(\rho_K)$$
$$\leq (\sigma^*-\chi) + \frac{\Upsilon}{\chi}Q_{3/4,K}\big[2\zeta(f-f^*)\big] - \Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big)Q_{1/4,K}\big[-\zeta^2\big] - \Upsilon\frac{4c'\epsilon}{5\mu_0}r^2(\rho_K) + \frac{c'\epsilon\kappa}{10\mu_0}r^2(\rho_K)$$
$$\leq (\sigma^*-\chi) + \Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big)(\sigma^{*2}+\gamma_1) + \Upsilon\Big(\frac{1}{\chi}Q_{3/4,K}\big[2\zeta(f-f^*)\big] - \frac{4c'\epsilon}{5\mu_0}r^2(\rho_K)\Big) + \frac{c'\epsilon\kappa}{10\mu_0}r^2(\rho_K).$$

Applying part 2 of Lemma 3.5, we get $Q_{3/4,K}\big[2\zeta(f-f^*)\big] \leq \alpha_1 \leq 2\epsilon r^2(\rho_K)$. Therefore,

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) &\leq (\sigma^*-\chi) + \Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big)(\sigma^{*2}+\gamma_1) + \Upsilon\Big(\frac{2\epsilon}{\chi}-\frac{4c'\epsilon}{5\mu_0}\Big)r^2(\rho_K) + \frac{c'\epsilon\kappa}{10\mu_0}r^2(\rho_K) \\ &\leq (\sigma^*-\chi) + \frac{\sigma^{*2}+\gamma_1}{\sigma^*} + \frac{1}{\chi}\Big(-\sigma^{*2}-\gamma_1+2\Upsilon\epsilon r^2(\rho_K)\Big) - \Upsilon\frac{4c'\epsilon}{5\mu_0}r^2(\rho_K) + \frac{c'\epsilon\kappa}{10\mu_0}r^2(\rho_K) \end{aligned}$$

The minimization is done on $\chi > \sigma_{\min}$, therefore we have $\chi^{-1} \leq \sigma_{\min}^{-1}$, and we get

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) &\leq (\sigma^*-\chi) + \frac{\sigma^{*2}+\gamma_1}{\sigma^*} + \frac{-\sigma^{*2}-\gamma_1}{\chi} + \frac{2\Upsilon\epsilon r^2(\rho_K)}{\sigma_{\min}} - \Upsilon\frac{4c'\epsilon}{5\mu_0}r^2(\rho_K) + \frac{c'\epsilon\kappa}{10\mu_0}r^2(\rho_K) \\ &\leq (\sigma^*-\chi) + \frac{\sigma^{*2}+\gamma_1}{\sigma^*} + \frac{-\sigma^{*2}-\gamma_1}{\chi} + \left(\frac{2\epsilon}{\sigma_{\min}}-\frac{4c'\epsilon}{5\mu_0}\right)\Upsilon r^2(\rho_K) + \frac{c'\epsilon\kappa}{10\mu_0}r^2(\rho_K) \end{aligned}$$

Using the inequalities $\mu_0 \leq 4c'\sigma_{\min}/5$, and $\Upsilon > \kappa$, we get

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) &\leq (\sigma^*-\chi) + \frac{\sigma^{*2}+\gamma_1}{\sigma^*} + \frac{-\sigma^{*2}-\gamma_1}{\chi} + \left(\frac{2\epsilon}{\sigma_{\min}}-\frac{4c'\epsilon}{5\mu_0}\right)\kappa r^2(\rho_K) + \frac{c'\epsilon\kappa}{10\mu_0}r^2(\rho_K) \\ &\leq (\sigma^*-\chi) + \frac{\sigma^{*2}+\gamma_1}{\sigma^*} + \frac{-\sigma^{*2}-\gamma_1}{\chi} + \left(\frac{2\epsilon}{\sigma_{\min}}-\frac{7c'\epsilon}{10\mu_0}\right)\kappa r^2(\rho_K) \end{aligned}$$

The function $\chi \mapsto -\chi - a/\chi$ is increasing for $\chi \leq \sqrt{a}$, and $\sigma^* - \alpha_{2,\kappa} \leq \sqrt{a}$, where here $a = \sigma^{*2}+\gamma_1$. Therefore, the maximum of the former function is attained for $\chi = \sigma^* - \alpha_{2,\kappa}$, which yields

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) &\leq \alpha_{2,\kappa} + (\sigma^{*2}+\gamma_1)\left(\frac{1}{\sigma^*}-\frac{1}{\sigma^*-\alpha_{2,\kappa}}\right) + \left(\frac{2\epsilon}{\sigma_{\min}}-\frac{7c'\epsilon}{10\mu_0}\right)\kappa r^2(\rho_K) \\ &\leq \left(\frac{\sigma^{*2}+\gamma_1}{\sigma^{*2}}-1\right)\alpha_{2,\kappa} + \left(\frac{2\epsilon}{\sigma_{\min}}-\frac{7c'\epsilon}{10\mu_0}\right)\kappa r^2(\rho_K). \end{aligned}$$

3.6.9.2 Case $||f-f^*||_{L^2_P} > r(\rho_K)$

We have $||f-f^*|| = \rho_K$, therefore, it follows from parts 1 and 2 of Lemma 3.5,

$$\begin{aligned} &MOM_K\left(l_{f^*}\Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big) + \frac{\Upsilon}{\chi}\Big(2\zeta(f-f^*)-(f-f^*)^2\Big)\right) \\ &\quad\leq \frac{\Upsilon}{\chi}\Big(Q_{7/8,K}\big[2\zeta(f-f^*)\big] - Q_{1/4,K}\big[(f-f^*)^2\big]\Big) - \Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big)Q_{1/8,K}\big[-\zeta^2\big] \\ &\quad\leq \frac{\Upsilon}{\chi}\Big(\alpha_1 - (4\theta_0)^{-2}||f-f^*||^2_{L_2}\Big) + \Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big)(\sigma^{*2}+\gamma_1). \\ &\quad\leq \frac{\Upsilon}{\chi}\big(2\epsilon-(4\theta_0)^{-2}\big)||f-f^*||^2_{L_2} + \Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big)(\sigma^{*2}+\gamma_1). \end{aligned}$$

Plugging this back in Equation 3.22, we get

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq (\sigma^*-\chi) &+ \frac{\Upsilon}{\chi}\big(2\epsilon-(4\theta_0)^{-2}\big)||f-f^*||^2_{L_2} + \Big(\frac{1}{\sigma^*}-\frac{1}{\chi}\Big)(\sigma^{*2}+\gamma_1) \\ &- \mu\Upsilon \sup_{z^*\in\Gamma_{f^*}(\rho_K)} z^*(f-f^*) + \frac{\mu\kappa\rho_K}{10}. \end{aligned}$$

As $2\epsilon-(4\theta_0)^{-2} \leq 0$, the function $\chi \mapsto -\chi - a/\chi$ is increasing for $\chi \leq \sqrt{a}$, and $\sigma^* - \alpha_{2,\kappa} \leq \sqrt{a}$, where here, $a = \sigma^{*2}+\gamma_1-2\epsilon+(4\theta_0)^{-2}$, we get

$$\begin{aligned} T_{K,\mu}(g,\chi,f^*,\sigma^*) &\leq \alpha_{2,\kappa} + \frac{\Upsilon}{\sigma-\alpha_{2,\kappa}}\big(2\epsilon-(4\theta_0)^{-2}\big)||f-f^*||^2_{L_2} + \frac{\sigma^{*2}+\gamma_1}{\sigma^{*2}}\alpha_{2,\kappa} \\ &\quad - \mu\Upsilon \sup_{z^*\in\Gamma_{f^*}(\rho_K)} z^*(f-f^*) + \frac{\mu\kappa\rho_K}{10} \\ &\leq \Big(2+\frac{\gamma_1}{\sigma^{*2}}\Big)\alpha_{2,\kappa} + \Upsilon\left(\frac{2\epsilon-(4\theta_0)^{-2}}{\sigma-\alpha_{2,\kappa}}r^2(\rho_K) + \mu\rho_K\right) + \frac{\mu\kappa\rho_K}{10} \end{aligned}$$

$$\leq \left(2+\frac{\gamma_1}{\sigma^{*2}}\right)\alpha_{2,\kappa} + \Upsilon\left(\frac{2\epsilon-(4\theta_0)^{-2})}{\sigma-\alpha_{2,\kappa}}+\frac{c'\epsilon}{\mu_0}\right)r^2(\rho_K)+\frac{c'\kappa\epsilon r^2(\rho_K)}{10\mu_0}$$
$$\leq \left(2+\frac{\gamma_1}{\sigma^{*2}}\right)\alpha_{2,\kappa} + \kappa\left(\frac{2\epsilon-(4\theta_0)^{-2})}{\sigma-\alpha_{2,\kappa}}+\frac{11c'\epsilon}{10\mu_0}\right)r^2(\rho_K).$$

□

3.7 Proofs of main results

3.7.1 Proof of Theorem 3.4

We begin the proof by applying Lemmas 3.5 and 3.9, and on the rest of the proof, we will reason on the set $\Omega(K) := \Omega_1(K)\cap\Omega_2(K)$. By Lemmas 3.5 and 3.9, $\mathbb{P}\big(\Omega(K)\big) \geq 1-4\exp(-K/4320) - \exp(-5K/13824) \geq 1-5\exp(-K/4320)$. Applying the definition of the estimators in Equation (3.4), we have on $\Omega(K)$

$$\begin{aligned}\mathcal{C}_{K,\mu}(\hat{f},\hat{\sigma}) &\leq \mathcal{C}_{K,\mu}(f^*,\sigma^*) = \sup_{g\in F,\,\chi>\sigma_{\min}} T_{K,\mu}(g,\chi,f^*,\sigma^*)\\ &\leq \max_{i\in[5]}\sup_{(g,\chi)\in F_i^{(1)}} T_{K,\mu}(g,\chi,f^*,\sigma^*) \leq \max_{i\in[5]} B_{i,1}\\ &\leq B_{1,1},\end{aligned}$$

where the last inequality follows by combining Lemma 3.11 (for $\kappa=1$) and the first part of Lemma 3.12.

The proof is completed by application of the following lemma.

Lemma 3.13. *Assume that $\mathcal{C}_{K,\mu}(\hat{f},\hat{\sigma}) \leq B_{1,1}$. Then Equations (3.14)-(3.16) hold and*

$$R(\hat{f}) \leq R(f^*) + \left(1+2\epsilon+\frac{\sigma^{*2}+\gamma_1}{\sigma^{*2}}\right)r^2(2\rho_K) + (\sigma^*+r(2\rho_K))B_{1,1} + (\sigma^*+\alpha_{2,2})(\alpha_{2,2}+2\mu\rho_K).$$

Proof: For any $x\in\mathbb{R}^K$, $Q_{1/2}(x) \geq -Q_{1/2}(-x)$. As a consequence,

$$B_{1,1} \geq \mathcal{C}_{K,\mu}(\hat{f},\hat{\sigma}) = \sup_{g\in F,\,\chi>\sigma_{\min}} T_{K,\mu}(g,\chi,\hat{f},\hat{\sigma}) \geq T_{K,\mu}(f^*,\sigma^*,\hat{f},\hat{\sigma}) = -T_{K,\mu}(\hat{f},\hat{\sigma},f^*,\sigma^*).$$

We deduce that on $\Omega(K)$, $(\hat{f},\hat{\sigma}) \in \{(g,\chi)\in F\times\mathbb{R}_+^* : T_{K,\mu}(g,\chi,f^*,\sigma^*) \geq -B_{1,1}\}$. Applying the second part of Lemma 3.12, we have $-B_{1,1} > \sup_{i=2,\dots 9} B_{i,2}$ and, combining this with Lemma 3.11 (for $\kappa=2$), we get that $(\hat{f},\hat{\sigma}) \in F_1^{(2)}$. By definition of $F_1^{(2)}$, we have $||\hat{f}-f^*|| \leq 2\rho_K$, $||\hat{f}-f^*||_{L_2} \leq r(2\rho_K)$, and $|\sigma-\sigma^*| \leq \alpha_{2,2}$, as claimed.

Finally, we prove the control on the excess risk. We apply Lemma 3.7 with $\rho := 2\rho_K$ and $\alpha_\sigma := \alpha_{2,2}$

$$\begin{aligned}R(\hat{f})-R(f^*) &= ||\hat{f}-f^*||_{L_2}^2 + P[-2\zeta(f-f^*)]\\ &\leq r^2(2\rho_K) + (\sigma^*+r(2\rho_K))T_{K,\lambda}(f^*,\sigma^*,\hat{f},\hat{\sigma}) + 2\epsilon\cdot\max\left(r_M^2(\rho,\gamma_M)\,,\,\frac{384\theta_m^2}{\epsilon^2}\frac{K}{n}\,,\,||f-f^*||_{L_2}\right)\\ &\qquad + (\sigma^{*2}+\gamma_1)\frac{r(2\rho_K)}{\sigma^{*2}} + (\sigma^*+\alpha_{2,2})(\alpha_{2,2}+2\mu\rho_K).\end{aligned}$$

We bound $T_{K,\lambda}(f^*,\sigma^*,\hat{f},\hat{\sigma})$ by $\mathcal{C}_{K,\lambda}(\hat{f},\hat{\sigma})$, and $r_M^2(\rho,\gamma_M)\,,\,\frac{384\theta_m^2}{\epsilon^2}\frac{K}{n}\,,\,||f-f^*||_{L_2}$ by $r^2(2\rho_K)$.

$$\begin{aligned}&R(\hat{f})-R(f^*)\\ &\quad\leq \left(1+2\epsilon+\frac{\sigma^{*2}+\gamma_1}{\sigma^{*2}}\right)r^2(2\rho_K) + (\sigma^*+r(2\rho_K))\mathcal{C}_{K,\lambda}(\hat{f},\hat{\sigma}) + (\sigma^*+\alpha_{2,2})(\alpha_{2,2}+2\mu\rho_K)\end{aligned}$$

$$\leq \left(1 + 2\epsilon + \frac{\sigma^{*2} + \gamma_1}{\sigma^{*2}}\right) r^2(2\rho_K) + (\sigma^* + r(2\rho_K))B_{1,1} + (\sigma^* + \alpha_{2,2})(\alpha_{2,2} + 2\mu\rho_K).$$

□

3.7.2 Proof of Theorem 3.1

We will apply Theorem 3.4. Following [91, pages 39-40] and with $q_0 = 4$ (in their notation), it is possible to choose the function $r(\cdot)$ and ρ such that

$$\rho^* = c_{L,1}||\zeta||_{L^4} s\sqrt{\frac{1}{n}\log\left(\frac{ed}{s}\right)} \text{ and } r^2(\rho^*) = c_{L,2}||\zeta||^2_{L^4}\frac{s}{n}\log\left(\frac{ed}{s}\right),$$

where $c_{L,1}, c_{L,2}$ are constants depending on C_1. Therefore, we can see that, with $c_1 := c'\epsilon c_{L,2}/c_{L,1}$ and $\epsilon := 1/(833\theta_0^2)$, our choice of μ satisfies

$$\mu := c_1\sqrt{\frac{1}{n}\log\left(\frac{ed}{s}\right)} = \frac{c'\epsilon}{\mu_0}\frac{r^2(\rho^*)}{\rho^*},$$

where $\mu_0 := ||\zeta||_{L^4}$. Note that $\mu_0 = \mathbb{E}[\zeta^4]^{1/4} = \kappa^{*1/4}\sigma^*$.

Therefore, conditions in Theorem 3.4 are satisfied, which give the bounds

$$|\hat{\beta} - \beta^*|_1 \leq 2\rho^* \sim c_{L,1}||\zeta||_{L^4} s\sqrt{\frac{1}{n}\log\left(\frac{ed}{s}\right)},$$
$$|\hat{\beta} - \beta^*|_2 \leq 2r^2(\rho^*) \sim c_{L,2}||\zeta||^2_{L^4}\frac{s}{n}\log\left(\frac{ed}{s}\right).$$

The results follows by application of the norm interpolation inequality.

$$|\hat{\beta} - \beta^*|_p \leq |\hat{\beta} - \beta^*|_1^{-1+2/p}|\hat{\beta} - \beta^*|_2^{2-2/p} \lesssim ||\zeta||_{L^4} s^{1/p}\sqrt{\frac{1}{n}\log\left(\frac{ed}{s}\right)}. \square$$

Part II

Conditional copula estimation

Chapter 4

About tests of the “simplifying” assumption for conditional copulas

Abstract

We discuss the so-called “simplifying assumption” of conditional copulas in a general framework. We introduce several tests of the latter assumption for non- and semiparametric copula models. Some related test procedures based on conditioning subsets instead of point-wise events are proposed. The limiting distribution of such test statistics under the null are approximated by several bootstrap schemes, most of them being new. We prove the validity of a particular semiparametric bootstrap scheme. Some simulations illustrate the relevance of our results.

Keywords: Conditional copula, simplifying assumption, bootstrap.

Based on [38]: Derumigny, A., & Fermanian, J. D., About tests of the “simplifying” assumption for conditional copulas. *Dependence Modeling*, 5(1), 154-197, 2017.

4.1 Introduction

In statistical modelling and applied science more generally, it is very common to distinguish two subsets of variables: a random vector of interest (also called explained/exogenous variables) and a vector of covariates (explanatory/endogenous variables). The objective is to predict the law of the former vector given the latter vector belongs to some subset, possibly a singleton. This basic idea constitutes the first step towards forecasting some important statistical sub-products as conditional means, quantiles, volatilities, etc. Formally, consider a d-dimensional random vector $\mathbf{X}$. We are faced with two random sub-vectors $\mathbf{X}_I$ and $\mathbf{X}_J$, s.t. $\mathbf{X} = (\mathbf{X}_I, \mathbf{X}_J)$, $I \cup J = \{1, \dots, d\}$, $I \cap J = \emptyset$, and our models of interest specify the conditional law of $\mathbf{X}_I$ knowing $\mathbf{X}_J = \mathbf{x}_J$ or knowing $\mathbf{X}_J \in A_J$ for some subset $A_J \subset \mathbb{R}^{|J|}$. We use the standard notation for vectors: for any set of indices I, $\mathbf{x}_I$ means the $|I|$-dimensional vector whose arguments are the x_k, $k \in I$. For convenience and without a loss of generality, we will set $I = \{1, \dots, p\}$ and $J = \{p+1, \dots, d\}$.

Besides, the problem of dependence among the components of d-dimensional random vectors has been extensively studied in the academic literature and among practitioners in a lot of different fields. The raise of copulas for more than twenty years illustrates the need of flexible and realistic multivariate

models and tools. When covariates are present and with our notation, the challenge is to study the dependence among the components of $\mathbf{X}_I$ given $\mathbf{X}_J$. Logically, the concept of conditional copulas has emerged. First introduced for pointwise (atomic) conditioning events by Patton ([111, 112]), the definition has been generalized in [52] for arbitrary measurable conditioning subsets. In this paper, we rely on the following definition: for any borel subset $A_J \subset \mathbb{R}^{d-p}$, a conditional copula of $\mathbf{X}_I$ given $(\mathbf{X}_J \in A_J)$ is denoted by $C_{I|J}^{A_J}(\cdot|\mathbf{X}_J \in A_J)$. This is the cdf of the random vector $(F_{1|J}(X_1|\mathbf{X}_J \in A_J), \ldots, F_{p|J}(X_p|\mathbf{X}_J \in A_J))$ given $(\mathbf{X}_J \in A_J)$. Here, $F_{k|J}(\cdot|\mathbf{X}_J \in A_J)$ denotes the conditional law of X_k knowing $\mathbf{X}_J \in A_J$, $k = 1, \ldots, p$. The latter conditional distributions will be assumed continuous in this paper, implying the existence and uniqueness of $C_{I|J}^{A_J}$ (Sklar's theorem). In other words, for any $\mathbf{x}_I \in \mathbb{R}^p$,

$$\mathbb{P}\left(\mathbf{X}_I \leq \mathbf{x}_I | \mathbf{X}_J \in A_J\right) = C_{I|J}^{A_J}\Big(F_{1|J}(x_1|\mathbf{X}_J \in A_J), \ldots, F_{p|J}(x_p|\mathbf{X}_J \in A_J) \,\Big|\, \mathbf{X}_J \in A_J\Big).$$

Note that the influence of A_J on $C_{I|J}^{A_J}$ is twofold: when A_J changes, the conditioning event $(\mathbf{X}_J \in A_J)$ changes, but the conditioned random vector $(F_{1|J}(X_1|\mathbf{X}_J \in A_J), \ldots, F_{p|J}(X_p|\mathbf{X}_J \in A_J))$ changes too.

In particular, when the conditioning events are reduced to singletons, we get that the conditional copula of $\mathbf{X}_I$ knowing $\mathbf{X}_J = \mathbf{x}_J$ is a cdf $C_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J)$ on $[0,1]^p$ s.t., for every $\mathbf{x}_I \in \mathbb{R}^p$,

$$\mathbb{P}\left(\mathbf{X}_I \leq \mathbf{x}_I | \mathbf{X}_J = \mathbf{x}_J\right) = C_{I|J}\left(F_{1|J}(x_1|\mathbf{X}_J = \mathbf{x}_J), \ldots, F_{p|J}(x_p|\mathbf{X}_J = \mathbf{x}_J) \,|\, \mathbf{X}_J = \mathbf{x}_J\right).$$

With generalized inverse functions, an equivalent definition of a conditional copula is as follows:

$$C_{I|J}\left(\mathbf{u}_I \,|\, \mathbf{X}_J = \mathbf{x}_J\right) = F_{I|J}\big(F_{1|J}^{-}(u_1|\mathbf{X}_J = \mathbf{x}_J), \ldots, F_{p|J}^{-}(u_p|\mathbf{X}_J = \mathbf{x}_J)|\mathbf{X}_J = \mathbf{x}_J\big),$$

for every $\mathbf{u}_I$ and $\mathbf{x}_J$, setting $F_{I|J}(\mathbf{x}_I|\mathbf{X}_J = \mathbf{x}_J) := \mathbb{P}\left(\mathbf{X}_I \leq \mathbf{x}_I|\mathbf{X}_J = \mathbf{x}_J\right)$.

Most often, the dependence of $C_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J)$ w.r.t. to $\mathbf{x}_J$ is a source of significant complexities, in terms of model specification and inference. Therefore, most authors assume that the following "simplifying assumption" is fulfilled.

Simplifying assumption $(\mathcal{H}_0)$: the conditional copula $C_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J)$ does not depend on $\mathbf{x}_J$, i.e., for every $\mathbf{u}_I \in [0,1]^p$, the function $\mathbf{x}_J \in \mathbb{R}^{d-p} \mapsto C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J)$ is a constant function (that depends on $\mathbf{u}_I$).

Under the simplifying assumption, we will set $C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J) =: C_{s,I|J}(\mathbf{u}_I)$. The latter identity means that the dependence on $\mathbf{X}_J$ across the components of $\mathbf{X}_I$ is passing only through their conditional margins. Note that $C_{s,I|J}$ is different from the usual copula of $\mathbf{X}_I$: $C_I(\cdot)$ is always the cdf of the vector $(F_1(X_1), \ldots, F_p(X_p))$ whereas, under $\mathcal{H}_0$, $C_{s,I|J}$ is the cdf of the vector $\mathbf{Z}_{I|J} := (F_{1|J}(X_1|X_J), \ldots, F_{p|J}(X_p|X_J))$ (see Proposition 4.4 below). Note that the latter copula is identical to the partial copula introduced by Bergsma [17], and recently studied in [61, 128] in particular. Such a partial copula is always be defined (whether $\mathcal{H}_0$ is satisfied or not) as the cdf of $\mathbf{Z}_{I|J}$. Note that it is equal to $\int_{\mathbb{R}^{d-p}} C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J) dP_J(\mathbf{x}_J)$.

Remark 4.1. *The simplifying assumption $\mathcal{H}_0$ does not imply that $C_{s,I|J}(\cdot)$ is $C_I(\cdot)$, the usual copula of $\mathbf{X}_I$. This can be checked with a simple example: let $\mathbf{X} = (X_1, X_2, X_3)$ be a trivariate random vector s.t., given X_3, $X_1 \sim \mathcal{N}(X_3, 1)$ and $X_2 \sim \mathcal{N}(X_3, 1)$. Moreover, X_1 and X_2 are independent given X_3. The latter variable may be $\mathcal{N}(0,1)$, to fix the ideas. Obviously, with our notation, $I = \{1,2\}$, $J = \{3\}$, $d = 3$ and $p = 2$. Therefore, for any couple $(u_1, u_2) \in [0,1]^2$ and any real number x_3, $C_{1,2|3}(u_1, u_2|x_3) = u_1 u_2$ and does not depend on x_3. Assumption $\mathcal{H}_0$ is then satisfied. But the copula of (X_1, X_2) is not the independence copula, simply because X_1 and X_2 are not independent.*

Basically, it is far from obvious to specify and estimate relevant conditional copula models in practice, especially when the conditioning and/or conditioned variables are numerous. The simplifying assumption is particularly relevant with vine models (see [1], among others). Indeed, to build vines from a d-dimensional random vector $\mathbf{X}$, it is necessary to consider sequences of conditional bivariate copulas $C_{I|J}$, where $I = \{i_1, i_2\}$ is a couple of indices in $\{1, \ldots, d\}$, $J \subset \{1, \ldots, d\}$, $I \cap J = \emptyset$, and $(i_1, i_2|J)$ is a node of the vine. In other words, a bivariate conditional copula is needed at every node of any vine, and the sizes of the conditioning subsets of variables are increasing along the vine. Without additional assumptions, the modelling task becomes rapidly very cumbersome (inference and estimation by maximum likelihood). Therefore, most authors adopt the simplifying assumption $\mathcal{H}_0$ at every node of the vine. Note that the curse of dimensionality still apparently remains because conditional marginal cdfs $F_{k|J}(\cdot|\mathbf{X}_J)$ are invoked with different subsets J of increasing sizes. But this curse can be avoided by calling recursively the non-parametric copulas that have been estimated before (see [105]).

Nonetheless, the simplifying assumption has appeared to be rather restrictive, even if it may be seen as acceptable for practical reasons and in particular situations. The debate between pro and cons of the simplifying assumption is still largely open, particularly when it is called in some vine models. On one side, [68] affirms that this simplifying assumption is not only required for fast, flexible, and robust inference, but that it provides "a rather good approximation, even when the simplifying assumption is far from being fulfilled by the actual model". On the other side, [5] maintain that "this view is too optimistic". They propose a visual test of $\mathcal{H}_0$ when $d = 3$ and in a parametric framework. Their technique was based on local linear approximations and sequential likelihood maximizations. They illustrate the limitations of $\mathcal{H}_0$ by simulation and through real datasets. They note that "an uncritical use of the simplifying assumption may be misleading". Nonetheless, they do not provide formal test procedures. Beside, [4] have proposed a formal likelihood test of the simplifying assumption but when the conditional marginal distributions are known, a rather restrictive situation. Some authors have exhibited classes of parametric distributions for which $\mathcal{H}_0$ is satisfied: see [68], significantly extended by [130]. Nonetheless, such families are rather strongly constrained. Therefore, these two papers propose to approximate some conditional copula models by others for which the simplifying assumption is true. This idea has been developed in [128] in a vine framework, because they recognize that "it is very unlikely that the unknown data generating process satisfies the simplifying assumption in a strict mathematical sense."

Therefore, there is a need for formal universal tests of the simplifying assumption. It is likely that the latter assumption is acceptable in some circumstances, whereas it is too rough in others. This means, for given subsets of indices I and J, we would like to test

$$\mathcal{H}_0 : C_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J) \text{ does not depend on } \mathbf{x}_J,$$

against that opposite assumption. Hereafter, we will propose several test statistics of $\mathcal{H}_0$, possibly assuming that the conditional copula belongs to some parametric family.

Note that several papers have already proposed estimators of conditional copula. [141], [63] and [52] have studied some nonparametric kernel based estimators. [32], [119] studied bayesian additive models of conditional copulas. Recently, [120] invoke B-splines to manage vectors of conditioning variables. In a semiparametric framework, i.e. assuming an underlying parametric family of conditional copulas, numerous models and estimators have been proposed, notably [3], [2], [50] (single-index type models), [140] (additive models), among others. But only a few of these papers have a focus on testing the simplifying assumption $\mathcal{H}_0$ specifically, although convergence of the proposed estimators are necessary to lead such a task in theory. Actually, some tests of $\mathcal{H}_0$ is invoked "in passing" in these papers as

potential applications, but without a general approach and/or without some guidelines to evaluate p-values in practice. As exceptions, in very recent papers, [62] have tackled the simplifying assumption directly through comparisons between conditional and unconditional Kendall's tau. Moreover, [87] have proposed tests of the latter assumption for vine models.

Example 4.2. *To illustrate the problem, let us consider a simple example of $\mathcal{H}_0$ in dimension 3. Assume that $p = 2$ and $d = 3$. For simplicity, let us assume that (X_1, X_2) follows a Gaussian distribution conditionally on X_3, that is :*

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \Big| X_3 = x_3 \sim \mathcal{N}\left(\begin{pmatrix} \mu_1(x_3) \\ \mu_2(x_3) \end{pmatrix}, \begin{pmatrix} \sigma_1^2(x_3) & \rho(x_3)\sigma_1(x_3)\sigma_2(x_3) \\ \rho(x_3)\sigma_1(x_3)\sigma_2(x_3) & \sigma_2^2(x_3) \end{pmatrix} \right). \tag{4.1}$$

Obviously, $\alpha(\cdot) := (\mu_1, \mu_2, \sigma_1, \sigma_2)(\cdot)$ is a parameter that only affects the conditional margins. Moreover, the conditional copula of (X_1, X_2) given $X_3 = x_3$ is gaussian with the parameter $\rho(x_3)$. Six possible cases can then be distinguished:

a. *All variables are mutually independent.*

b. *(X_1, X_2) is independent of X_3, but X_1 and X_2 are not independent.*

c. *X_1 and X_2 are both marginally independent of X_3, but the conditional copula of X_1 and X_2 depends on X_3.*

d. *X_1 (or X_2) and X_3 are not independent but X_1 and X_2 are independent conditionally given X_3.*

e. *X_1 (or X_2) and X_3 are not independent but the conditional copula of X_1 and X_2 is independent of X_3.*

f. *X_1 (or X_2) and X_3 are not independent and the conditional copula of X_1 and X_2 is dependent of X_3.*

These six cases are summarized in the following table:

	$\rho(\cdot) = 0$	$\rho(\cdot) = \rho_0$	$\rho(\cdot)$ *is not constant*
$\alpha(\cdot) = \alpha_0$	*a*	*b*	*c*
$\alpha(\cdot)$ *is not constant*	*d*	*e*	*f*

In the conditional Gaussian model (4.1), the simplifying assumption $\mathcal{H}_0$ consists in assuming that we live in one of the cases $\{a, b, d, e\}$, whereas the alternative cases are c and f. In this model, the conditional copula is entirely determined by the conditional correlation. Note that, in some other models, the conditional correlation can vary only because of the conditional margins, while the conditioning copula stay constant: see [128].

Note that, in general, there is no reason why the conditional margins would be constant in the conditioning variable (and in most applications, they are not). Nevertheless, if we knew the marginal cdfs' were constant with respect to the conditioning variable, then the test of $\mathcal{H}_0$ (i.e. b against c) would become a classical test of independence between $\mathbf{X}_I$ and $\mathbf{X}_J$.

Testing $\mathcal{H}_0$ is closely linked to the m-sample copula problem, for which we have m different and independent samples of a p-dimensional variable $\mathbf{X}_I = (X_1, \ldots, X_p)$. In each sample k, the observations are i.i.d., with their own marginal laws and their own copula $C_{I,k}$. The m-sample copula problem consists on testing whether the m latter copulas $C_{I,k}$ are equal. Note that we could merge all samples into a single

one, and create discrete variables Y_i that are equal to k when i lies in the sample k. Therefore, the m-sample copula problem is formally equivalent to testing $\mathcal{H}_0$ with the conditioning variable $\mathbf{X}_J := Y$.

Conversely, assume we have defined a partition $\{A_{1,J}, \dots, A_{m,J}\}$ of $\mathbb{R}^{d-p}$ composed of borelian subsets such that $\mathbb{P}(\mathbf{X}_J \in A_{k,J}) > 0$ for all $k = 1, \dots, m$, and we want to test

$$\overline{\mathcal{H}}_0 : k \in \{1, \dots, m\} \mapsto C_{I|J}^{A_{k,J}}(\cdot | \mathbf{X}_J \in A_{k,J}) \text{ does not depend on } k.$$

Then, divide the sample in m different sub-samples, where any sub-sample k contains the observations for which the conditioning variable belongs to $A_{k,J}$. Then, $\overline{\mathcal{H}}_0$ is equivalent to a m-sample copula problem. Note that $\overline{\mathcal{H}}_0$ looks like a "consequence" of $\mathcal{H}_0$ when it is not the case in general (see Section 4.3.1), for continuous $\mathbf{X}_J$ variables.

Nonetheless, $\overline{\mathcal{H}}_0$ conveys the same intuition as $\mathcal{H}_0$. Since it can be led more easily in practice (no smoothing is required), some researchers could prefer the former assumption than the latter. That is why it will be discussed hereafter. Note that the 2-sample copula problem has already been addressed by [115], and the m-sample by [23]. However, both paper are designed only in a nonparametric framework, and these authors have not noticed the connection with the simplifying assumption.

The goal of the paper is threefold: first, to write a "state-of-the art" of the simplifying assumption problem; second to propose some "reasonable" test statistics of the simplifying assumption in different contexts; third, to introduce a new approach of the latter problem, through "box-related" zero assumptions and some associated test statistics. Since it is impossible to state the theoretical properties of all these test statistics, we will rely on "ad-hoc arguments" to convince the reader they are relevant, without trying to establish specific results. Globally, this paper can be considered also as a work program around the simplifying assumption $\mathcal{H}_0$ for the next years.

In Section 4.2, we introduce different ways of testing $\mathcal{H}_0$. We propose different test statistics under a fully nonparametric perspective, i.e. when $C_{I|J}$ is not supposed to belong into a particular parametric copula family, through some comparisons between empirical cdfs' in Subsection 4.2.1, or by invoking a particular independence property in Subsection 4.2.2. In Subsection 4.2.3, new tools are needed if we assume underlying parametric copulas. To evaluate the limiting distributions of such tests, we propose several bootstrap techniques (Subsection 4.2.4). Section 4.3 is related to testing $\overline{\mathcal{H}}_0$. In Subsection 4.3.1, we detail the relations between $\mathcal{H}_0$ and $\overline{\mathcal{H}}_0$. Then, we provide tests statistics of $\overline{\mathcal{H}}_0$ for both the nonparametric (Subsection 4.3.2) and the parametric framework (Subsection 4.3.3), as well as bootstrap methods (Subsection 4.3.4). In particular, we prove the validity of the so-called "parametric independent" bootstrap when testing $\overline{\mathcal{H}}_0$. The performances of the latter tests are assessed and compared by simulation in Section 4.4. A table of notation is available in Appendix 4.6 and some of the proofs are collected in Appendix 4.7.

4.2 Tests of the simplifying assumption

4.2.1 "Brute-force" tests of the simplifying assumption

A first natural idea is to build a test of $\mathcal{H}_0$ based on a comparison between some estimates of the conditional copula $C_{I|J}$ with and without the simplifying assumption, for different conditioning events. Such estimates will be called $\hat{C}_{I|J}$ and $\hat{C}_{s,I|J}$ respectively. Then, introducing some distance $\mathcal{D}$ between

conditional distributions, a test can be based on the statistics $\mathcal{D}(\hat{C}_{I|J}, \hat{C}_{s,I|J})$. Following most authors, we immediately think of Kolmogorov-Smirnov-type statistics

$$\mathcal{T}^0_{KS,n} := \|\hat{C}_{I|J} - \hat{C}_{s,I|J}\|_\infty = \sup_{\mathbf{u}_I \in [0,1]^p} \sup_{\mathbf{x}_J \in \mathbb{R}^{d-p}} |\hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}_{s,I|J}(\mathbf{u}_I)|, \tag{4.2}$$

or Cramer von-Mises-type test statistics

$$\mathcal{T}^0_{CvM,n} := \int \left(\hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}_{s,I|J}(\mathbf{u}_I)\right)^2 w(d\mathbf{u}_I, d\mathbf{x}_J), \tag{4.3}$$

for some weight function of bounded variation w, that could be chosen as random (see below).

To evaluate $\hat{C}_{I|J}$, we propose to invoke the nonparametric estimator of conditional copulas proposed by [52]. Alternative kernel-based estimators of conditional copulas can be found in [63], for instance.

Let us start with an iid d-dimensional sample $(\mathbf{X}_i)_{i=1,\dots,n}$. Let $\hat{F}_k$ be the marginal empirical distribution function of X_k, based on the sample $(X_{1,k}, \dots, X_{n,k})$, for any $k = 1, \dots, d$. Our estimator of $C_{I|J}$ will be defined as

$$\hat{C}_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J) := \hat{F}_{I|J}\left(\hat{F}^-_{1|J}(u_1|\mathbf{X}_J = \mathbf{x}_J), \dots, \hat{F}^-_{p|J}(u_p|\mathbf{X}_J = \mathbf{x}_J)|\mathbf{X}_J = \mathbf{x}_J\right),$$

$$\hat{F}_{I|J}(\mathbf{x}_I|\mathbf{X}_J = \mathbf{x}_J) := \frac{1}{n}\sum_{i=1}^n K_n(\mathbf{X}_{i,J}, \mathbf{x}_J)\mathbb{1}(\mathbf{X}_{i,I} \leq \mathbf{x}_I), \tag{4.4}$$

where

$$K_n(\mathbf{X}_{i,J}, \mathbf{x}_J) := K_h\left(\hat{F}_{p+1}(X_{i,p+1}) - \hat{F}_{p+1}(x_{p+1}), \dots, \hat{F}_d(X_{i,d}) - \hat{F}_d(x_d)\right),$$

$$K_h(\mathbf{x}_J) := h^{-(d-p)}K\left(x_{p+1}/h, \dots, x_d/h\right),$$

and K is a $(d-p)$-dimensional kernel. Obviously, for $k \in I$, we have introduced some estimates of the marginal conditional cdfs' similarly:

$$\hat{F}_{k|J}(x|\mathbf{X}_J = \mathbf{x}_J) := \frac{\sum_{i=1}^n K_n(\mathbf{X}_{i,J}, \mathbf{x}_J)\mathbb{1}(\mathbf{X}_{i,I} \leq \mathbf{x}_I)}{\sum_{j=1}^n K_n(\mathbf{X}_{j,J}, \mathbf{x}_J)}. \tag{4.5}$$

Obviously, $h = h(n)$ is the term of a usual bandwidth sequence, where $h(n) \to 0$ when n tends to the infinity. Since $\hat{F}_{I|J}$ is a nearest-neighbors estimator, it does not necessitate a fine-tuning of local bandwidths (except for those values $\mathbf{x}_J$ s.t. $F_J(\mathbf{x}_J)$ is close to one or zero), contrary to more usual Nadaraya-Watson techniques. In other terms, a single convenient choice of h would provide "satisfying" estimates of $\hat{C}_{I|J}(\mathbf{x}_I|\mathbf{X}_J = \mathbf{x}_J)$ for most values of $\mathbf{x}$. For practical reasons, it is important that $\hat{F}_{k|J}(x_k|\mathbf{x}_J)$ belongs to $[0,1]$ and that $\hat{F}_{k|J}(\cdot|\mathbf{x}_J)$ is a true distribution. This is the reason why we use a normalized version for the estimator of the conditional marginal cdfs.

To calculate the latter statistics (4.2) and (4.3), it is necessary to provide an estimate of the underlying conditional copula under $\mathcal{H}_0$. This could be done naively by particularizing a point $\mathbf{x}^*_J \in \mathbb{R}^{d-p}$ and by setting $\hat{C}^{(1)}_{s,I|J}(\cdot) := \hat{C}_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}^*_J)$. Since the choice of $\mathbf{x}^*_J$ is too arbitrary, an alternative could be to set

$$\hat{C}^{(2)}_{s,I|J}(\cdot) := \int \hat{C}_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J)\, w(d\mathbf{x}_J),$$

for some function w that is of bounded variation, and $\int w(d\mathbf{x}_J) = 1$. Unfortunately, the latter choice induce $(d-p)$-dimensional integration procedures, that becomes a numerical problem rapidly when $d-p$ is larger than three.

Therefore, let us randomize the "weight" functions w, to avoid multiple integrations. For instance, choose the empirical distribution of $\mathbf{X}_J$ as w, providing

$$\hat{C}^{(3)}_{s,I|J}(\cdot) := \int \hat{C}_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J)\, \hat{F}_J(d\mathbf{x}_J) = \frac{1}{n}\sum_{i=1}^{n} \hat{C}_{I|J}(\cdot|\mathbf{X}_J = \mathbf{X}_{i,J}). \tag{4.6}$$

An even simpler estimate of $C_{s,I|J}$, the conditional copula of $\mathbf{X}_I$ given $\mathbf{X}_J$ under the simplifying assumption, can be obtained by noting that, under $\mathcal{H}_0$, $C_{s,I|J}$ is the joint law of $\mathbf{Z}_{I|J} := (F_1(X_1|\mathbf{X}_J), \ldots, F_p(X_p|\mathbf{X}_J))$ (see Property 4.4 below). Therefore, it is tempting to estimate $C_{s,I|J}(\mathbf{u}_I)$ by

$$\hat{C}^{(4)}_{s,I|J}(\mathbf{u}_I) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\Big(\hat{F}_{1|J}(X_{i,1}|\mathbf{X}_{i,J}) \leq u_1, \ldots, \hat{F}_{p|J}(X_{i,p}|\mathbf{X}_{i,J}) \leq u_p\Big), \tag{4.7}$$

when $\mathbf{u}_I \in [0,1]^p$, for some consistent estimates $\hat{F}_{k|J}(x_k|\mathbf{x}_J)$ of $F_{k|J}(x_k|\mathbf{x}_J)$. A similar estimator has been promoted and studied in [60] or in [114], but they have considered the empirical copula associated to the pseudo sample $((\hat{F}_1(X_{i1}|\mathbf{X}_{iJ}), \ldots, \hat{F}_p(X_{ip}|\mathbf{X}_{iJ})))_{i=1,\ldots,n}$ instead of its empirical cdf. It will be called $\hat{C}^{(5)}_{s,I|J}$. Hereafter, we will denote $\hat{C}_{s,I|J}$ one of the "averaged" estimators $\hat{C}^{(k)}_{s,I|J}$, $k > 1$ and we can forget the naive pointwise estimator $\hat{C}^{(1)}_{s,I|J}$. Therefore, under some conditions of regularity, we guess that our estimators $\hat{C}_{s,I|J}(\mathbf{u}_I)$ of the conditional copula under $\mathcal{H}_0$ will be $\sqrt{n}$-consistent and asymptotically normal. It has been proved for $C^{(5)}_{s,I|J}$ in [60] or in [114], as a byproduct of the weak convergence of the associated process.

Under $\mathcal{H}_0$, we would like that the previous test statistics $\mathcal{T}^0_{KS,n}$ or $\mathcal{T}^0_{CvM,n}$ are convergent. Typically, such a property is given as a sub-product by the weak convergence of a relevant empirical process, here $(\mathbf{u}_I, \mathbf{x}_J) \in [0,1]^p \times \mathbb{R}^{d-p} \mapsto \sqrt{nh_n^{d-p}}(\hat{C}_{I|J} - C_{I|J})(\mathbf{u}_I|\mathbf{x}_J)$. Unfortunately, this will not be the case in general seing the previous process as a function indexed by $\mathbf{x}_J$, at least for wide ranges of bandwidths. Due to the difficulty of checking the tightness of the process indexed by $\mathbf{x}_J$, some alternative techniques may be required as Gaussian approximations (see [28], e.g.). Nonetheless, they would lead us far beyond the scope of this paper. Therefore, we simply propose to slightly modify the latter test statistics, to manage only a *fixed* set of arguments $\mathbf{x}_J$. For instance, in the case of the Kolmogorov-Smirnov-type test, consider a simple grid $\chi_J := \{\mathbf{x}_{1,J}, \ldots, \mathbf{x}_{m,J}\}$, and the modified test statistics

$$\mathcal{T}^{0,m}_{KS,n} := \sup_{\mathbf{u}_I \in [0,1]^p} \sup_{\mathbf{x}_J \in \chi_J} |\hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}_{s,I|J}(\mathbf{u}_I)|.$$

In the case of the Cramer von-Mises-type test, we can approximate any integral by finite sums, possibly after a change of variable to manage a compactly supported integrand. Actually, this is how they are calculated in practice! For instance, invoking Gaussian quadratures, the modified statistics would be

$$\mathcal{T}^{0,m}_{CvM,n} := \sum_{j=1}^{m} \omega_j \left(\hat{C}_{I|J}(\mathbf{u}_{j,I}|\mathbf{x}_{j,J}) - \hat{C}_{s,I|J}(\mathbf{u}_{j,I})\right)^2, \tag{4.8}$$

for some conveniently chosen constants ω_j, $j = 1, \ldots, m$. Note that the numerical evaluation of $\hat{C}_{I|J}$ is relatively costly. Since quadrature techniques require a lot less points m than "brute-force" equally spaced grids (in dimension d, here), they have to be preferred most often.

Therefore, at least for such modified test statistics, we can insure the tests are convergent. Indeed, under some conditions of regularity, it can be proved that $\hat{C}_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J)$ is consistent and asymptotically normal, for every choice of $\mathbf{u}_I$ and $\mathbf{x}_J$ (see [52]). And a relatively straightforward extension of their Corollary 1 would provide that, under $\mathcal{H}_0$ and for all $\mathcal{U} := (\mathbf{u}_{I,1}, \ldots, \mathbf{u}_{I,q}) \in [0,1]^{p(q+r)}$ and

$\mathcal{X} := (\mathbf{x}_{J,1}, \ldots, \mathbf{x}_{J,q}) \in \mathbb{R}^{(d-p)q}$,

$$\Big\{ \sqrt{nh_n^{d-p}}(\hat{C}_{I|J} - C_{s,I|J})(\mathbf{u}_{I,1}|\mathbf{X}_J = \mathbf{x}_{J,1}), \ldots, \sqrt{nh_n^{d-p}}(\hat{C}_{I|J} - C_{s,I|J})(\mathbf{u}_{I,q}|\mathbf{X}_J = \mathbf{x}_{J,q}),$$
$$\sqrt{n}(\hat{C}_{s,I|J} - C_{s,I|J})(\mathbf{u}_{I,q+1}), \ldots, \sqrt{n}(\hat{C}_{s,I|J} - C_{s,I|J})(\mathbf{u}_{I,q+r}) \Big\},$$

converges in law towards a Gaussian random vector. As a consequence, $\sqrt{nh_n^{d-p}}\mathcal{T}_{KS,n}^{0,m}$ and $nh_n^{d-p}\mathcal{T}_{CvM,n}^{0,m}$ tends to a complex but not degenerate law under the $\mathcal{H}_0$.

Remark 4.3. *Other test statistics of $\mathcal{H}_0$ can be obtained by comparing directly the functions $\hat{C}_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J)$, for different values of $\mathbf{x}_J$. For instance, let us define*

$$\tilde{\mathcal{T}}_{KS,n}^0 := \sup_{\mathbf{x}_J, \mathbf{x}'_J \in \mathbb{R}^{d-p}} \|\hat{C}_{I|J}(\cdot|\mathbf{x}_J) - \hat{C}_{I|J}(\cdot|\mathbf{x}'_J)\|_\infty$$
$$= \sup_{\mathbf{x}_J, \mathbf{x}'_J \in \mathbb{R}^{d-p}} \sup_{\mathbf{u}_I \in [0,1]^p} |\hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}'_J)|, \tag{4.9}$$

or

$$\tilde{\mathcal{T}}_{CvM,n}^0 := \int \Big(\hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}'_J)\Big)^2 \, w(d\mathbf{u}_I, d\mathbf{x}_J, d\mathbf{x}'_J), \tag{4.10}$$

for some function of bounded variation w. As above, modified versions of these statistics can be obtained considering fixed $\mathbf{x}_J$-grids. Since these statistics involve higher dimensional integrals/sums than previously, they will not be studied more in depth.

The L^2-type statistics $\mathcal{T}_{CvM,n}^0$ and $\tilde{\mathcal{T}}_{CvM,n}^0$ involve at least d summations or integrals, which can become numerically expensive when the dimension of $\mathbf{X}$ is "large". Nonetheless, we are free to set convenient weight functions. To reduce the computational cost, several versions of $\mathcal{T}_{CvM,n}^0$ are particularly well-suited, by choosing conveniently the functions w. For instance, consider

$$\mathcal{T}_{CvM,n}^{(1)} := \int \Big(\hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}_{s,I|J}(\mathbf{u}_I)\Big)^2 \, \hat{C}_I(d\mathbf{u}_I) \, \hat{F}_J(d\mathbf{x}_J),$$

where $\hat{F}_J$ and $\hat{C}_I$ denote the empirical cdf of $(\mathbf{X}_{i,J})$ and the empirical copula of $(\mathbf{X}_{i,I})$ respectively. Therefore, $\mathcal{T}_{CvM,n}^{(1)}$ simply becomes

$$\mathcal{T}_{CvM,n}^{(1)} = \frac{1}{n^2} \sum_{j=1}^n \sum_{i=1}^n \Big(\hat{C}_{I|J}(\hat{U}_{i,I}|\mathbf{X}_J = \mathbf{X}_{j,J}) - \hat{C}_{s,I|J}(\hat{U}_{i,I})\Big)^2, \tag{4.11}$$

where $\hat{U}_{i,I} = (\hat{F}_1(X_{i,1}), \ldots, \hat{F}_p(X_{i,p}))$, $i = 1, \ldots, n$. Similarly, we can choose

$$\tilde{\mathcal{T}}_{CvM,n}^{(1)} := \int \Big(\hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}'_J)\Big)^2 \, \hat{C}_I(d\mathbf{u}_I) \, \hat{F}_J(d\mathbf{x}_J) \, \hat{F}_J(d\mathbf{x}'_J)$$
$$= \frac{1}{n^3} \sum_{j=1}^n \sum_{j'=1}^n \sum_{i=1}^n \Big(\hat{C}_{I|J}(\hat{U}_{i,I}|\mathbf{X}_J = \mathbf{X}_{j,J}) - \hat{C}_{I|J}(\hat{U}_{i,I}|\mathbf{X}_J = \mathbf{X}_{j',J})\Big)^2.$$

To deal with a single summations only, it is even possible to propose to set

$$\mathcal{T}_{CvM,n}^{(2)} := \int \Big(\hat{C}_{I|J}(\hat{F}_{1|J}(x_1|\mathbf{x}_J), \ldots, \hat{F}_{p|J}(x_p|\mathbf{x}_J)|\mathbf{x}_J)$$
$$- \hat{C}_{s,I|J}(\hat{F}_{1|J}(x_1|\mathbf{x}_J), \ldots, \hat{F}_{p|J}(x_p|\mathbf{x}_J))\Big)^2 \, \hat{F}(d\mathbf{x}_I, d\mathbf{x}_J),$$

where $\hat{F}$ denotes the empirical cdf of $\mathbf{X}$. This means

$$\mathcal{T}_{CvM,n}^{(2)} = \frac{1}{n} \sum_{i=1}^n \Big(\hat{C}_{I|J}\big(\hat{F}_{1|J}(X_{i,1}|\mathbf{X}_{i,J}), \ldots, \hat{F}_{p|J}(X_{i,p}|\mathbf{X}_{i,J})|\mathbf{X}_J = \mathbf{X}_{i,J}\big)$$
$$- \hat{C}_{s,I|J}\big(\hat{F}_{1|J}(X_{i,1}|\mathbf{X}_{i,J}), \ldots, \hat{F}_{p|J}(X_{i,p}|\mathbf{X}_{i,J})\big)\Big)^2.$$

We have introduced some tests based on comparisons between empirical cdfs'. Obviously, the same idea could be applied to associated densities, as in [49] for instance, or even to other functions of the underlying distributions.

Since the previous test statistics are complicated functionals of some "semi-smoothed" empirical process, it is very challenging to evaluate their asymptotic laws under $\mathcal{H}_0$ analytically. In every case, these limiting laws will not be distribution free, and their calculation would be very tedious. Therefore, as usual with copulas, it is necessary to evaluate the limiting distributions of such tests statistics by a convenient bootstrap procedure (parametric or nonparametric). These bootstrap techniques will be presented in Section 4.2.4.

4.2.2 Tests based on the independence property

Actually, testing $\mathcal{H}_0$ is equivalent to a test of the independence between the random vectors $\mathbf{X}_J$ and $\mathbf{Z}_{I|J} := (F_1(X_1|\mathbf{X}_J), \ldots, F_p(X_p|\mathbf{X}_J))$ strictly speaking, as proved in the following proposition.

Proposition 4.4. *The vectors $\mathbf{Z}_{I|J}$ and $\mathbf{X}_J$ are independent iff $C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J)$ does not depend on $\mathbf{x}_J$ for every vectors $\mathbf{u}_I$ and $\mathbf{x}_J$. In this case, the cdf of $\mathbf{Z}_{I|J}$ is $C_{s,I|J}$.*

Proof: For any vectors $\mathbf{u}_I \in [0,1]^p$ and any subset $A_J \subset \mathbb{R}^{d-p}$,

$$\begin{aligned}
\mathbb{P}(\mathbf{Z}_{I|J} \leq \mathbf{u}_I, \mathbf{X}_J \in A_J) &= \mathbb{E}\left[\mathbb{1}((\mathbf{X}_J \in A_J)\mathbb{P}(\mathbf{Z}_{I|J} \leq \mathbf{u}_I|\mathbf{X}_J)\right] \\
&= \int \mathbb{1}((\mathbf{x}_J \in A_J)\mathbb{P}(\mathbf{Z}_{I|J} \leq \mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J) \\
&= \int_{A_J} \mathbb{P}(F_k(X_k|\mathbf{X}_J = \mathbf{x}_J) \leq u_k, \forall k \in I|\mathbf{X}_J = \mathbf{x}_J)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J) \\
&= \int_{A_J} C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J).
\end{aligned}$$

If $\mathbf{Z}_{I|J}$ and $\mathbf{X}_J$ are independent, then

$$\mathbb{P}(\mathbf{Z}_{I|J} \leq \mathbf{u}_I)\mathbb{P}(\mathbf{X}_J \in A_J) = \int \mathbb{1}((\mathbf{x}_J \in A_J)C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J),$$

for every $\mathbf{u}_I$ and A_J. This implies $\mathbb{P}(\mathbf{Z}_{I|J} \leq \mathbf{u}_I) = C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J)$ for every $\mathbf{u}_I \in [0,1]^p$ and every $\mathbf{x}_J$ in the support of $\mathbf{X}_J$. This means that $C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J)$ does not depend on $\mathbf{x}_J$, because $\mathbf{Z}_{I|J}$ does not depend on any $\mathbf{x}_J$ by definition.

Reciprocally, under $\mathcal{H}_0$, $C_{s,I|J}$ is the cdf of $\mathbf{Z}_{I|J}$. Indeed,

$$\begin{aligned}
\mathbb{P}(\mathbf{Z}_{I|J} \leq \mathbf{u}_I) &= \mathbb{P}\left(F_k(X_k|\mathbf{X}_J) \leq u_k, \forall k \in I\right) \\
&= \int \mathbb{P}\left(F_k(X_k|\mathbf{X}_J = \mathbf{x}_J) \leq u_k, \forall k \in I|\, \mathbf{X}_J = \mathbf{x}_J\right)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J) \\
&= \int C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J) = \int C_{s,I|J}(\mathbf{u}_I)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J) = C_{s,I|J}(\mathbf{u}_I).
\end{aligned}$$

Moreover, due to Sklar's Theorem, we have

$$\begin{aligned}
\mathbb{P}(\mathbf{Z}_{I|J} \leq \mathbf{u}_I, \mathbf{X}_J \in A_J) &= \int \mathbb{1}(\mathbf{x}_J \in A_J)C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J) \\
&= \int \mathbb{1}(\mathbf{x}_J \in A_J)C_{s,I|J}(\mathbf{u}_I)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J) = \mathbb{P}(\mathbf{Z}_{I|J} \leq \mathbf{u}_I)\mathbb{P}(\mathbf{X}_J \in A_J),
\end{aligned}$$

implying the independence between $\mathbf{Z}_{I|J}$ and $\mathbf{X}_J$. □

Then, testing $\mathcal{H}_0$ is formally equivalent to testing

$$\mathcal{H}_0^* : \mathbf{Z}_{I|J} = (F_1(X_1|\mathbf{X}_J), \ldots, F_p(X_p|\mathbf{X}_J)) \text{ and } \mathbf{X}_J \text{ are independent.}$$

Since the conditional marginal cdfs' are not observable, keep in mind that we have to work with pseudo-observations in practice, i.e. vectors of observations that are not independent. In other words, our tests of independence should be based on pseudo-samples

$$\left(\hat{F}_{1|J}(X_{i,1}|\mathbf{X}_{i,J}), \ldots, \hat{F}_{p|J}(X_{i,p}|\mathbf{X}_{i,J})\right)_{i=1,\ldots,n} := (\hat{\mathbf{Z}}_{i,I|J})_{i=1,\ldots,n}, \tag{4.12}$$

for some consistent estimate $\hat{F}_{k|J}(\cdot|\mathbf{X}_J)$, $k \in I$ of the conditional cdfs', for example as defined in Equation (4.5). The chance of getting distribution-free asymptotic statistics will be very tiny, and we will have to rely on some bootstrap techniques again. To summarize, we should be able to apply some usual tests of independence, but replacing iid observations with (dependent) pseudo-observations.

Most of the tests of $\mathcal{H}_0^*$ rely on the joint law of $(\mathbf{Z}_{I|J}, \mathbf{X}_J)$, that may be evaluated empirically as

$$\begin{aligned} G_{I,J}(\mathbf{x}_I, \mathbf{x}_J) &:= \mathbb{P}(\mathbf{Z}_{I|J} \leq \mathbf{x}_I, \mathbf{X}_J \leq \mathbf{x}_J) \\ &\simeq \quad \hat{G}_{I,J}(\mathbf{x}) := n^{-1} \sum_{i=1}^{n} \mathbb{1}(\hat{\mathbf{Z}}_{i,I|J} \leq \mathbf{x}_I, \mathbf{X}_{i,J} \leq \mathbf{x}_J). \end{aligned}$$

Now, let us propose some classical strategies to build independence tests.

- Chi-square-type tests of independence: Let $B_1, \ldots, B_N$ (resp. $A_1, \ldots, A_m$) some disjoint subsets in $\mathbb{R}^p$ (resp. $\mathbb{R}^{d-p}$).

$$\mathcal{I}_{\chi,n} = n \sum_{k=1}^{N} \sum_{l=1}^{m} \frac{\left(\hat{G}_{I,J}(B_k \times A_l) - \hat{G}_{I,J}(B_k \times \mathbb{R}^{d-p})\hat{G}_{I,J}(\mathbb{R}^p \times A_l)\right)^2}{\hat{G}_{I,J}(B_k \times \mathbb{R}^{d-p})\hat{G}_{I,J}(\mathbb{R}^p \times A_l)}. \tag{4.13}$$

- Distance between distributions:

$$\mathcal{I}_{KS,n} = \sup_{\mathbf{x} \in \mathbb{R}^d} |\hat{G}_{I,J}(\mathbf{x}) - \hat{G}_{I,J}(\mathbf{x}_I, \infty^{d-p})\hat{G}_{I,J}(\infty^p, \mathbf{x}_J)|, \text{ or} \tag{4.14}$$

$$\mathcal{I}_{2,n} = \int \left(\hat{G}_{I,J}(\mathbf{x}) - \hat{G}_{I,J}(\mathbf{x}_I, \infty^{d-p})\hat{G}_{I,J}(\infty^p, \mathbf{x}_J)\right)^2 \omega(\mathbf{x})\, d\mathbf{x}, \tag{4.15}$$

  for some (possibly random) weight function ω. Particularly, we can propose the single sum

$$\begin{aligned} \mathcal{I}_{CvM,n} &= \int \left(\hat{G}_{I,J}(\mathbf{x}) - \hat{G}_{I,J}(\mathbf{x}_I, \infty^{d-p})\hat{G}_{I,J}(\infty^p, \mathbf{x}_J)\right)^2 \hat{G}_{I,J}(d\mathbf{x}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\hat{G}_{I,J}(\hat{\mathbf{Z}}_{i,I|J}, \mathbf{X}_{i,J}) - \hat{G}_{I,J}(\hat{\mathbf{Z}}_{i,I|J}, \infty^{d-p})\hat{G}_{I,J}(\infty^p, \mathbf{X}_{i,J})\right)^2. \end{aligned} \tag{4.16}$$

- Tests of independence based on comparisons of copulas: let $\breve{C}_{I,J}$ and $\hat{C}_J$ be the empirical copulas based on the pseudo-sample $(\hat{\mathbf{Z}}_{i,I|J}, \mathbf{X}_{i,J})_{i=1,\ldots,n}$, and $(\mathbf{X}_{i,J})_{i=1,\ldots,n}$ respectively. Set

$$\breve{\mathcal{I}}_{KS,n} = \sup_{\mathbf{u} \in [0,1]^d} |\breve{C}_{I,J}(\mathbf{u}) - \hat{C}^{(k)}_{s,I|J}(\mathbf{u}_I)\hat{C}_J(\mathbf{u}_J)|, k = 1, \ldots, 5, \text{ or}$$

$$\breve{\mathcal{I}}_{2,n} = \int_{\mathbf{u} \in [0,1]^d} \left(\breve{C}_{I,J}(\mathbf{u}) - \hat{C}^{(k)}_{s,I|J}(\mathbf{u}_I)\hat{C}_J(\mathbf{u}_J)\right)^2 \omega(\mathbf{u})\, d\mathbf{u},$$

  and in particular

$$\breve{\mathcal{I}}_{CvM,n} = \int_{\mathbf{u} \in [0,1]^d} \left(\breve{C}_{I,J}(\mathbf{u}) - \hat{C}^{(k)}_{s,I|J}(\mathbf{u}_I)\hat{C}_J(\mathbf{u}_J)\right)^2 \breve{C}_{I,J}(d\mathbf{u}).$$

The underlying ideas of the test statistics $\check{\mathcal{I}}_{KS,n}$ and $\check{\mathcal{I}}_{CvM,n}$ are similar to those that have been proposed by Deheuvels ([33],[34]) in the case of unconditional copulas. Nonetheless, in our case, we have to calculate pseudo-samples of the pseudo-observations $(\hat{\mathbf{Z}}_{i,I|J})$ and $(\mathbf{X}_{i,J})$, instead of a usual pseudo-sample of $(\mathbf{X}_i)$.

Note that the latter techniques require the evaluation of some conditional distributions, for instance by kernel smoothing. Therefore, the level of numerical complexity of these test statistics of $\mathcal{H}_0^*$ is comparable with those we have proposed before to test $\mathcal{H}_0$ directly.

4.2.3 Parametric tests of the simplifying assumption

In practice, modelers often assume a priori that the underlying copulas belong to some specified parametric family $\mathcal{C} := \{C_\theta, \theta \in \Theta \subset \mathbb{R}^m\}$. Let us adapt our tests under this parametric assumption. Apparently, we would like to test

$$\check{\mathcal{H}}_0 : C_{I|J}(\cdot|\mathbf{X}_J) = C_\theta(\cdot), \text{ for some } \theta \in \Theta \text{ and almost every } \mathbf{X}_J.$$

Actually, $\check{\mathcal{H}}_0$ requires two different things: the fact that the conditional copula is a constant copula w.r.t. its conditioning events (test of $\mathcal{H}_0$) and, additionally, that the right copula belongs to $\mathcal{C}$ (classical composite Goodness-of-Fit test). Under this point of view, we would have to adapt "omnibus" specification tests to manage conditional copulas and pseudo observations. For instance, and among of alternatives, we could consider an amended version of Andrews's ([7]) specification test

$$CK_n := \frac{1}{\sqrt{n}} \max_{j \leq n} \left| \sum_{i=1}^{n} \left[\mathbb{1}(\hat{\mathbf{Z}}_{i,I|J} \leq \hat{\mathbf{Z}}_{j,I|J}) - C_{\hat{\theta}_0}(\hat{\mathbf{Z}}_{j,I|J}) \right] \mathbb{1}(\mathbf{X}_{i,J} \leq \mathbf{X}_{j,J}) \right|,$$

recalling the notation in (4.12). For other ideas of the same type, see [145] and the references therein.

The latter global approach is probably too demanding. Here, we prefer to isolate the initial problem that was related to the simplifying assumption only. Therefore, let us assume that, for every $\mathbf{x}_J$, there exists a parameter $\theta(\mathbf{x}_J)$ such that $C_{I|J}(\cdot|\mathbf{x}_J) = C_{\theta(\mathbf{x}_J)}(\cdot)$. To simplify, we assume the function $\theta(\cdot)$ is continuous. Our problem is then reduced to testing the constancy of θ, i.e.

$$\mathcal{H}_0^c : \text{the function } \mathbf{x}_J \mapsto \theta(\mathbf{x}_J) \text{ is a constant, called } \theta_0.$$

For every $\mathbf{x}_J$, assume we estimate $\theta(\mathbf{x}_J)$ consistently. For instance, this can be done by modifying the standard semiparametric Canonical Maximum Likelihood methodology ([56, 137]): set

$$\hat{\theta}(\mathbf{x}_J) := \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log c_\theta \left(\hat{F}_{1|J}(X_{i,1}|\mathbf{X}_J = \mathbf{X}_{i,J}), \ldots, \hat{F}_{p|J}(X_{i,p}|\mathbf{X}_J = \mathbf{X}_{i,J}) \right) \cdot K_n(\mathbf{X}_{i,J}, \mathbf{x}_J),$$

through usual kernel smoothing in $\mathbb{R}^{d-p}$, where $c_\theta(\mathbf{u}) := \partial^p C_\theta(\mathbf{u}) / \partial u_1 \cdots \partial u_p$ for $\theta \in \Theta$ and $\mathbf{u} \in [0,1]^p$. Alternatively, we could consider

$$\tilde{\theta}(\mathbf{x}_J) := \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log c_\theta \left(\hat{F}_{1|J}(X_{i,1}|\mathbf{X}_J = \mathbf{x}_J), \ldots, \hat{F}_{p|J}(X_{i,p}|\mathbf{X}_J = \mathbf{x}_J) \right) \cdot K_n(\mathbf{X}_{i,J}, \mathbf{x}_J),$$

instead of $\hat{\theta}(\mathbf{x}_J)$. See [2] concerning the theoretical properties of $\tilde{\theta}(\mathbf{x}_J)$ and some choice of conditional cdfs'. Those of $\hat{\theta}(\mathbf{x}_J)$ remain to be stated precisely, to the best of our knowledge. But there is no doubt both methodologies provide consistent estimators, even jointly, under some conditions of regularity.

Under $\mathcal{H}_0^c$, the natural "unconditional" copula parameter θ_0 of the copula of the $\mathbf{Z}_{I|J}$ will be estimated by

$$\hat{\theta}_0 := \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log c_\theta \left(\hat{F}_{1|J}(X_{i,1}|\mathbf{X}_{i,J}), \ldots, \hat{F}_{p|J}(X_{i,p}|\mathbf{X}_{i,J}) \right). \tag{4.17}$$

Surprisingly, the theoretical properties of the latter estimator do not seem to have been established in the literature explicitly. Nonetheless, the latter M-estimator is a particular case of those considered in [50] in the framework of single-index models when the link function is a known function (that does not depend on the index). Therefore, by adapting their assumption in the current framework, we easily obtain that $\hat{\theta}_0$ is consistent and asymptotically normal if c_θ is sufficiently regular, for convenient choices of bandwidths and kernels.

Now, there are some challengers to test $\mathcal{H}_0^c$:

- Tests based on the comparison between $\hat{\theta}(\cdot)$ and $\hat{\theta}_0$:

$$\mathcal{T}_\infty^c := \sup_{\mathbf{x}_J\in\mathbb{R}^{d-p}} \|\hat{\theta}(\mathbf{x}_J) - \hat{\theta}_0\|, \text{ or } \mathcal{T}_2^c := \int \|\hat{\theta}(\mathbf{x}_J) - \hat{\theta}_0\|^2 \omega(\mathbf{x}_J)\, d\mathbf{x}_J, \tag{4.18}$$

  for some weight function ω.

- Tests based on the comparison between $C_{\hat{\theta}(\cdot)}$ and $C_{\hat{\theta}_0}$:

$$\mathcal{T}_{dist}^c := \int dist\left(C_{\hat{\theta}(\mathbf{x}_J)}, C_{\hat{\theta}_0} \right) \omega(\mathbf{x}_J)\, d\mathbf{x}_J, \tag{4.19}$$

  for some distance $dist(\cdot,\cdot)$ between cdfs'.

- Tests based on the comparison between copula densities (when they exist):

$$\mathcal{T}_{dens}^c := \int \left(c_{\hat{\theta}(\mathbf{x}_J)}(\mathbf{u}_I) - c_{\hat{\theta}_0}(\mathbf{u}_I) \right)^2 \omega(\mathbf{u}_I, \mathbf{x}_J)\, d\mathbf{u}_I\, d\mathbf{x}_J. \tag{4.20}$$

Remark 4.5. *It might be difficult to compute some of these integrals numerically, because of unbounded supports. One solution is to to make change of variables. For example,*

$$\mathcal{T}_2^c = \int \|\hat{\theta}(F_J^-(\mathbf{u}_J)) - \hat{\theta}_0\|^2 \omega(F_J^-(\mathbf{u}_J)) \frac{d\mathbf{u}_J}{f_J(F_J^-(\mathbf{u}_J))}.$$

Therefore, the choice $\omega = f_J$ allows us to simplify the latter statistics to $\int \|\hat{\theta}(F_J^-(\mathbf{u}_J)) - \hat{\theta}_0\|^2 d\mathbf{u}_J$, which is rather easy to evaluate. We used this trick in the numerical section below.

4.2.4 Bootstrap techniques for tests of $\mathcal{H}_0$

It is necessary to evaluate the limiting laws of the latter test statistics under the null. As a matter of fact, we generally cannot exhibit explicit - and distribution-free a fortiori - expressions for these limiting laws. The common technique is provided by bootstrap resampling schemes.

More precisely, let us consider a general statistics $\mathcal{T}$, built from the initial sample $\mathcal{S} := (\mathbf{X}_1, \ldots, \mathbf{X}_n)$. The main idea of the bootstrap is to construct N new samples $\mathcal{S}^* := (\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*)$ following a given resampling scheme given $\mathcal{S}$. Then, for each bootstrap sample $\mathcal{S}^*$, we will evaluate a bootstrapped test statistics $\mathcal{T}^*$, and the empirical law of all these N statistics is used as an approximation of the limiting law of the initial statistic $\mathcal{T}$.

4.2.4.1 Some resampling schemes

The first natural idea is to invoke Efron's usual "nonparametric bootstrap", where we draw independently with replacement $\mathbf{X}_i^*$ for $i = 1, \dots, n$ among the initial sample $\mathcal{S} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$. This provides a bootstrap sample $\mathcal{S}^* := (\mathbf{X}_1^*, \dots, \mathbf{X}_n^*)$.

The nonparametric bootstrap is an "omnibus" procedure whose theoretical properties are well-known but that may not be particularly adapted to the problem at hand. Therefore, we will propose alternative sampling schemes that should be of interest, even if we do not state their validity on the theoretical basis. Such a task is left for further researches.

An natural idea would be to use some properties of $\mathbf{X}$ under $\mathcal{H}_0$, in particular the characterization given in Proposition 4.4: under $\mathcal{H}_0$, we known that $\mathbf{Z}_{i,I|J}$ and $\mathbf{X}_{i,J}$ are independent. This will be only relevant for the tests of Subsection 4.2.2, and for a few tests of Subsection 4.2.1, where such statistics are based on the pseudo-sample $(\hat{\mathbf{Z}}_{i,I|J}, \mathbf{X}_{i,J})_{i=1,\dots,n}$. Therefore, we propose the following so-called "pseudo-independent bootstrap" scheme:

Repeat, for $i = 1$ to n,

1. draw $\mathbf{X}_{i,J}^*$ among $(\mathbf{X}_{j,J})_{j=1,\dots,n}$;
2. draw $\hat{\mathbf{Z}}_{i,I|J}^*$ independently, among the observations $\hat{\mathbf{Z}}_{j,I|J}$, $j = 1, \dots, n$.

This provides a bootstrap sample $\mathcal{S}^* := \big((\hat{\mathbf{Z}}_{1,I|J}^*, \mathbf{X}_{1,J}^*), \dots, (\hat{\mathbf{Z}}_{n,I|J}^*, \mathbf{X}_{n,J}^*)\big)$.

Note that we could invoke the same idea, but with a usual nonparametric bootstrap perspective: draw with replacement a n-sample among the pseudo-observations $(\hat{\mathbf{Z}}_{i,I|J}, \mathbf{X}_{i,J})_{i=1,\dots,n}$ for each bootstrap sample. This can be called a "pseudo-nonparametric bootstrap" scheme.

Moreover, note that we cannot draw independently $\mathbf{X}_{i,J}^*$ among $(\mathbf{X}_{j,J})_{j=1,\dots,n}$, and beside $\mathbf{X}_{i,I}^*$ among $(\mathbf{X}_{j,I})_{j=1,\dots,n}$ independently. Indeed, $\mathcal{H}_0$ does not imply the independence between $\mathbf{X}_I$ and $\mathbf{X}_J$. At the opposite, it makes sense to build a "conditional bootstrap" as follows:

Repeat, for $i = 1$ to n,

1. draw $\mathbf{X}_{i,J}^*$ among $(\mathbf{X}_{j,J})_{j=1,\dots,n}$;
2. draw $\hat{\mathbf{X}}_{i,I}^*$ independently, along the estimated conditional law of $\mathbf{X}_I$ given $\mathbf{X}_J = \mathbf{X}_{i,J}^*$. This can be down by drawing a realization along the law $\hat{F}_{I|J}(\cdot|\mathbf{X}_J = \mathbf{X}_{i,J}^*)$, for instance (see (4.4)). This can be done easily because the latter law is purely discrete, with unequal weights that depend on $\mathbf{X}_{i,J}^*$ and $\mathcal{S}$.

This provides a bootstrap sample $\mathcal{S}^* := \big((\hat{\mathbf{X}}_{1,I}^*, \mathbf{X}_{1,J}^*), \dots, (\hat{\mathbf{X}}_{n,I}^*, \mathbf{X}_{n,J}^*)\big)$.

Remark 4.6. *Note that the latter way of resampling is not far from the usual nonparametric bootstrap. Indeed, when the bandwidths tend to zero, once $\mathbf{x}_J^* = \mathbf{X}_{i,J}$ is drawn, the procedure above will select the other components of $\mathbf{X}_i$ (or close values), i.e. the probability that $\mathbf{x}_I^* = \mathbf{X}_{i,I}$ is "high".*

In the parametric framework, we might also want to use an appropriate resampling scheme. As a matter of fact, all the previous resampling schemes can be used, as in the nonparametric framework, but we would not take advantage of the parametric hypothesis, i.e. the fact that all conditional copulas belong to a known family. We have also to keep in mind that even if the conditional copula has a

parametric form, the global model is not fully parametric, because we have not provided a parametric model neither for the conditional marginal cdfs $F_{k|J}$, $k = 1, \dots, p$, nor for the cdf of $\mathbf{X}_J$.

Therefore, we can invoke the null hypothesis $\mathcal{H}_0^c$ and approximate the real copula C_{θ_0} of $\mathbf{Z}_{I|J}$ by $C_{\hat{\theta}_0}$. This leads us to define the following "parametric independent bootstrap":

Repeat, for $i = 1$ to n,

1. draw $\mathbf{X}^*_{i,J}$ among $(\mathbf{X}_{j,J})_{j=1,\dots,n}$;
2. sample $\mathbf{Z}^*_{i,I|J,\hat{\theta}_0}$ from the copula with parameter $\hat{\theta}_0$ independently.

This provides a bootstrap sample $\mathcal{S}^* := \big((\mathbf{Z}^*_{1,I|J,\hat{\theta}_0}, \mathbf{X}^*_{1,J}), \dots, (\mathbf{Z}^*_{n,I|J,\hat{\theta}_0}, \mathbf{X}^*_{n,J})\big)$.

Remark 4.7. *At first sight, this might seem like a strange mixing of parametric and nonparametric bootstrap. If $|J| = 1$, we can nonetheless do a "full parametric bootstrap", by observing that all estimators of our previous test statistics do not depend on $\mathbf{X}_J$, but on realizations of $\hat{F}_J(\mathbf{X}_J)$ (see Equations (4.4) and (4.5)). Since the law of latter variable is close to a uniform distribution, it is tempting to sample $V^*_{i,J} \sim \mathcal{U}_{[0,1]}$ at the first stage, $i = 1, \dots, n$, and then to replace $\hat{F}_J(\mathbf{X}_{i,J})$ with $V^*_{i,J}$ to get an alternative bootstrap sample.*

Without using $\mathcal{H}_0^c$, we could define the "parametric conditional bootstrap" as:

Repeat, for $i = 1$ to n,

- draw $\mathbf{X}^*_{i,J}$ among $(\mathbf{X}_{j,J})_{j=1,\dots,n}$;
- sample $\mathbf{Z}^*_{i,I|J,\theta^*_i}$ from the copula with parameter $\hat{\theta}(\mathbf{X}^*_{i,J})$.

This provides a bootstrap sample $\mathcal{S}^* := \big((\mathbf{Z}^*_{1,I|J,\theta^*_i}, \mathbf{X}^*_{1,J}), \dots, (\mathbf{Z}^*_{n,I|J,\theta^*_i}, \mathbf{X}^*_{n,J})\big)$.

Note that, in several resampling schemes, we should be able to keep the same $\mathbf{X}_J$ as in the original sample, and simulate only $\mathbf{Z}^*_{i,I|J}$ in step 2, as in [7], pages 10-11. Such an idea has been proposed by [108], in a slightly different framework and univariate conditioning variables. They proved that such a bootstrap scheme "works", after a fine-tuning of different smoothing parameters: see their Theorem 1.

4.2.4.2 Bootstrapped test statistics

The problem is now to evaluate the law of a given test statistic, say $\mathcal{T}$, under $\mathcal{H}_0$ by the some bootstrap techniques. We recall the main technique in the case of the classical nonparametric bootstrap. We conjecture that the idea is still theoretically sound under the other resampling schemes that have been proposed in Subsection 4.2.4.1.

The principle for the nonparametric bootstrap is based on the weak convergence of the underlying empirical process. Formally, if $\mathcal{S} := \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ in an iid sample in $\mathbb{R}^d$, $\mathbf{X} \sim F$ and if F_n denotes its empirical distribution, it is well-known that $\sqrt{n}\,(F_n - F)$ tends weakly in ℓ^∞ towards a d-dimensional Brownian bridge $\mathbb{B}_F$. And the nonparametric bootstrap works in the sense that $\sqrt{n}\,(F^*_n - F_n)$ converges weakly towards a process $\mathbb{B}'_F$, an independent version of $\mathbb{B}_F$, given the initial sample $\mathcal{S}$.

Due to the Delta Method, for every Hadamard-differentiable functional χ from $\ell^\infty(\mathbb{R}^d)$ to $\mathbb{R}$, there exists a random variable H_χ s.t. $\sqrt{n}\,(\chi(F_n) - \chi(F)) \Rightarrow H_\chi$. Assume a test statistics $\mathcal{T}_n$ of $\mathcal{H}_0$ can be written as a sufficiently regular functional of the underlying empirical process as

$$\mathcal{T}_n := \psi\left(\sqrt{n}\,(\chi_s(F_n) - \chi(F_n))\right),$$

where $\chi_s(F) = \chi(F)$ under the null assumption. Then, under $\mathcal{H}_0$, we can rewrite this expression as

$$\mathcal{T}_n := \psi\left(\sqrt{n}\left(\chi_s(F_n) - \chi_s(F) + \chi(F) - \chi(F_n)\right)\right). \tag{4.21}$$

Given any bootstrap sample $\mathcal{S}^*$ and the associated empirical distribution F_n^*, the usual bootstrap equivalent of $\mathcal{T}_n$ is

$$\mathcal{T}_n^* := \psi\left(\sqrt{n}\left(\chi_s(F_n^*) - \chi_s(F_n) + \chi(F_n) - \chi(F_n^*)\right)\right),$$

from Equation (4.21). See [139], Section 3.9, for details and mathematically sound statements.

Applying these ideas, we can guess the bootstrapped statistics corresponding to the tests statistics of $\mathcal{H}_0$, at least when the usual nonparametric bootstrap is invoked.

Let us illustrate the idea with $\mathcal{T}^0_{KM,n}$. Note that $\hat{C}_{I|J}(\cdot|\mathbf{X}_J = \cdot) = \chi_{KM}(F_n)(\cdot)$ and $\hat{C}_{s,I|J} = \chi_{s,KM}(F_n)$ for some smoothed functional χ_{KM} and $\chi_{s,KM}$. Under $\mathcal{H}_0$, $\chi_{KM} = \chi_{s,KM}$ and $\mathcal{T}^0_{KS,n} := \|\chi_{KM}(F_n) - \chi_{KM}(F) - \chi_{s,KM}(F_n) + \chi_{s,KM}(F)\|_\infty$. Therefore, its bootstrapped version is

$$\begin{aligned}\mathcal{T}^{0,*}_{KS,n} &:= \|\chi_{KM}(F_n^*) - \chi_{KM}(F_n) - \chi_{s,KM}(F_n^*) + \chi_{s,KM}(F_n)\|_\infty \\ &= \|\hat{C}^*_{I|J} - \hat{C}_{I|J} - \hat{C}^*_{s,I|J} + \hat{C}_{s,I|J}\|_\infty.\end{aligned}$$

Obviously, the functions $\hat{C}^*_{I|J}$ and $\hat{C}^*_{s,I|J}$ have been calculated as $\hat{C}_{I|J}$ and $\hat{C}_{s,I|J}$ respectively, but replacing $\mathcal{S}$ by $\mathcal{S}^*$. Similarly, the bootstrapped versions of some Cramer von-Mises-type test statistics are

$$\mathcal{T}^{0,*}_{CvM,n} := \int \left(\hat{C}^*_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}^*_{s,I|J}(\mathbf{u}_I) + \hat{C}_{s,I|J}(\mathbf{u}_I)\right)^2 w(d\mathbf{u}_I, d\mathbf{x}_J).$$

When playing with the weight functions w, it is possible to keep the same weights for the bootstrapped versions, or to replace them with some functionals of F_n^*. For instance, asymptotically, it is equivalent to consider

$$\mathcal{T}^{(1),*}_{CvM,n} := \int \left(\hat{C}^*_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}^*_{s,I|J}(\mathbf{u}_I) + \hat{C}_{s,I|J}(\mathbf{u}_I)\right)^2 \hat{C}_n(d\mathbf{u}_I)\,\hat{F}_J(d\mathbf{x}_J), \text{ or}$$

$$\mathcal{T}^{(1),*}_{CvM,n} := \int \left(\hat{C}^*_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}^*_{s,I|J}(\mathbf{u}_I) + \hat{C}_{s,I|J}(\mathbf{u}_I)\right)^2 \hat{C}^*_n(d\mathbf{u}_I)\,\hat{F}^*_J(d\mathbf{x}_J).$$

Similarly, the limiting law of

$$\begin{aligned}\mathcal{T}^{(2),*}_{CvM,n} := \int \Big(&\hat{C}^*_{I|J}(\hat{F}^*_{n,1}(x_1|\mathbf{x}_J), \ldots, \hat{F}^*_{n,p}(x_p|\mathbf{x}_J)|\mathbf{x}_J) \\ &- \hat{C}_{I|J}(\hat{F}^*_{n,1}(x_1|\mathbf{x}_J), \ldots, \hat{F}^*_{n,p}(x_p|\mathbf{x}_J)|\mathbf{x}_J) - \hat{C}^*_{s,I|J}(\hat{F}^*_{n,1}(x_1|\mathbf{x}_J), \ldots, \hat{F}^*_{n,p}(x_p|\mathbf{x}_J)) \\ &+ \hat{C}_{s,I|J}(\hat{F}^*_{n,1}(x_1|\mathbf{x}_J), \ldots, \hat{F}^*_{n,p}(x_p|\mathbf{x}_J))\Big)^2 H_n(d\mathbf{x}_I, d\mathbf{x}_J),\end{aligned}$$

given F_n is unchanged replacing H_n by H_n^*.

The same ideas apply concerning the tests of Subsection 4.2.2, but they require some modifications. Let H be some cdf on $\mathbb{R}^d$. Denote by H_I and H_J the associated cdf on the first p and $d-p$ components respectively. Denote by $\hat{H}$, $\hat{H}_I$ and $\hat{H}_J$ their empirical counterparts. Under $\mathcal{H}_0$, and for any measurable subsets B_I and A_J, $H(B_I \times A_J) = H(B_I)H(A_J)$. Our tests will be based on the difference

$$\begin{aligned}\hat{H}(B_I \times A_J) - \hat{H}_I(B_I)\hat{H}_J(A_J) &= (\hat{H} - H)(B_I \times A_J) \\ &- (\hat{H}_I - H_I)(B_I)\hat{H}_J(A_J) - (\hat{H}_J - H_J)(A_J)H_I(B_I).\end{aligned}$$

Therefore, a bootstrapped approximation of the latter quantity will be

$$(\hat{H}^* - \hat{H})(B_I \times A_J) - (\hat{H}^*_I - \hat{H}_I)(B_I)\hat{H}^*_J(A_J) - (\hat{H}^*_J - \hat{H}_J)(A_J)\hat{H}_I(B_I).$$

To be specific, the bootstrapped versions of our tests are specified as below.

- Chi-square-type test of independence:

$$\mathcal{I}^*_{\chi,n} := n\sum_{k=1}^{N}\sum_{l=1}^{m}\frac{1}{\hat{G}^*_{I,J}(B_k\times\mathbb{R}^{d-p})\hat{G}^*_{I,J}(\mathbb{R}^p\times A_l)}\Big((\hat{G}^*_{I,J}-\hat{G}_{I,J})(B_k\times A_l)$$
$$-\quad \hat{G}^*_{I,J}(B_k\times\mathbb{R}^{d-p})\hat{G}^*_{I,J}(\mathbb{R}^p\times A_l)+\hat{G}_{I,J}(B_k\times\mathbb{R}^{d-p})\hat{G}_{I,J}(\mathbb{R}^p\times A_l)\Big)^2.$$

- Distance between distributions:

$$\mathcal{I}^*_{KS,n} = \sup_{\mathbf{x}\in\mathbb{R}^d}|(\hat{G}^*_{I,J}-\hat{G}_{I,J})(\mathbf{x})-\hat{G}^*_{I,J}(\mathbf{x}_I,\infty^{d-p})\hat{G}^*_{I,J}(\infty^p,\mathbf{x}_J)+\hat{G}_{I,J}(\mathbf{x}_I,\infty^{d-p})\hat{G}_{I,J}(\infty^p,\mathbf{x}_J)|$$

$$\mathcal{I}^*_{2,n} = \int\left((\hat{G}^*_{I,J}-\hat{G}_{I,J})(\mathbf{x})-\hat{G}^*_{I,J}(\mathbf{x}_I,\infty^{d-p})\hat{G}^*_{I,J}(\infty^p,\mathbf{x}_J)+\hat{G}_{I,J}(\mathbf{x}_I,\infty^{d-p})\hat{G}_{I,J}(\infty^p,\mathbf{x}_J)\right)^2\omega(\mathbf{x})\,d\mathbf{x},$$

and $\mathcal{I}^*_{CvM,n}$ is obtained replacing $\omega(\mathbf{x})\,d\mathbf{x}$ by $\hat{G}^*_{I,J}(d\mathbf{x})$ (or even $\hat{G}_{I,J}(d\mathbf{x})$).

- A test of independence based on the independence copula: Let $\breve{C}^*_{I,J}$, $\breve{C}^*_{I|J}$ and $\hat{C}^*_J$ be the empirical copulas based on a bootstrapped version of the pseudo-sample $(\hat{\mathbf{Z}}_{i,I|J},\mathbf{X}_{i,J})_{i=1,\dots,n}$, $(\hat{\mathbf{Z}}_{i,I|J})_{i=1,\dots,n}$ and $(\mathbf{X}_{i,J})_{i=1,\dots,n}$ respectively. This version can be obtained by nonparametric bootstrap, as usual, providing new vectors $\hat{\mathbf{Z}}^*_{i,I|J}$ at every draw. The associated bootstrapped statistics are

$$\breve{\mathcal{I}}^*_{KS,n} = \sup_{\mathbf{u}\in[0,1]^d}|(\breve{C}^*_{I,J}-\breve{C}_{I,J})(\mathbf{u})-\breve{C}^*_{I|J}(\mathbf{u}_I)\hat{C}^*_J(\mathbf{u}_J)+\breve{C}_{I|J}(\mathbf{u}_I)\hat{C}_J(\mathbf{u}_J)|,$$
$$\breve{\mathcal{I}}^*_{2,n} = \int_{\mathbf{u}\in[0,1]^d}\Big((\breve{C}^*_{I,J}-\breve{C}_{I,J})(\mathbf{u})-\breve{C}^*_{I|J}(\mathbf{u}_I)\hat{C}^*_J(\mathbf{u}_J)+\breve{C}_{I|J}(\mathbf{u}_I)\hat{C}_J(\mathbf{u}_J)\Big)^2\omega(\mathbf{u})\,d\mathbf{u},$$
$$\breve{\mathcal{I}}^*_{CvM,n} = \int_{\mathbf{u}\in[0,1]^d}\Big((\breve{C}^*_{I,J}-\breve{C}_{I,J})(\mathbf{u})-\breve{C}^*_{I|J}(\mathbf{u}_I)\hat{C}^*_J(\mathbf{u}_J)+\breve{C}_{I|J}(\mathbf{u}_I)\hat{C}_J(\mathbf{u}_J)\Big)^2\breve{C}^*_{I,J}(d\mathbf{u}).$$

In the case of the parametric statistics, the situation is pretty much the same, as long as we invoke the nonparametric bootstrap. For instance, the bootstrapped versions of some previous test statistics are

$$(\mathcal{T}^c_2)^* := \int\|\hat{\theta}^*(\mathbf{x}_J)-\hat{\theta}(\mathbf{x}_J)-\hat{\theta}^*_0+\hat{\theta}_0\|^2\omega(\mathbf{x}_J)\,d\mathbf{x}_J,\ \text{or}$$

$$(\mathcal{T}^c_{dens})^* := \int\Big(c_{\hat{\theta}^*(\mathbf{x}_J)}(\mathbf{u}_I)-c_{\hat{\theta}(\mathbf{x}_J)}(\mathbf{u}_I)-c_{\hat{\theta}^*_0}(\mathbf{u}_I)+c_{\hat{\theta}_0}(\mathbf{u}_I)\Big)^2\omega(\mathbf{u}_I,\mathbf{x}_J)d\mathbf{u}_I\,d\mathbf{x}_J.$$

in the case of the nonparametric bootstrap. We conjecture that the previous techniques can be applied with the other resampling schemes that have been proposed in Subsection 4.2.4.1. Nonetheless, a complete theoretical study of all these alternative schemes and the statement of the validity of their associated bootstrapped statistics is beyond the scope of this paper.

Remark 4.8. *For the "parametric independent" bootstrap scheme, we have observed that the test powers are a lot better by considering*

$$(\mathcal{T}^c_2)^{**} := \int\|\hat{\theta}^*(\mathbf{x}_J)-\hat{\theta}^*_0\|^2\omega(\mathbf{x}_J)\,d\mathbf{x}_J,\ \textit{or}$$

$$(\mathcal{T}^c_{dens})^{**} := \int\Big(c_{\hat{\theta}^*(\mathbf{x}_J)}(\mathbf{u}_I)-c_{\hat{\theta}^*_0}(\mathbf{u}_I)\Big)^2\omega(\mathbf{u}_I,\mathbf{x}_J)d\mathbf{u}_I\,d\mathbf{x}_J,$$

instead. The relevance of such statistics may be theoretically justified in the slightly different context of "box-type" tests in the next Section (see Theorem 4.14). Since our present case is close to the situation of "many small boxes", it is not surprising that we observe similar features. Note that, contrary to the nonparametric bootstrap or the "parametric conditional" bootstrap, the "parametric independent"

bootstrap scheme uses $\mathcal{H}_0$. More generally, and following the same idea, we found that using the statistic $\mathcal{T}^{**} := \psi\left(\sqrt{n}\left(\chi_s(F_n^*) - \chi(F_n^*)\right)\right)$ for the pseudo-independent bootstrap yields much better performance than $\mathcal{T}^*$. In our simulations, we will therefore use $\mathcal{T}^{**}$ as the bootstrap test statistic (see Figures 4.1 and 4.2).

Remark 4.9. *In a vine model, every node is associated with a bivariate conditional copula, and it is desirable that they satisfy $\mathcal{H}_0$. Unfortunately, the arguments of such copulas are defined through conditional distributions $F_i(X_i|\mathbf{X}_K)$ for some subsets $K \subset \{1,\dots,d\}$. Therefore, we do not observe realizations of such arguments, except for the first level. In practice, they have to be replaced with pseudo-observations in our previous test statistics. Their calculation involves the bivariate conditional copulas that are associated with the previous nodes in a recursive way. The theoretical analysis of the associated bootstrap schemes is challenging and falls beyond the scope of the current work.*

4.3 Tests with "boxes"

4.3.1 The link with the simplifying assumption

As we have seen in Remark 4.1, we do not have $C_{s,I|J} = C_I$ in general. This is the hint there are some subtle relations between conditional copulas when the conditioning event is pointwise or when it is a measurable subset. Actually, to test $\mathcal{H}_0$ in Section 4.2, we have relied on kernel estimates and smoothing parameters, at least to evaluate conditional marginal distributions empirically. To avoid the curse of dimension (when $d-p$ is "large" i.e. larger than three in practice), it is tempting to replace the pointwise conditioning events $\mathbf{X}_J = \mathbf{x}_J$ with $\mathbf{X}_J \in A_J$ for some borelian subsets $A_J \subset \mathbb{R}^{d-p}$, $\mathbb{P}(\mathbf{X}_J \in A_J) > 0$. As a shorthand notation, we shall write $\mathcal{A}_J$ the set of all such A_J. We call them "boxes" because choosing $d-p$-dimensional rectangles (i.e. intersections of half-spaces separated by orthogonal hyperplans) is natural, but our definitions are still valid for arbitrary borelian subsets in $\mathbb{R}^{d-p}$. Technically speaking, we will assume that the functions $\mathbf{x}_J \mapsto \mathbb{1}(\mathbf{x}_J \in A_J)$ are Donsker, to apply uniform CLTs' without any hurdle. Actually, working with $\mathbf{X}_J$-"boxes" instead of pointwise will simplify a lot the picture. Indeed, the evaluation of conditional cdfs' given $\mathbf{X}_J \in A_J$ does not require kernel smoothing, bandwidth choices, or other techniques of curve estimation that deteriorate the optimal rates of convergence.

Note that, by definition of the conditional copula of $\mathbf{X}_I$ given $(\mathbf{X}_J \in A_J)$, we have

$$\begin{aligned}&\mathbb{P}(\mathbf{X}_I \le \mathbf{x}_I | \mathbf{X}_J \in A_J)\\ &\quad = \quad C_{I|J}^{A_J}\left(\mathbb{P}(X_1 \le x_1|\mathbf{X}_J \in A_J), \dots, \mathbb{P}(X_p \le x_p|\mathbf{X}_J \in A_J)|\mathbf{X}_J \in A_J\right),\end{aligned}$$

for every point $\mathbf{x}_I \in \mathbb{R}^p$ and every subset A_J in $\mathcal{A}_J$. So, it is tempting to replace $\mathcal{H}_0$ by

$$\widetilde{\mathcal{H}}_0 : C_{I|J}^{A_J}(\mathbf{u}_I|\mathbf{X}_J \in A_J) \text{ does not depend on } A_J \in \mathcal{A}_J, \text{for any } \mathbf{u}_I.$$

For any $\mathbf{x}_J$, consider a sequence of boxes $(A_J^{(n)}(\mathbf{x}_J))$ s.t. $\cap_n A_J^{(n)}(\mathbf{x}_J) = \{\mathbf{x}_J\}$. If the law of $\mathbf{X}$ is sufficiently regular, then $\lim_n C_{I|J}^{A_J^{(n)}}(\mathbf{u}_I|\mathbf{X}_J \in A_J^{(n)}) = C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J)$ for any $\mathbf{u}_I$. Therefore, $\tilde{\mathcal{H}}_0$ implies $\mathcal{H}_0$. This is stated formally in the next proposition.

Proposition 4.10. *Assume that the function $h : \mathbb{R}^d \to [0,1]$, defined by $h(\mathbf{y}) := \mathbb{P}(\mathbf{X}_I \leq \mathbf{y}_I | \mathbf{X}_J = \mathbf{y}_J)$ is continuous everywhere. Let $\mathbf{x}_J \in \mathbb{R}^{d-p}$ such that $F_{i|J}(\cdot|\mathbf{x}_J)$ is strictly increasing for every $i = 1, \ldots, d$. Then, for any sequence of boxes $(A_J^{(n)}(\mathbf{x}_J))$ such that $\cap_n A_J^{(n)}(\mathbf{x}_J) = \{\mathbf{x}_J\}$, we have*

$$\lim_n C_{I|J}^{A_J^{(n)}(\mathbf{x}_J)}(\mathbf{u}_I | \mathbf{X}_J \in A_J^{(n)}(\mathbf{x}_J)) = C_{I|J}(\mathbf{u}_I | \mathbf{X}_J = \mathbf{x}_J),$$

for every $\mathbf{u}_I \in [0,1]^p$.

Proof: Consider a particular $\mathbf{u}_I \in [0,1]^p$. If one component of $\mathbf{u}_I$ is zero, the result is obviously satisfied. If one component of $\mathbf{u}_I$ is one, this component does not play any role. Therefore, we can restrict ourselves on $\mathbf{u}_I \in (0,1)^p$. By continuity, there exists $\mathbf{x}_I \in \mathbb{R}^p$ s.t. $u_i = F_{i|J}(x_i|\mathbf{x}_J)$ for every $i = 1, \ldots, p$. Let the sequences $(x_i^{(n)})$ such that $u_i = F_{i|J}(x_i^{(n)} | \mathbf{X}_J \in A_J^{(n)})$ for every n and every $i = 1, \ldots, p$. First, let us show that $x_i^{(n)} \to x_i$ when n tends to the infinity. Indeed, by the definition of conditional probabilities ([124], p.220), we have

$$u_i = \mathbb{P}(X_i \leq x_i^{(n)} | \mathbf{X}_J \in A_J^{(n)}) = \frac{1}{\mathbb{P}(\mathbf{X}_J \in A_J^{(n)})} \int_{\{\mathbf{y}_J \in A_J^{(n)}\}} \mathbb{P}(X_i \leq x_i^{(n)} | \mathbf{X}_J = \mathbf{y}_J)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{y}_J),$$

and

$$\begin{aligned} u_i = \mathbb{P}(X_i \leq x_i | \mathbf{X}_J = \mathbf{x}_J) = & \frac{1}{\mathbb{P}(\mathbf{X}_J \in A_J^{(n)})} \int_{\{\mathbf{y}_J \in A_J^{(n)}\}} \mathbb{P}(X_i \leq x_i^{(n)} | \mathbf{X}_J = \mathbf{x}_J)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{y}_J) \\ & + \mathbb{P}(X_i \leq x_i | \mathbf{X}_J = \mathbf{x}_J) - \mathbb{P}(X_i \leq x_i^{(n)} | \mathbf{X}_J = \mathbf{x}_J). \end{aligned}$$

By substracting the two latter identities, we deduce

$$\begin{aligned} & \frac{1}{\mathbb{P}(\mathbf{X}_J \in A_J^{(n)})} \int_{\{\mathbf{y}_J \in A_J^{(n)}\}} \left[\mathbb{P}(X_i \leq x_i^{(n)} | \mathbf{X}_J = \mathbf{y}_J) - \mathbb{P}(X_i \leq x_i^{(n)} | \mathbf{X}_J = \mathbf{x}_J) \right] d\mathbb{P}_{\mathbf{X}_J}(\mathbf{y}_J) \\ = \quad & \mathbb{P}(X_i \leq x_i | \mathbf{X}_J = \mathbf{x}_J) - \mathbb{P}(X_i \leq x_i^{(n)} | \mathbf{X}_J = \mathbf{x}_J). \end{aligned} \tag{4.22}$$

But, by assumption, $F_{i|J}(t|\mathbf{y}_J)$ tends towards $F_{i|J}(t|\mathbf{x}_J)$ when $\mathbf{y}_J$ tends to $\mathbf{x}_J$, for any t (pointwise convergence). Actually, the latter convergence is uniform on $\mathbb{R}$: $\|F_{i|J}(\cdot|\mathbf{y}_J) - F_{i|J}(\cdot|\mathbf{x}_J)\|_\infty$ tends to zero when $\mathbf{y}_J \to \mathbf{x}_J$. This is a straightforward consequence of Pólya's Theorem (also called second Dini's Theorem in the literature): see Subsection (A.1) in [18] for instance. From (4.22), we deduce that $\mathbb{P}(X_i \leq x_i^{(n)} | \mathbf{X}_J = \mathbf{x}_J) \to \mathbb{P}(X_i \leq x_i | \mathbf{X}_J = \mathbf{x}_J)$. By the continuity of $F_{i|J}(\cdot|\mathbf{x}_J)$, we get $x_i^{(n)} \to x_i$, for any $i = 1, \ldots, p$.

Second, let us come back to conditional copulas: setting $\mathbf{x}_I^{(n)} := (x_1^{(n)}, \ldots, x_p^{(n)})$, we have

$$\begin{aligned} & C_{I|J}^{A_J^{(n)}}(\mathbf{u}_I | A_J^{(n)}) - C_{I|J}(\mathbf{u}_I | \mathbf{x}_J) \\ = \quad & C_{I|J}^{A_J^{(n)}}(F_{1|J}(x_1^{(n)}|A_J^{(n)}), \ldots, F_{p|J}(x_p^{(n)}|A_J^{(n)}) | A_J^{(n)}) - C_{I|J}(F_{1|J}(x_1|\mathbf{x}_J), \ldots, F_{p|J}(x_p|\mathbf{x}_J) | \mathbf{x}_J) \\ = \quad & F_{I|J}(\mathbf{x}_I^{(n)} | A_J^{(n)}) - F_{I|J}(\mathbf{x}_I | \mathbf{x}_J) \\ = \quad & \frac{1}{\mathbb{P}(\mathbf{X}_J \in A_J^{(n)})} \int_{\{\mathbf{y}_J \in A_J^{(n)}\}} \left[\mathbb{P}(\mathbf{X}_I \leq \mathbf{x}_I^{(n)} | \mathbf{X}_J = \mathbf{y}_J) - \mathbb{P}(\mathbf{X}_I \leq \mathbf{x}_I | \mathbf{X}_J = \mathbf{x}_J) \right] d\mathbb{P}_{\mathbf{X}_J}(\mathbf{y}_J). \end{aligned}$$

Since $\mathbf{x}_I^{(n)}$ tends to $\mathbf{x}_I$ when $n \to \infty$ and invoking the continuity of h at $(\mathbf{x}_I, \mathbf{x}_J)$, we get $C_{I|J}^{A_J^{(n)}}(\mathbf{u}_I | A_J^{(n)}) \to C_{I|J}(\mathbf{u}_I | \mathbf{x}_J)$ when $n \to \infty$. □

Unfortunately, the opposite is false. Counter-intuitively, $\tilde{\mathcal{H}}_0$ does not lead to a consistent test of the simplifying assumption. Indeed, under $\mathcal{H}_0$, we can see that $C_{I|J}^{A_J}(\mathbf{u}_I | \mathbf{X}_J \in A_J)$ **depends on** A_J in general, even if $C_{I|J}(\mathbf{u}_I | \mathbf{X}_J = \mathbf{x}_J)$ **does not depend on** $\mathbf{x}_J$!

This is due to the nonlinear transform between conditional (univariate and multivariate) distributions and conditional copulas. In other words, for a usual d-dimensional cdf H, we have

$$H(\mathbf{x}_I|\mathbf{X}_J \in A_J) = \frac{1}{\mathbb{P}(A_J)} \int_{A_J} H(\mathbf{x}_I|\mathbf{X}_J = \mathbf{x}_J)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J), \tag{4.23}$$

for every measurable subset $A_J \in \mathcal{A}_J$ and $\mathbf{x}_I \in \mathbb{R}^p$. At the opposite and in general, for conditional copulas,

$$C_{I|J}^{A_J}(\mathbf{u}_I|\mathbf{X}_J \in A_J) \neq \frac{1}{\mathbb{P}(A_J)} \int_{A_J} C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J), \tag{4.24}$$

for $\mathbf{u}_I \in [0,1]^p$. And even if we assume $\mathcal{H}_0$, we have in general,

$$C_{I|J}^{A_J}(\mathbf{u}_I|\mathbf{X}_J \in A_J) \neq \frac{1}{\mathbb{P}(A_J)} \int_{A_J} C_{s,I|J}(\mathbf{u}_I)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J) = C_{s,I|J}(\mathbf{u}_I). \tag{4.25}$$

As a particular case, taking $A_J = \mathbb{R}^{d-p}$, this means again that $C_I(\mathbf{u}_I) \neq C_{s,I|J}(\mathbf{u}_I)$.

Let us check this rather surprising feature with the example of Remark 4.1 for another subset A_J. Recall that $\mathcal{H}_0$ is true and that $C_{s,1,2|3}(u,v) = uv$ for every $u, v \in [0,1]$. Consider the subset $(X_3 \leq a)$, for any real number a. The probability of this event is $\Phi(a)$. Now, let us verify that

$$uv \neq H(F_{1|3}^-(u|X_3 \leq a), F_{2|3}^-(v|X_3 \leq a)|X_3 \leq a),$$

for some u, v in $(0,1)$. Clearly, for every real number x_k, we have

$$\mathbb{P}(X_k \leq x_k|X_3 \leq a) = \frac{1}{\Phi(a)} \int_{-\infty}^{a} \Phi(x_k - z)\phi(z)\, dz, k = 1, 2, \text{ and}$$

$$\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2|X_3 \leq a) = \frac{1}{\Phi(a)} \int_{-\infty}^{a} \Phi(x_1 - z)\Phi(x_2 - z)\phi(z)\, dz.$$

In particular, $\mathbb{P}(X_k \leq 0|X_3 \leq a) = (1 + \Phi(-a))/2$. Therefore, set $u^* = v^* = (1 + \Phi(-a))/2$ and we get

$$\begin{aligned} & H(F_{1|3}^-(u^*|X_3 \leq a), F_{2|3}^-(v^*|X_3 \leq a)|X_3 \leq a) = H(0,0|X_3 \leq a) \\ = \quad & \frac{1}{3}\left(1 + \Phi(-a) + \Phi^2(-a)\right) \neq u^* v^*. \end{aligned}$$

In this example, $C_{s,1,2|3}(\cdot) \neq C_{1,2|3}^{]-\infty,a]}(\cdot|X_3 \leq a)$, for every a, even if $\mathcal{H}_0$ is satisfied.

Nonetheless, getting back to the general case, we can easily provide an equivalent of Equation (4.23) for general conditional copulas, i.e. without assuming $\mathcal{H}_0$.

Proposition 4.11. *For all $\mathbf{u}_I \in [0,1]^p$ and all $A_J \in \mathcal{A}_J$,*

$$C_{I|J}^{A_J}(\mathbf{u}_I|\mathbf{X}_J \in A_J) = \frac{1}{\mathbb{P}(A_J)} \int_{A_J} \psi(\mathbf{u}_I, \mathbf{x}_J, A_J) d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J), \text{ with}$$

$$\begin{aligned} & \psi(\mathbf{u}_I, \mathbf{x}_J, A_J) \\ & = C_{I|J}\left(F_{1|J}\Big(F_{1|J}^-(u_1|\mathbf{X}_J \in A_J)\Big|\mathbf{X}_J = \mathbf{x}_J\Big), \ldots, F_{p|J}\Big(F_{p|J}^-(u_p|\mathbf{X}_J \in A_J)\Big|\mathbf{X}_J = \mathbf{x}_J\Big) \,\middle|\, \mathbf{X}_J = \mathbf{x}_J \right). \end{aligned}$$

Proof: From (4.23), we get :

$$\begin{aligned}
&H(\mathbf{x}_I|\mathbf{X}_J \in A_J) \\
&= \frac{1}{\mathbb{P}(A_J)} \int_{A_J} H(\mathbf{x}_I|\mathbf{X}_J = \mathbf{x}_J)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J) \\
&= \frac{1}{\mathbb{P}(A_J)} \int_{A_J} C_{I|J}\Big(F_{1|J}(x_1|\mathbf{X}_J = \mathbf{x}_J), \ldots, F_{p|J}(x_p|\mathbf{X}_J = \mathbf{x}_J) \,\Big|\, \mathbf{X}_J = \mathbf{x}_J\Big)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J).
\end{aligned}$$

We can conclude by using the following definition of the conditional copula

$$C_{I|J}^{A_J}(\mathbf{u}_I|\mathbf{X}_J \in A_J) = H(F_{1|J}^{-}(u_1|\mathbf{X}_J \in A_J), \ldots, F_{p|J}^{-}(u_p|\mathbf{X}_J \in A_J)|\mathbf{X}_J \in A_J). \ \square$$

Now, we understand why (4.24) (and (4.25) under $\mathcal{H}_0$) are not identities: the conditional copulas, given the subset A_J, still depend on the conditional margins of $\mathbf{X}_I$ given $\mathbf{X}_J$ pointwise in general.

Note that, if X_i is independent of $\mathbf{X}_J$ for every $i = 1, \ldots, p$, then, for any such i,

$$F_{i|J}\Big(F_{i|J}^{-}(u_i|\mathbf{X}_J \in A_J)\big|\mathbf{X}_J = \mathbf{x}_J\Big) = F_i\Big(F_i^{-}(u_i)\Big) = u_i,$$

and we can revisit the identity of Proposition 4.11: under $\mathcal{H}_0$, we have

$$\begin{aligned}
C_{I|J}^{A_J}(\mathbf{u}_I|\mathbf{X}_J \in A_J) &= \frac{1}{\mathbb{P}(A_J)} \int_{A_J} C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J) \\
&= \frac{1}{\mathbb{P}(A_J)} \int_{A_J} C_{s,I|J}(\mathbf{u}_I)\, d\mathbb{P}_{\mathbf{X}_J}(\mathbf{x}_J) = C_{s,I|J}(\mathbf{u}_I).
\end{aligned}$$

This means $\mathcal{H}_0$ and $\tilde{\mathcal{H}}_0$ are equivalent. We consider such circumstances as very peculiar and do not have to be confused with a test of $\mathcal{H}_0$. Therefore, we advise to lead a preliminary test of independence between $\mathbf{X}_I$ and $\mathbf{X}_J$ (or at least between X_i and $\mathbf{X}_J$ for any $i = 1, \ldots, p$) before trying to test $\mathcal{H}_0$ itself.

Now, let us revisit the characterisation of $\mathcal{H}_0$ in terms of the independence property, as in Subsection 4.2.2. The latter analysis is confirmed by the equivalent of Proposition 4.4 in the case of conditioning subsets A_J. Now, the relevant random vector would be

$$\mathbf{Z}_{I|A_J} := \big(F_{1|J}(X_1|\mathbf{X}_J \in A_J), \ldots, F_{p|J}(X_p|\mathbf{X}_J \in A_J)\big),$$

that has straightforward empirical counterparts. Then, it is tempting to test

$$\widetilde{\mathcal{H}}_0^* : \mathbf{Z}_{I|A_J} \text{ and } (X_J \in A_J) \text{ are independent for every borelian subset } A_J \subset \mathbb{R}^{d-p}.$$

Nonetheless, it can be proved easily that this is not a test of $\mathcal{H}_0$, unfortunately.

Proposition 4.12. *$\mathbf{Z}_{I|A_J}$ and $(X_J \in A_J)$ are independent for every measurable subset $A_J \subset \mathbb{R}^{d-p}$ iff $\mathbf{X}_I$ and $\mathbf{X}_J$ are independent.*

Proof: For any measurable subset A_J and any $\mathbf{u}_I \in [0,1]^p$, under $\widetilde{\mathcal{H}}_0^*$, we have

$$\mathbb{P}\left(\mathbf{Z}_{I|A_J} \leq \mathbf{u}_I, \mathbf{X}_J \in A_J\right) = \mathbb{P}\left(\mathbf{Z}_{I|A_J} \leq \mathbf{u}_I\right) \mathbb{P}(\mathbf{X}_J \in A_J).$$

Consider $\mathbf{x}_I \in \mathbb{R}^p$. Due to the continuity of the conditional cdfs', there exists u_k s.t. $F_k(x_k|\mathbf{X}_J \in A_J) = u_k$, $k = 1, \ldots, p$. Then, using the invertibility of $x \mapsto F_k(x|\mathbf{X}_J \in A_J)$, we get $\mathbb{P}\left(\mathbf{Z}_{I|A_J} \leq \mathbf{u}_I, \mathbf{X}_J \in A_J\right) = \mathbb{P}\left(\mathbf{X}_I \leq \mathbf{x}_I, \mathbf{X}_J \in A_J\right)$. This implies that $\widetilde{\mathcal{H}}_0^*$ is equivalent to the following property: for every $\mathbf{x}_I \in \mathbb{R}^p$ and A_J,

$$\mathbb{P}\left(X_I \leq \mathbf{x}_I, \mathbf{X}_J \in A_J\right) = \mathbb{P}\left(\mathbf{X}_I \leq \mathbf{x}_I\right) \mathbb{P}\left(\mathbf{X}_J \in A_J\right). \square$$

The previous result shows that a test of $\tilde{\mathcal{H}}_0^*$ is a test of independence between $\mathbf{X}_I$ and $\mathbf{X}_J$. When the latter assumption is satisfied, $\tilde{\mathcal{H}}_0$ and then $\mathcal{H}_0$ are true too, but the opposite is false.

Previously, we have exhibited a simple trivariate model where $\mathcal{H}_0$ is satisfied when $\mathbf{X}_I$ and $\mathbf{X}_J$ are not independent. Then, we see that it is not reasonable to test whether the mapping $A_J \mapsto C_{I|J}^{A_J}(\cdot|\mathbf{X}_J \in A_J)$ is constant over $\mathcal{A}_J$, the set of **all** A_J such that $\mathbb{P}_{\mathbf{X}_J}(A_J) > 0$, with the idea of testing $\mathcal{H}_0$.

Nonetheless, one can weaken the latter assumption, and restrict oneself to a **finite** family $\overline{\mathcal{A}}_J$ of subsets with positive probabilities. For such a family, we could test the assumption

$$\overline{\mathcal{H}}_0 : A_J \mapsto C_{I|J}^{A_J}(\cdot\,|\mathbf{X}_J \in A_J) \text{ is constant over } \overline{\mathcal{A}}_J.$$

To fix the ideas and w.l.o.g., we will consider a given family of disjoint subsets $\overline{\mathcal{A}}_J = \{A_{1,J}, \ldots, A_{m,J}\}$ in $\mathbb{R}^{d-p}$ hereafter. Note the following consequence of Proposition 4.11.

Proposition 4.13. *Assume that, for all $A_J \in \overline{\mathcal{A}}_J$ and for all $i \in I$,*

$$F_{i|J}(x|\mathbf{X}_J = \mathbf{x}_J) = F_{i|J}(x|\mathbf{X}_J \in A_J), \quad \forall \mathbf{x}_J \in A_J, x \in \mathbb{R}. \tag{4.26}$$

Then, $\mathcal{H}_0$ implies $\overline{\mathcal{H}}_0$.

Obviously, if the family $\overline{\mathcal{A}}_J$ is too big, then (4.26) will be too demanding: $\overline{\mathcal{H}}_0$ will be close to a test of independence between $\mathbf{X}_I$ and $\mathbf{X}_J$, and no longer a test of $\mathcal{H}_0$. Moreover, the chosen subsets in the family $\overline{\mathcal{A}}_J$ do not need to be disjoint, even if this would be a natural choice. As a special case, if $\mathbb{R}^{d-p} \in \overline{\mathcal{A}}_J$, the previous condition is equivalent to the independence between X_i and $\mathbf{X}_J$ for every $i \in I$.

Note that (4.26) does not imply that the vector of explanatory variables $\mathbf{X}_J$ should be discretized. Indeed, the full model requires the specification of the underlying conditional copula too, independently of the conditional margins and arbitrarily. For instance, we can choose a Gaussian conditional copula whose parameter is a continuous function of $\mathbf{X}_J$, even if (4.26) is fulfilled. And the law of $\mathbf{X}_I$ given $\mathbf{X}_J$ will depend on the current value of X_J.

A test of $\overline{\mathcal{H}}_0$ may be relevant in a lot of situations, beside technical arguments as the absence of smoothing. First, the case of discrete (or discretized) explanatory variables $\mathbf{X}_J$ is frequent. When $\mathbf{X}_J$ is discrete and takes a value among $\{\mathbf{x}_{1,J}, \ldots, \mathbf{x}_{m,J}\}$, set $A_{k,J} = \{\mathbf{x}_{k,J}\}$, $k = 1, \ldots, m$. Then, there is identity between testing $\mathcal{H}_0$ and $\overline{\mathcal{H}}_0$, with $\overline{\mathcal{A}}_J = \{A_{1,J}, \ldots, A_{m,J}\}$. Second, the level of precision and sharpness of a copula model is often lower than the models for (conditional) margins. To illustrate this idea, a lot of complex and subtle models to explain the dynamics of asset volatilities are available when the dynamics of cross-assets dependencies are often a lot more basic and without clear-cut empirical findings. Therefore, it makes sense to simplify conditional copula models compared to conditional marginal models. This can be done by considering only a few possible conditional copulas, associated to some events $(\mathbf{X}_J \in A_{k,J})$, $k = 1, \ldots, m$. For example, Jondeau and Rockinger [78] (the first paper that introduced conditional dependence structures, beside Patton [112]) proposed a Gaussian copula parameter that take a finite of values randomly, based on the realizations of some past asset returns. Third, similar situation occur with most Markov-switching copula models, where a finite set of copulas is managed. In such models, the (unobservable, in general) underlying state of the economy determines the index of the box: see [31],[142],[129],[53], among others.

Therefore, testing $\overline{\mathcal{H}}_0$ is of interest per se. Even if this is not equivalent to $\mathcal{H}_0$ (i.e. the simplifying assumption) formally, the underlying intuitions are close. And, particularly when the components of

the conditioning variable $\mathbf{X}_J$ are numerous, it can make sense to restrict the information set of the underlying conditional copula to a fixed number of conveniently chosen subsets A_J. And the constancy of the underlying copula when $\mathbf{X}_J$ belongs to such subsets is valuable in a lot of practical situations. Therefore, in the next subsections, we study some specific tests of $\overline{\mathcal{H}}_0$ itself.

4.3.2 Non-parametric tests with "boxes"

To specify such tests, we need first to estimate the conditional marginal cdfs', for instance by

$$\hat{F}_{k|J}(x|\mathbf{X}_J \in A_J) := \frac{\sum_{i=1}^n \mathbb{1}(X_{i,k} \leq x, \mathbf{X}_{i,J} \in A_J)}{\sum_{i=1}^n \mathbb{1}(\mathbf{X}_{i,J} \in A_J)},$$

for every real x and $k = 1, \ldots, p$. Similarly the joint law of $\mathbf{X}_I$ given $(\mathbf{X}_J \in A_J)$ may be estimated by

$$\hat{F}_{I|J}(\mathbf{x}_I|\mathbf{X}_J \in A_J) := \frac{\sum_{i=1}^n \mathbb{1}(X_{i,I} \leq \mathbf{x}_I, \mathbf{X}_{i,J} \in A_J)}{\sum_{i=1}^n \mathbb{1}(\mathbf{X}_{i,J} \in A_J)}.$$

The conditional copula given $(\mathbf{X}_J \in A_J)$ will be estimated by

$$\hat{C}_{I|J}^{A_J}(\mathbf{u}_I|\mathbf{X}_J \in A_J) := \hat{F}_{I|J}(\hat{F}_{1|J}^{-}(u_1|\mathbf{X}_J \in A_J), \ldots, \hat{F}_{p|J}^{-}(u_p|\mathbf{X}_J \in A_J)|\mathbf{X}_J \in A_J).$$

Therefore, it is easy to imagine tests of $\overline{\mathcal{H}}_0$, for instance

$$\overline{\mathcal{T}}_{KS,n} := \sup_{\mathbf{u}_I \in [0,1]^d} \sup_{k,l=1,\ldots,m} |\hat{C}_{I|J}^{A_{k,J}}(\mathbf{u}_I|\mathbf{X}_J \in A_{k,J}) - \hat{C}_{I|J}^{A_{l,J}}(\mathbf{u}_I|\mathbf{X}_J \in A_{l,J})|, \tag{4.27}$$

$$\overline{\mathcal{T}}_{CvM,n} := \sum_{k,l=1}^m \int \left(\hat{C}_{I|J}^{A_{k,J}}(\mathbf{u}_I|\mathbf{X}_J \in A_{k,J}) - \hat{C}_{I|J}^{A_{l,J}}(\mathbf{u}_I|\mathbf{X}_J \in A_{l,J})\right)^2 w(d\mathbf{u}_I), \tag{4.28}$$

for some nonnegative weight functions w, or even

$$\overline{\mathcal{T}}_{dist,n} := \sum_{k,l=1}^m dist\left(\hat{C}_{I|J}^{A_{k,J}}(\cdot|\mathbf{X}_J \in A_{k,J}), \hat{C}_{I|J}^{A_{l,J}}(\cdot|\mathbf{X}_J \in A_{l,J})\right), \tag{4.29}$$

where $dist(\cdot,\cdot)$ denotes a distance between cdfs' on $[0,1]^p$. More generally, define the matrix

$$\widehat{M}(\overline{\mathcal{A}}_J) := \left[\mathbb{1}(k \neq l)\, dist\Big(\hat{C}_{I|J}^{A_{k,J}}(\cdot|\mathbf{X}_J \in A_{k,J}), \hat{C}_{I|J}^{A_{l,J}}(\cdot|\mathbf{X}_J \in A_{l,J})\Big)\right]_{1\leq k,l\leq m},$$

and any statistic of the form $||\widehat{M}(\overline{\mathcal{A}}_J)||$ can be used as a test statistics of $\overline{\mathcal{H}}_0$, when $||\cdot||$ is a norm on the set of $m \times m$-matrices. Obviously, it is easy to introduce similar statistics based on copula densities instead of cdfs'.

4.3.3 Parametric test statistics with "boxes"

When we work with subsets $A_J \in \mathbb{R}^{d-p}$ instead of pointwise conditioning events $(\mathbf{X}_J = \mathbf{x}_J)$, we can adapt all the previous parametric test statistics of Subsection 4.2.3. Nonetheless, the framework will be slightly modified.

Let us assume that, for every $A_J \in \overline{\mathcal{A}}_J$, $C_{I|J}^{A_J}(\cdot|\mathbf{X}_J \in A_J)$ belongs to the same parametric copula family $\mathcal{C} = \{C_\theta, \theta \in \Theta\}$. In other words, $C_{I|J}^{A_J}(\cdot|\mathbf{X}_J \in A_J) = C_{\theta(A_J)}(\cdot)$ for every $A_J \in \overline{\mathcal{A}}_J$. Therefore, we could test the constancy of the mapping $A_J \mapsto \theta(A_J)$, i.e. to test

$$\overline{\mathcal{H}}_0^c : \text{the function } k \in \{1, \ldots, m\} \mapsto \theta(A_{k,J}) \text{ is a constant called } \theta_0^b.$$

Clearly, for every $A_J \in \overline{\mathcal{A}}_J$, we can estimate $\theta(A_J)$ by

$$\hat{\theta}(A_J) := \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log c_\theta \left(\hat{F}_{1|J}(X_{i,1}|\mathbf{X}_{i,J} \in A_J), \ldots, \hat{F}_{p|J}(X_{i,p}|\mathbf{X}_{i,J} \in A_J) \right) \mathbb{1}(\mathbf{X}_{i,J} \in A_J).$$

It can be proved that the estimate $\hat{\theta}(A_J)$ is consistent and asymptotically normal, by revisiting the proof of Theorem 1 in [137]. Here, the single difference w.r.t. the latter paper is induced by the random sample size, modifying the limiting distributions. The proof is left to the reader.

Under the zero assumption $\overline{\mathcal{H}}_0^c$, the parameter of the copula of $(F_1(X_1|\mathbf{X}_J \in A_{k,J}), \ldots, F_p(X_p|\mathbf{X}_J \in A_{k,J}))$ given $(\mathbf{X}_J \in A_{k,J})$ is the same for any $k = 1, \ldots, m$. It will be denoted by θ_0^b, and we can still estimate it by the semi-parametric procedure

$$\hat{\theta}_0^b := \arg\max_{\theta\in\Theta} \sum_{k=1}^{m} \sum_{i=1}^{n} \log c_\theta \left(\hat{F}_{1|J}(X_{i,1}|\mathbf{X}_{i,J} \in A_{k,J}), \ldots, \hat{F}_{p|J}(X_{i,p}|\mathbf{X}_{i,J} \in A_{k,J}) \right) \mathbb{1}(\mathbf{X}_{i,J} \in A_{k,J}).$$

Obviously, under some conditions of regularity and under $\overline{\mathcal{H}}_0^c$, it can be proved that $\hat{\theta}_0^b$ is consistent and asymptotically normal, by adapting the results of [137].

For convenience, let us define the "box index" function $k(\mathbf{x}_J) := \sum_{k=1}^{m} k\mathbb{1}\{\mathbf{x}_J \in A_{k,J}\}$, for any $\mathbf{x}_J \in \mathbb{R}^{d-p}$. In other words, k is the index of the box $A_{k,J}$ that contains $\mathbf{x}_J$. It equals zero, when no box in $\overline{\mathcal{A}}_J$ contains $\mathbf{x}_J$. Let us introduce the r.v. $Y_i := k(\mathbf{X}_{i,J})$, that stores only all the needed information concerning the conditioning with respect to the variables $\mathbf{X}_{i,J}$. We can then define the empirical pseudo-observations as

$$\begin{aligned}
\mathbf{Z}_{i,I|Y} &:= \sum_{k=1}^{m} \left(F_{1|J}(X_{i,1}|\mathbf{X}_J \in A_{k,J}), \ldots, F_{p|J}(X_{i,p}|\mathbf{X}_J \in A_{k,J}) \right) \mathbb{1}\{\mathbf{X}_{i,J} \in A_{k,J}\} \\
&= \left(F_{1|J}(X_{i,1}|\mathbf{X}_J \in A_{k(\mathbf{X}_{i,J}),J}), \ldots, F_{p|J}(X_{i,p}|\mathbf{X}_J \in A_{k(\mathbf{X}_{i,J}),J}) \right) \\
&= \left(F_{1|Y}(X_{i,1}|Y_i), \ldots, F_{p|Y}(X_{i,p}|Y_i) \right),
\end{aligned}$$

for any $i = 1, \ldots, n$. Since we do not observe the conditional marginal cdfs', we define the observed pseudo-observations that we calculate in practice: for $i = 1, \ldots, n$,

$$\hat{\mathbf{Z}}_{i,I|Y} := \left(\hat{F}_{1|J}(X_{i,1}|\mathbf{X}_J \in A_{Y_i,J}), \ldots, \hat{F}_{p|J}(X_{i,p}|\mathbf{X}_J \in A_{Y_i,J}) \right).$$

Note that we can then rewrite the previous estimators as

$$\hat{\theta}(A_{k,J}) = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log c_\theta \left(\hat{\mathbf{Z}}_{i,I|Y} \right) \mathbb{1}(Y_i = k), \text{ and } \hat{\theta}_0^b = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log c_\theta \left(\hat{\mathbf{Z}}_{i,I|Y} \right).$$

Now, let us revisit some of the previously proposed test statistics in the case of "boxes".

- Tests based on the comparison between $\hat{\theta}(\cdot)$ and $\hat{\theta}_0$:

$$\overline{\mathcal{T}}_\infty^c := \sqrt{n} \max_{k=1,\ldots,m} \|\hat{\theta}(A_{k,J}) - \hat{\theta}_0\|, \; \overline{\mathcal{T}}_2^c := n \sum_{k=1}^{m} \|\hat{\theta}(A_{k,J}) - \hat{\theta}_0\|^2 \omega_k, \tag{4.30}$$

  for some weights ω_k.

- Tests based on the comparison between $C_{\hat{\theta}(\cdot)}$ and $C_{\hat{\theta}_0}$:

$$\overline{\mathcal{T}}_{dist}^c := \sum_{k=1}^{m} dist(C_{\hat{\theta}(A_k)}, C_{\hat{\theta}_0}) \omega_k, \tag{4.31}$$

and others.

4.3.4 Bootstrap techniques for tests with boxes

In the same way as in the previous section, we will need bootstrap schemes to evaluate the limiting laws of the test statistics of $\overline{\mathcal{H}}_0$ or $\overline{\mathcal{H}}_0^c$ under the null. All the nonparametric resampling schemes of Subsection 4.2.4.1 (in particular Efron's usual bootstrap) can be used in this framework, replacing the conditional pseudo-observations $\hat{\mathbf{Z}}_{i,I|J}$ by $\hat{\mathbf{Z}}_{i,I|Y}$, $i = 1, \ldots, n$. The parametric resampling schemes of Subsection 4.2.4.1 can also be applied to the framework of "boxes", replacing $\hat{\theta}_0$ by $\hat{\theta}_0^b$ and $\hat{\theta}(\mathbf{x}_J)$ by $\hat{\theta}(A_J)$. In the parametric case, the bootstrapped estimates are denoted by $\hat{\theta}_0^*$ and $\hat{\theta}^*(A_J)$. They are the equivalents of $\hat{\theta}_0^b$ and $\hat{\theta}_n(A_J)$, replacing $(\hat{\mathbf{Z}}_{i,I|J}, Y_i)$ by $(\mathbf{Z}_i^*, Y_i^*)$.

The bootstrapped statistics will also be changed accordingly. Writing them explicitly is a rather straightforward exercise and we do not provide the details, contrary to Subsection 4.2.4. For example, the bootstrapped statistics corresponding to (4.30) is

$$\left(\overline{\mathcal{T}}_2^c\right)^* := n \sum_{k=1}^{m} \|\hat{\theta}^*(A_{k,J}) - \hat{\theta}(A_{k,J}) - \hat{\theta}_0^* + \hat{\theta}_0^b\|^2 \omega_k,$$

where $\hat{\theta}_0^*$ is the result of the program $\arg\max_\theta \sum_{i=1}^n \log c_\theta \left(\hat{\mathbf{Z}}_{i,I|Y}^*\right)$, in the case of Efron's nonparametric bootstrap.

As we noticed in Remark 4.8, some changes are required when dealing with the "parametric independent" bootstrap. Indeed, under the alternative, we observe $\hat{\theta}^*(A_{k,J}) - \hat{\theta}_0^* \approx 0$, because we have precisely generated a bootstrap sample under $\overline{\mathcal{H}}_0^c$. As a consequence, the law of $\left(\overline{\mathcal{T}}_2^c\right)^*$ would be close to the law of $\overline{\mathcal{T}}_2^c$ but under the alternative, providing very small powers. Therefore, convenient bootstrapped test statistics of $\overline{\mathcal{H}}_0$ under the "parametric independent" scheme will be of the type

$$\left(\overline{\mathcal{T}}_2^c\right)^{**} := n \sum_{k=1}^{m} \|\hat{\theta}^*(A_{k,J}) - \hat{\theta}_0^*\|^2 \omega_k.$$

Such a result is justified theoretically by the following theorem.

Theorem 4.14. *Assume that $\overline{\mathcal{H}}_0^c$ is satisfied, and that we apply the parametric independent bootstrap. Set*

$$\Theta n, 0 := \sqrt{n}\big(\hat{\theta}_0 - \theta_0\big), \Theta_{n,k} := \sqrt{n}\big(\hat{\theta}(A_{k,J}) - \theta_0\big), k = 1, \ldots, m,$$
$$\Theta_{n,0}^* := \sqrt{n}\big(\hat{\theta}_0^* - \theta_0\big), \text{ and } \Theta_{n,k}^* := \sqrt{n}\big(\hat{\theta}^*(A_{k,J}) - \theta_0\big), k = 1, \ldots, m.$$

Then there exists two independent and identically distributed random vectors $(\Theta_0, \ldots, \Theta_m)$ and $(\Theta_0^\perp, \ldots, \Theta_m^\perp)$, and a real number a_0 such that

$$\Big(\Theta_{n,0}, \ldots, \Theta_{n,m}, \Theta_{n,0}^*, \ldots, \Theta_{n,m}^*\Big) \Longrightarrow \Big(\Theta_0, \ldots, \Theta_m, \Theta_0^\perp + a_0\Theta_0, \ldots, \Theta_m^\perp + a_0\Theta_0\Big).$$

The proof of this theorem has been postponed in Appendix 4.7.

As a consequence of the latter result, applying the parametric independent bootstrap procedures for some test statistics based on comparisons between $\hat{\theta}_0$ and the $\hat{\theta}(A_{k,J})$ is valid. For instance, $\overline{\mathcal{T}}_2^c$ and $\left(\overline{\mathcal{T}}_2^c\right)^{**}$ will converge jointly in distribution to a pair of independent and identically distributed variables. Indeed, we have

$$\begin{aligned}
\left(\overline{\mathcal{T}}_2^c\,,\, \left(\overline{\mathcal{T}}_2^c\right)^{**}\right) &= \left(n\sum_{k=1}^{m} \|\hat{\theta}_{n,0}^b - \hat{\theta}_n(A_{k,J})\|^2\omega_k\,,\, n\sum_{k=1}^{m} \|\hat{\theta}_{n,0}^* - \hat{\theta}_n^*(A_{k,J})\|^2\omega_k\right) \\
&= \left(n\sum_{k=1}^{m} \|\hat{\theta}_{n,0}^b - \theta_0 + \theta_0 - \hat{\theta}_n(A_{k,J})\|^2\omega_k\,,\, n\sum_{k=1}^{m} \|\hat{\theta}_{n,0}^* - \theta_0 + \theta_0 - \hat{\theta}_n^*(A_{k,J})\|^2\omega_k\right) \\
&\Longrightarrow \left(\sum_{k=1}^{m} \|\Theta_0 - \Theta_k\|^2\omega_k\,,\, \sum_{k=1}^{m} \|\Theta_0^\perp + a_0\Theta_0 - \Theta_k^\perp - a_0\Theta_0\|^2\omega_k\right).
\end{aligned}$$

The same reasoning applies with $\overline{\mathcal{T}}^c_{\infty}$ and $\overline{\mathcal{T}}^c_{dist}$, for sufficiently regular copula families.

Remark 4.15. *We have to stress that the first-level bootstrap, i.e. resampling among the conditioning variables $\mathbf{X}_{i,J}$, $i = 1, \ldots, n$ is surely necessary to obtain the latter result. Indeed, it can be seen that the key proposition 4.16 is no longer true otherwise, because the limiting covariance functions of the two corresponding processes $\mathbb{G}_n$ and $\mathbb{G}^*_n$ will not be the same: see remark 4.22 below.*

4.4 Numerical applications

Now, we would like to evaluate the empirical performances of some of the previous tests by simulation. Such an exercise has been led by [58] or [16] extensively in the case of goodness-of-fit test for unconditional copulas. Our goal is not to replicate such experiments in the case of conditional copulas and for tests of the simplifying assumption. Indeed, we have proposed dozens of test statistics and numerous bootstrap schemes. Moreover, testing the simplifying assumption through $\mathcal{H}_0$ or some "box-type" problems through $\overline{\mathcal{H}}_0$ doubles the scale of the task. Finally, in the former case, we depend on smoothing parameters that induce additional degrees of freedom for the fine tuning of the experiments (note that [58] and [16] have renounced to consider tests that require additional smoothing parameters, as the pivotal test statistics proposed in [49]. In our opinion, an exhaustive simulation experiment should be the topic of (at least) one additional paper. Here, we will restrict ourselves to some partial numerical elements. They should convince readers that the methods and techniques we have discussed previously provide fairly good results and can be implemented in practice safely.

Hereafter, we consider bivariate conditional copulas and a single conditioning variable, i.e. $p = 2$ and $d = 3$. The sample sizes will be $n = 500$, except if it is differently specified. Concerning the bootstrap, we will resample $N = 200$ times to calculate approximated p-values. Each experiment has been repeated 500 times to calculate the percentages of rejection. The computations have been made on a standard laptop, and, for the non-parametric bootstrap, they took an average time of 14.1 seconds for $\mathcal{I}_{\chi,n}$; 26.9s for $\mathcal{T}^{0,m}_{CvM,n}$, 103s for $\mathcal{I}_{2,n}$, 265s for $\mathcal{T}^c_2$ and 0.922s for $\overline{\mathcal{T}}^c_2$.

In terms of model specification, the margins of $\mathbf{X} = (X_1, X_2, X_3)$ will depend on X_3 as

$$X_1 \sim \mathcal{N}(X_3, 1),\ X_2 \sim \mathcal{N}(X_3, 1) \text{ and } X_3 \sim \mathcal{N}(0, 1).$$

We have studied the following conditional copula families: given $X_3 = x$,

- the Gaussian copula model, with a correlation parameter $\theta(x)$,
- the Student copula model, with 4 degrees of freedom and a correlation parameter $\theta(x)$,
- the Clayton copula model, with a parameter $\theta(x)$,
- the Gumbel copula model, with a parameter $\theta(x)$,
- the Frank copula model, with a parameter $\theta(x)$.

In every case, we calibrate $\theta(x)$ such that the conditional Kendall's tau $\tau(x)$ satisfies $\tau(x) = \Phi(x)\tau_{\max}$, for some constant $\tau_{\max} \in (0, 1)$. By default, $\tau_{\max}$ is equal to one. In this case, the random Kendall's tau are uniformly distributed on $[0, 1]$.

Test of $\mathcal{H}_0$: we calculate the percentage of rejections of $\mathcal{H}_0$, when the sample is drawn under the true law (level analysis) or when it is drawn under the same parametric copula family, but with varying

parameters (power analysis). For example, when the true law is a Gaussian copula with a constant parameter ρ corresponding to $\tau = 1/2$, we draw samples under the alternative through a bivariate Gaussian copula whose random parameters are given by $\tau(X_3) = \Phi(X_3)$. The chosen test statistics are $\mathcal{T}^0_{CvM}$, $\tilde{\mathcal{T}}^0_{CvM}$ (nonparametrics test of $\mathcal{H}_0$), $\mathcal{I}_{\chi,n}$ and $\mathcal{I}_{2,n}$ (nonparametric tests of $\mathcal{H}_0$ based on the independence property) and $\mathcal{T}^c_2$ (a parametric test of $\mathcal{H}^c_0$). To compute these statistics, we use the estimator of the partial copula defined in Equation (4.6).

Test of $\overline{\mathcal{H}}_0$: in the case of the test with boxes, the data-generating process will be

$$X_1 \sim \mathcal{N}(\gamma(X_3), 1),\ X_2 \sim \mathcal{N}(\gamma(X_3), 1) \text{ and } X_3 \sim \mathcal{N}(0, 1),$$

where $\gamma(x) = \Phi^{-1}\left(\lfloor m\Phi(X_3)\rfloor/m\right)$, so that the boxes are all of equal probability. As $m \to \infty$, we recover the continuous model for which $\gamma(x) = x$.

In the same way, we calibrate the parameter $\theta(x)$ of the conditional copulas such that the conditional Kendall's tau satisfies $\tau(X_3) = \lfloor m\Phi(X_3)\rfloor/m$.

The choice of "the best" boxes $A_{1,J}, \ldots, A_{m,J}$ is not an easy task. This problem happens frequently in statistics (think of Pearson's chi-square test of independence, for instance), and there is no universal answer. Nonetheless, in some applications, intuition can be fuelled by the context. For example, in finance, it makes sense to test whether past positive returns induce different conditional dependencies between current returns than past negative returns. And, as a general "by default" rule, we can divide the space of $\mathbf{X}_J$ into several boxes of equal (empirical) probabilities. This trick is particularly relevant when the conditioning variable is univariate. Therefore, in our example, we have chosen $m = 5$ boxes of equal empirical probability for X_3, with equal weights.

We have only evaluated $\overline{\mathcal{T}}^c_2$ for testing $\overline{\mathcal{H}}^c_0$. In the following tables, for the parametric tests,

- "bootNP" means the usual nonparametric bootstrap ;
- "bootPI" means the parametric independent bootstrap (where $\mathbf{Z}_{I|J}$ is drawn under $C_{\hat{\theta}_0}$ and $\mathbf{X}_J$ under the usual nonparametric bootstrap);
- "bootPC" means the parametric conditional bootstrap (nonparametric bootstrap for $\mathbf{X}_J$, and $\mathbf{X}_I$ is sampled from the estimated conditional copula $C_{\hat{\theta}(\mathbf{X}^*_J)}$);
- "bootPseudoInd" means the pseudo-independent bootstrap (nonparametric bootstrap for $\mathbf{X}_J$, and draw $\hat{\mathbf{Z}}^*_{I|J}$ independently, among the pseudo-observations $\hat{\mathbf{Z}}_{j,I|J}$);
- "bootCond" means the conditional bootstrap (nonparametric bootstrap for $\mathbf{X}_J$, and $\mathbf{X}_I$ is sampled from the estimated conditional law of $\mathbf{X}_I$ given $\mathbf{X}^*_J$).

Concerning tests of $\mathcal{H}_0$, the results are relatively satisfying. For the nonparametric tests and those based on the independence property (Tables 4.1 and 4.2) the rejection rates are large when $\tau_{\max} = 1$, and the theoretical levels (5%) are underestimated (a not problematic feature in practice). This is still the case for tests of the simplifying assumption under a parametric copula model through $\mathcal{T}^c_2$: see Tables 4.3 and 4.4. The three bootstrap schemes provide similar numerical results. Remind that the bootstrapped statistics is $(\mathcal{T}^c_2)^{**}$ with bootPI (Remark 4.8). Tests of $\overline{\mathcal{H}}_0$ under a parametric framework and through $\overline{\mathcal{T}}^c_2$ confirm such observations. To evaluate the accuracy of the bootstrap approximations asymptotically, we have compared the empirical distribution of some test statistics and their bootstrap versions under the null hypothesis for two bootstrap schemes (see Figures 4.5 and 4.6). For the nonparametric bootstrap,

the two distributions begin to match each other at $n = 5000$ whereas $n = 500$ is enough for the parametric independent bootstrap.

We have tested the influence of $\tau_{\max}$: the smaller is this parameter, the smaller is the percentage of rejections under the alternative, because the simulated model tends to induce lower dependencies of copula parameters w.r.t. X_3: see Figures 4.1, 4.2, 4.3, and 4.4. Note that, on each of these figures, the point at the left corresponds to a conditional Kendall's tau which is constant, and equal to 0 (because $\tau_{\max} = 0$) whereas the rejection percentages in Tables 4.1 and 4.3 correspond to a conditional Kendall's tau constant, and equal to 0.5. As the two data-generating process are not the same, the rejection percentages can differ even if both are under the null hypothesis. Nevertheless, in every case, our empirical sizes converge to 0.05 as the sample size increases. When $n = 5000$, we found that the percentage of rejections are between 4% and 6%.

We have not tried to exhibit an "asymptotically optimal" bandwidth selector for our particular testing problem. This could be the task for further research. We have preferred a basic ad-hoc procedure. In our test statistics, we smooth w.r.t. $F_3(X_3)$ (or its estimate, to be specific), whose law is uniform on $(0, 1)$. A reasonable bandwidth h is given by the so-called rule-of-thumb in kernel density estimation, i.e. $h^* = \sigma(F_3(X_3))/n^{1/5} = 1/(\sqrt{12}n^{1/5}) = 0.083$. Such a choice has provided reasonable results. The typical influence of the bandwidth choice on the test results is illustrated in Figure 4.7. In general, the latter h^* belongs to reasonably wide intervals of "convenient" bandwidth values, so that the performances of our considered tests are not very sensitive to the bandwidth choice.

To avoid boundary problems, we have slightly modified the test statistics: we remove the observations i such that $F_3(X_{i,3}) \leq h$ or $F_3(X_{i,3}) \geq 1 - h$. This corresponds to changing the integrals (resp. max) on $[0, 1]$ to integrals (resp. max) on $[h, 1 - h]$.

Family	$\mathcal{T}^0_{CvM,n}$ (4.8)	$\tilde{\mathcal{T}}^0_{CvM,n}$ (4.10)	$\mathcal{I}_{\chi,n}$ (4.13)	$\mathcal{I}_{2,n}$ (4.15)
Gaussian	0	0	0	0
Student	0	0	0	0
Clayton	0	0	1	0
Gumbel	1	1	0	1
Frank	0	0	0	0

Table 4.1: Rejection percentages under the null (nonparametric tests, nonparametric bootstrap bootNP).

Family	$\mathcal{T}^0_{CvM,n}$ (4.8)	$\tilde{\mathcal{T}}^0_{CvM,n}$ (4.10)	$\mathcal{I}_{\chi,n}$ (4.13)	$\mathcal{I}_{2,n}$ (4.15)
Gaussian	98	100	100	93
Student	100	99	98	90
Clayton	99	99	99	98
Gumbel	99	98	100	95
Frank	98	100	98	50

Table 4.2: Rejection percentages under the alternative (nonparametric tests, nonparametric bootstrap bootNP).

Family	$\mathcal{T}_2^c$ (4.18)			$\overline{\mathcal{T}}_2^c$ (4.30)		
	bootPI	bootPC	bootNP	bootPI	bootPC	bootNP
Gaussian	4	0	0	6	4	1
Student	6	0	2	4	5	3
Clayton	7	0	1	7	1	1
Gumbel	3	1	0	9	2	2
Frank	4	0	6	3	5	1

Table 4.3: Rejection percentages under the null (parametric tests).

Family	$\mathcal{T}_2^c$ (4.18)			$\overline{\mathcal{T}}_2^c$ (4.30)		
	bootPI	bootPC	bootNP	bootPI	bootPC	bootNP
Gaussian	100	100	100	100	100	100
Student	100	100	100	100	100	100
Clayton	100	62	98	100	98	100
Gumbel	100	100	34	100	99	76
Frank	100	100	100	100	100	100

Table 4.4: Rejection percentages under the alternative (parametric tests).

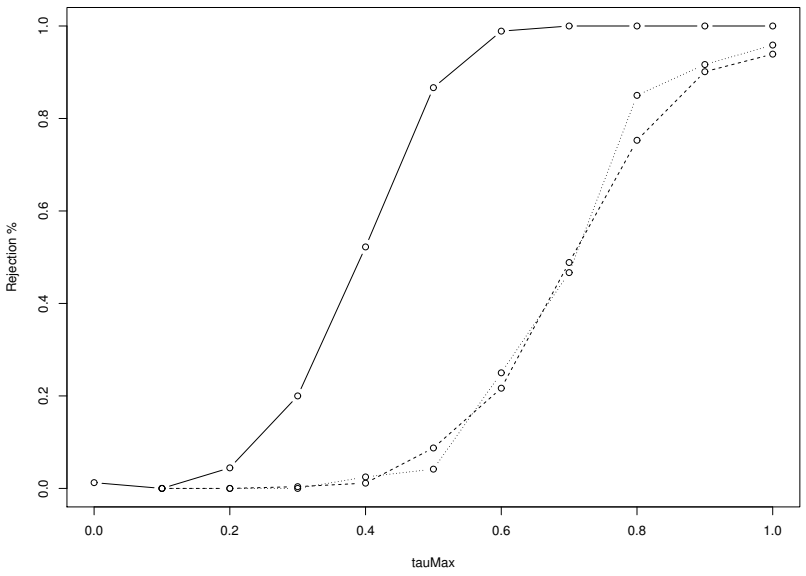


Figure 4.1: Rejection percentages for the statistics $\mathcal{I}_\chi$ (4.13) as a function of τ_{max}: we use the gaussian copula, with a conditional parameter $\theta(x)$ calibrated such that the conditional Kendall's tau $\tau(x)$ satisfies $\tau(x) = \tau_{max} \cdot \Phi(x)$. Solid line: bootNP. Dashed line : bootPseudoInd. Dotted line : bootCond.

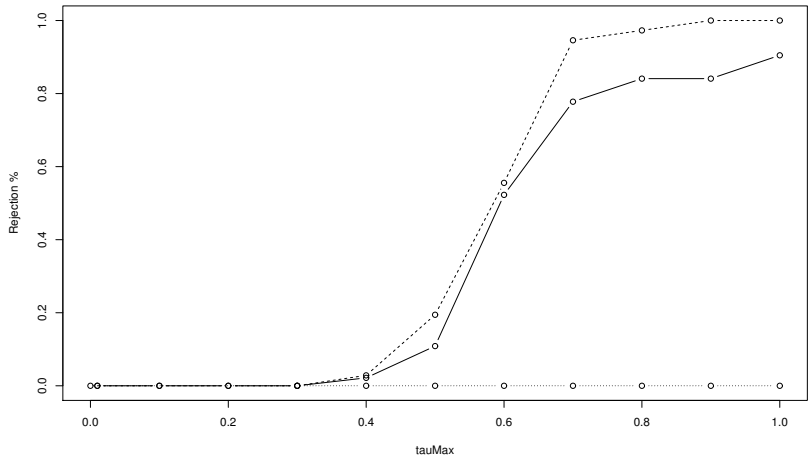


Figure 4.2: Rejection percentages for the statistics $I_{2,n}$ (4.15) as a function of τ_{max}: we use the gaussian copula, with a conditional parameter $\theta(x)$ calibrated such that the conditional Kendall's tau $\tau(x)$ satisfies $\tau(x) = \tau_{max} \cdot \Phi(x)$. Solid line: bootNP. Dashed line : bootPseudoInd. Dotted line : bootCond.

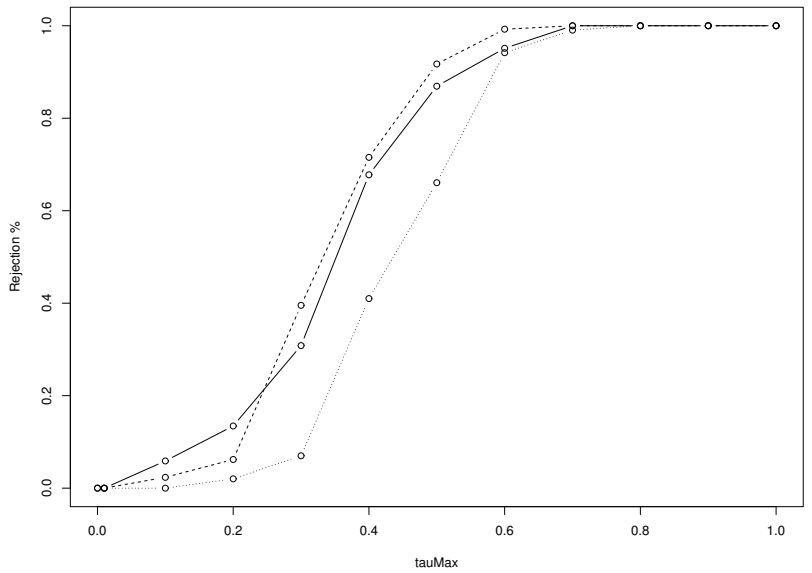


Figure 4.3: Rejection percentages for the statistics $\mathcal{T}_2^c$ (4.18) as a function of τ_{max}: we use the gaussian copula, with a conditional parameter $\theta(x)$ calibrated such that the conditional Kendall's tau $\tau(x)$ satisfies $\tau(x) = \tau_{max} \cdot \Phi(x)$. Solid line: bootNP. Dashed line : bootPI. Dotted line : bootPC.

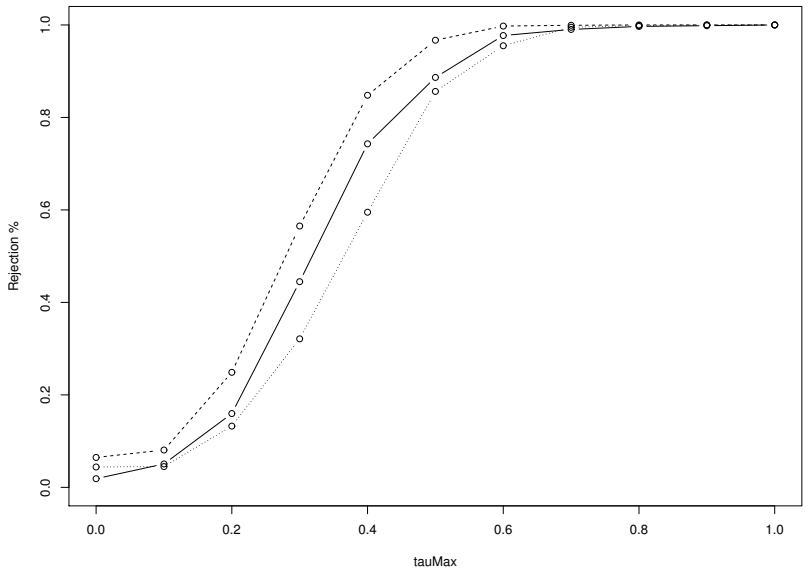


Figure 4.4: Rejection percentages for the statistics $\overline{\mathcal{T}}_2^c$ (4.30) as a function of τ_{max}: we use the gaussian copula, with a conditional parameter $\theta(x)$ calibrated such that the conditional Kendall's tau $\tau(x)$ satisfies $\tau(x) = \tau_{max} \cdot \lfloor m\Phi(X_3) \rfloor / m$. Solid line: bootNP. Dashed line : bootPI. Dotted line : bootPC.

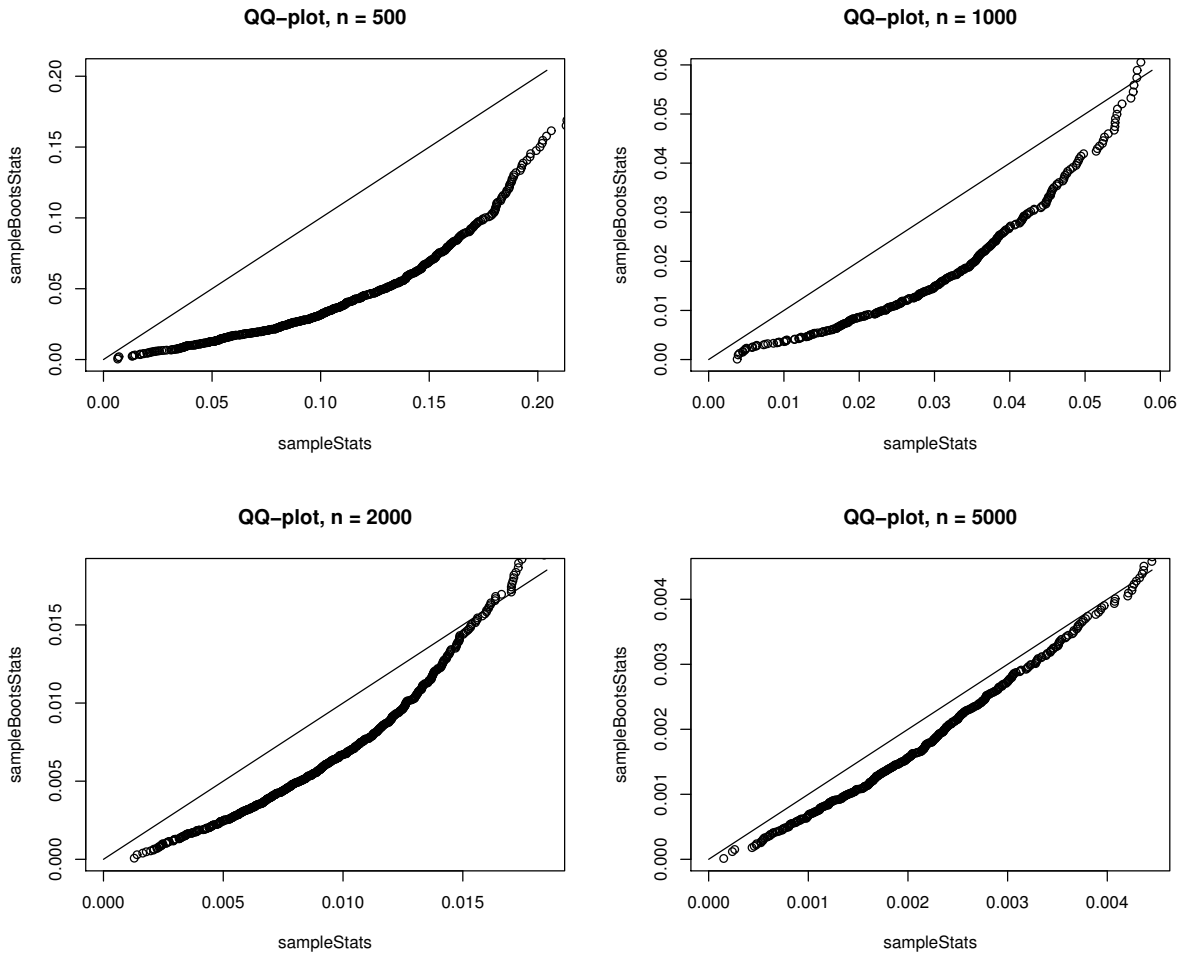


Figure 4.5: QQ-plot of a sample of the test statistic $\overline{\mathcal{T}}_2^c$ and a sample of the bootstrap test statistic $(\overline{\mathcal{T}}_2^c)^*$ using the non-parametric bootstrap for the gaussian copula, with different sample sizes and under $\overline{\mathcal{H}}_0^c$ (conditional Kendall's tau is constant and equal to 0.5).

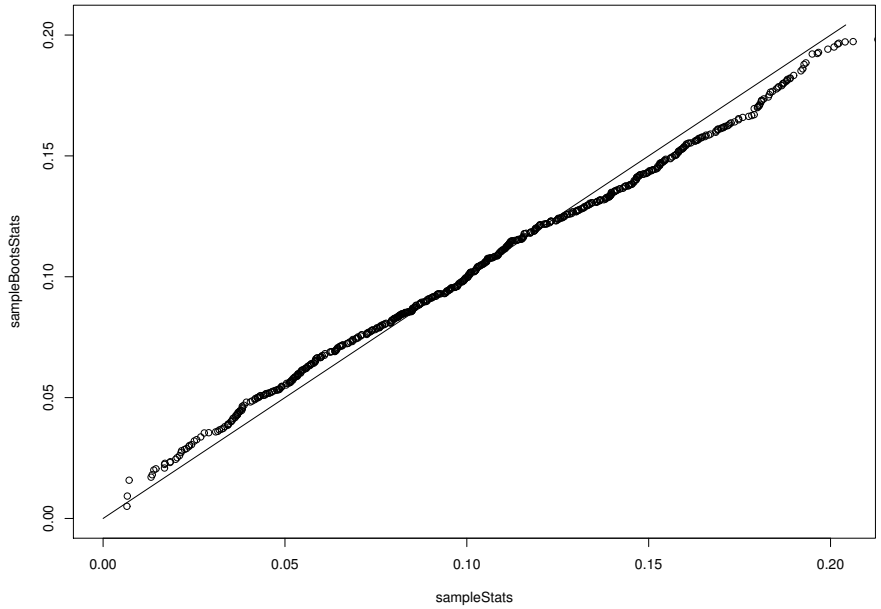


Figure 4.6: QQ-plot of a sample of the test statistic $\overline{\mathcal{T}}_{2,c}$ and a sample of the bootstrap test statistic $(\overline{\mathcal{T}}_{2,c})^{**}$ using the parametric independent bootstrap for the gaussian copula, with $n = 500$ and under $\overline{\mathcal{H}}_0^c$ (conditional Kendall's tau is constant and equal to 0.5).

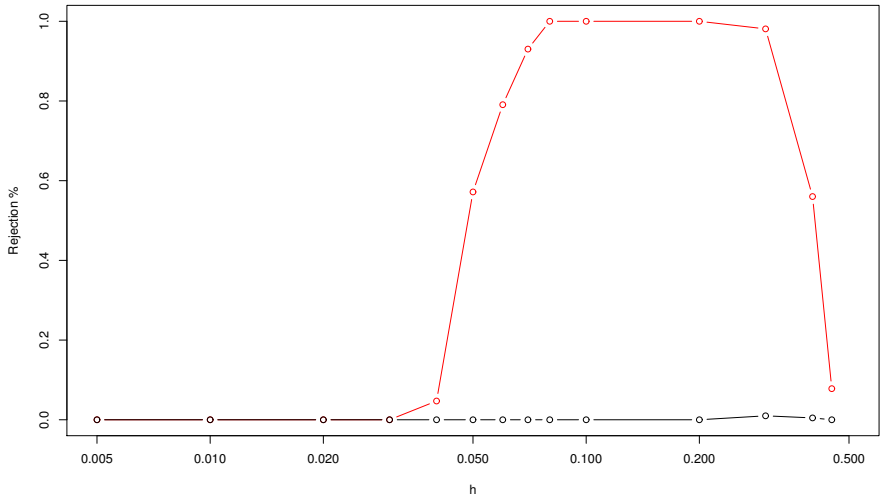


Figure 4.7: Rejection percentage for the statistic $\mathcal{T}_{CvM,n}^{0,m}$ (4.8) with $m = 20$ as a function of h. The red (resp. black) line corresponds to the alternative (resp. zero) assumption.

4.5 Conclusion

We have provided an overview of the simplifying assumption problem, under a statistical point of view. In the context of nonparametric or parametric conditional copula models (with unknown conditional marginal distributions), numerous testing procedures have been proposed. We have developed the theory towards a slightly different but related approach, where “box-type” conditioning events replace pointwise ones. This opens a new field for research that is interesting per se. Several new bootstrap procedures have been detailed, to evaluate p-values under the zero assumption in both cases. In particular, we have proved the validity of one of them (the “parametric independent” bootstrap scheme under $\overline{\mathcal{H}}_0$).

Clearly, there remains a lot of work. We have opened the Pandora box rather than provided definitive answers. Open questions are still numerous: precise theoretical convergence results of our test statistics (and others!), validity of these new bootstrap schemes, bandwidth choices, empirical performances,... All these dimensions would require further research. We have made a contribution to the landscape of problems related to the simplifying assumption, and proposed a working program for the whole copula community.

Acknowledgements: The authors acknowledge the two anonymous reviewers for their numerous and very useful comments and suggestions.

4.6 Notation

$\mathbf{X} = (\mathbf{X}_I, \mathbf{X}_J)$	random vector of size d
I, J	$\{1, \dots, p\}$ and $\{p+1, \dots, d\}$
$\mathcal{S} = (\mathbf{X}_{1,1:d}, \dots, \mathbf{X}_{n,1:d})$	initial sample of n i.i.d. observations
A_J	measurable subset in $\mathbb{R}^{d-p}$
$\mathcal{A}_J$	collection of all measurable subsets of $\mathbb{R}^{d-p}$ such that $\mathbf{X}_J$ is in each set with positive probability
$\overline{\mathcal{A}}_J = \{A_{1,J}, \dots, A_{m,J}\}$	partition of $\mathbb{R}^{d-p}$ into m sets such that $\mathbf{X}_J$ is in each set with positive probability
Y	box index, i.e. Y is the k such that $\mathbf{X}_J \in A_{k,J}$
$\hat{U}_{i,k}$	i-th pseudo-observation of the k-th variable
$\mathbf{Z}_{I\|J}$	conditional observation of $\mathbf{X}_I$ given $\mathbf{X}_J$
$\mathbf{Z}_{I\|Y}$	conditional observation of $\mathbf{X}_I$ given the box index Y
$\mathcal{C}$	copula family indexed by the elements of a set Θ
C_θ	copula of the family $\mathcal{C}$ with the parameter $\theta \in \Theta$
c_θ	density of the copula C_θ
θ_0	unconditional parameter of the copula of $\mathbf{Z}_{I\|J}$
$\theta(\mathbf{x}_J)$	parameter of the conditional copula of $\mathbf{Z}_{I\|J}$ given $\mathbf{X}_J = \mathbf{x}_J$
θ_0^b	unconditional parameter of the copula of $\mathbf{Z}_{I\|Y}$
$\theta(A_J)$	conditional parameter of the copula of $\mathbf{Z}_{I\|Y}$ given $X_J \in A_J$
$F_i(\cdot)$	marginal cdf of X_i, $i = 1, \dots, d$
$F_{i\|J}(\cdot\|\mathbf{X}_J \in A_J)$	conditional marginal cdf of X_i given $\mathbf{X}_J \in A_J$, $i = 1, \dots, p$
$F_{i\|J}(\cdot\|\mathbf{X}_J = \mathbf{x}_J)$	conditional marginal cdf of X_i given $\mathbf{X}_J = \mathbf{x}_J$, $i = 1, \dots, p$
$F_{I\|J}(\cdot\|\mathbf{X}_J \in A_J)$	conditional joint cdf of $\mathbf{X}_I$ given $\mathbf{X}_J \in A_J$
$F_{I\|J}(\cdot\|\mathbf{X}_J = \mathbf{x}_J)$	conditional joint cdf of $\mathbf{X}_I$ given $\mathbf{X}_J = \mathbf{x}_J$
$G_{I,J}(\cdot)$	joint cdf of $(\mathbf{Z}_{I\|J}, \mathbf{X}_J)$
$C_{I\|J}^{A_J}(\cdot\|\mathbf{X}_J \in A_J)$	conditional copula of $\mathbf{X}_I$ given $\mathbf{X}_J \in A_J$
$C_{I\|J}(\cdot\|\mathbf{X}_J = \mathbf{x}_J)$	conditional copula of $\mathbf{X}_I$ given $\mathbf{X}_J = \mathbf{x}_J$
$C_{s,I\|J}(\cdot)$	partial copula of $\mathbf{X}_I$ given $\mathbf{X}_J$

Table 4.5: Table of notation

$\mathcal{T}^0_{CvM,n}$ (4.3)	brute-force test statistic of $\mathcal{H}_0$, constructed with the L_2 distance between the conditional and the partial copula
(resp. $\mathcal{T}^0_{KS,n}$ (4.2))	(resp. L_∞ distance)
(resp. $\mathcal{T}^{0,m}_{CvM,n}$ (4.8))	(resp. L_2 distance using a fixed number m of points)
$\tilde{\mathcal{T}}^0_{CvM,n}$ (4.10)	brute-force test statistic of $\mathcal{H}_0$, constructed with the L_2-distance
(resp. $\tilde{\mathcal{T}}^0_{KS,n}$ (4.9))	(resp. L_∞-distance) between all pairs of conditional copulas
$\mathcal{I}_{\chi,n}$ (4.13)	chi-square-type test statistic of the independence between $\hat{\mathbf{Z}}_{I\|J}$ and $\mathbf{X}_J$
$\mathcal{I}_{KS,n}$ (4.14)	test statistic based on the distance between the joint empirical cdf of $(\hat{\mathbf{Z}}_{I\|J}, \mathbf{X}_J)$ and the product of their empirical cdf, using the L_∞ norm
(resp. $\mathcal{I}_{2,n}$ (4.15))	(resp. using the L_2 norm)
(resp. $\mathcal{I}_{CvM,n}$ (4.16))	(resp. using the L_2 norm, weighted by the joint empirical cdf as weight)
$\mathcal{T}^c_\infty$ (4.18)	test statistic based on the L_∞ distance between the parameter of the conditional copula and the constant parameter of the partial copula
(resp. $\mathcal{T}^c_2$ (4.18))	(resp. L_2 distance)
(resp. $\mathcal{T}^c_{dist}$ (4.19))	(resp. using some distance between the estimated copulas)
(resp. $\mathcal{T}^c_{dens}$ (4.20))	(resp. using the L_2 distance between the estimated copula densities)
$\overline{\mathcal{T}}_{dist,n}$ (4.29)	brute-force test statistic of $\overline{\mathcal{H}}_0$ constructed with the distance $dist(\cdot,\cdot)$ between all pairs of conditional copulas with Borelian subsets
(resp. $\overline{\mathcal{T}}_{KS,n}$ (4.27))	(resp. with the L_∞ distance)
(resp. $\overline{\mathcal{T}}_{CvM,n}$ (4.28))	(resp. with the L_2 distance)
$\overline{\mathcal{T}}^c_\infty$ (4.30)	test statistic based on the L_∞ distance between the parameters estimated on each set and the simplified parameter
(resp. $\overline{\mathcal{T}}^c_2$ (4.30))	(resp. based on the L_2 distance)
(resp. $\overline{\mathcal{T}}^c_{dist}$ (4.31))	(resp. based on some distance between the copulas whose parameters are estimated on each set and the copula with the simplified parameter)
$\mathcal{T}^*$, $\mathcal{T}^{**}$	bootstrap statistics corresponding to a general test statistic $\mathcal{T}$

Table 4.6: Table of main test statistics

4.7 Proof of Theorem 4.14

4.7.1 Preliminaires

Let $(\mathbf{Z}_i)_{i=1,\dots,n}$ be a sequence of i.i.d random vectors in $[0,1]^p$, $\mathbf{Z}_i$ being drawn from the true cdf C_{θ_0}. They have the same law as the previously called vectors $\mathbf{Z}_{i,I|A_J}$ or $\mathbf{Z}_{i,I|Y}$ under the zero assumption $\overline{\mathcal{H}}_0^c$. Let $(\mathbf{X}_{i,J})_{i=1,\dots,n}$ be a sequence of i.i.d random vectors in $\mathbb{R}^{d-p}$, $\mathbf{X}_{i,J} \sim F_J$. Let $(\mathbf{Z}_i^*)_{i=1,\dots,n}$ be an independent sequence of i.i.d random vectors in $[0,1]^p$, where $\mathbf{Z}_i^* \sim C_{\theta_0}$ exactly as $\mathbf{Z}_i$. The three samples $(\mathbf{Z}_i)$, $(\mathbf{X}_{i,J})$ and $(\mathbf{Z}_i^*)$ are mutually independent. Let $(\mathbf{X}_{i,J}^*)_{i=1,\dots,n}$ be a sequence of i.i.d random vectors in $\mathbb{R}^{d-p}$, which are drawn from $F_{n,J}$, the empirical cdf of $\mathbf{X}_{1,J},\dots\mathbf{X}_{n,J}$, and independently of both $(\mathbf{Z}_i)$ and $(\mathbf{Z}_i^*)$.

In the following, we shall use the notation $f\otimes g := (x,y)\mapsto f(x)g(y)$ when f, g are two real functions, possibly from different spaces. Set $l(\theta,\cdot) := \log c_\theta(\cdot)$. We will need some conditions of regularity.

Assumption (R): $(\theta,\mathbf{u}_I)\mapsto l(\theta,\mathbf{u}_I)$ is three times differentiable with respect to θ, for every $\mathbf{u}_I\in(0,1)^p$. Moreover, for every $\epsilon>0$,

$$\mathbb{E}\left[\sup_{\theta\,|\|\theta-\theta_0\|\le\epsilon}\ \sup_{\{\mathbf{z}\,|\|\mathbf{z}-\mathbf{Z}_i\|\le\|\hat{\mathbf{Z}}_{i,I|Y}-\mathbf{Z}_i\|\}} \left|\left|\frac{\partial^3 l}{\partial\theta^3}(\theta,\mathbf{z})\right|\right|\right] < +\infty.$$

The latter technical assumption can be weakened through some trimming techniques, as in [50]. Since this would require to change the definitions of the parametric estimators, we do not try to improve towards this direction. We will set $\dot{c}_\theta := \partial c_\theta/\partial\theta$ and $\ddot{c}_\theta := \partial^2 c_\theta/\partial\theta^2$.

We associate to every $\mathbf{X}_{i,J}$ (resp. $\mathbf{X}_{i,J}^*$) its corresponding index Y_i (resp. Y_i^*) s.t. $\mathbf{X}_{i,J}\in A_{Y_i}$ (resp. $\mathbf{X}_{i,J}^*\in A_{Y_i^*}$). For convenience, we assume that $(A_k)_{k=1,\dots,m}$ is a partition of $\mathbb{R}^{d-p}$. Otherwise, we have to restrict our sample to the observations for which $X_{i,J}$ belongs to some "box" A_k, $k=1,\dots,m$. Therefore, denote by C_n, C_n^*, $P_{n,Y}$ and $P_{n,Y}^*$ the empirical laws of $(\mathbf{Z}_i)$, $(\mathbf{Z}_i^*)$, (Y_i) and (Y_i^*) respectively. The joint law of $(\mathbf{Z}_1,Y_1)$ (resp. $(\mathbf{Z}_1,\mathbf{X}_{1,J})$) will be denoted by $\overline{G} := C_{\theta_0}\otimes P_Y$ (resp. $\overline{G} := C_{\theta_0}\otimes F_J$), with $P_Y(k)=\mathbb{P}(Y=k)$, $k=1,\dots,m$. Denote by G_n (resp. $\overline{G}_n$) the empirical law of $(\mathbf{Z}_i,Y_i)_{i=1,\dots,n}$ (resp. $(\mathbf{Z}_i,\mathbf{X}_{i,J})_{i=1,\dots,n}$) Moreover, G_n^* and $\overline{G}_n^*$ will be the empirical distributions of $(\mathbf{Z}_i^*,Y_i^*)_{i=1,\dots,n}$ and $(\mathbf{Z}_i^*,\mathbf{X}_{i,J}^*)_{i=1,\dots,n}$ respectively. Let $\mathcal{P}_n$ be the joint probability distribution of

$$\left(\mathbf{Z}_i,Y_i,\mathbf{Z}_i^*,Y_i^*\right)_{i=1,\dots,n}\in\left([0,1]^p\times\{1,\dots,m\}\right)^{\otimes 2n}.$$

The following proposition is key. It will be proved in Subsection 4.7.3.

Proposition 4.16. *Consider the empirical process defined on* $[0,1]^p\times\mathbb{R}^{d-p}$ *by*

$$\overline{\mathbb{G}}_n(\mathbf{z},\mathbf{x}_J) := \sqrt{n}(\overline{G}_n-\overline{G})(\mathbf{z},\mathbf{x}_J) := \frac{1}{\sqrt{n}}\sum_{i=1}^n\{\mathbb{1}\big((\mathbf{Z}_i,\mathbf{X}_{i,J})\le(\mathbf{z},\mathbf{x}_J)\big)-C_{\theta_0}(\mathbf{z})F_J(\mathbf{x}_J)\},$$

and the corresponding bootstrapped empirical process

$$\overline{\mathbb{G}}_n^*(\mathbf{z},\mathbf{x}_J) := \frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbb{1}\big((\mathbf{Z}_i^*,\mathbf{X}_{i,J}^*)\le(\mathbf{z},\mathbf{x}_J)\big)-C_{\theta_0}(\mathbf{z})\frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbb{1}(\mathbf{X}_{i,J}\le\mathbf{x}_J),$$

or, equivalently, $\overline{\mathbb{G}}_n^* = \sqrt{n}(\overline{G}_n^*-C_{\theta_0}\otimes F_{n,J})$. *Then there exist two independent and identically distributed Gaussian processes* $\overline{\mathbb{A}}_G$ *and* $\overline{\mathbb{A}}_G^\perp$ *such that* $\left(\overline{\mathbb{G}}_n\,,\,\overline{\mathbb{G}}_n^*\right)$ *converges to* $\left(\overline{\mathbb{A}}_G\,,\,\overline{\mathbb{A}}_G^\perp\right)$ *weakly in* $\Big(\ell^\infty\big([0,1]^p\times\mathbb{R}^{d-p}\big)\Big)^2$.

As a Corollary, we deduce the same results when the discrete variables Y_i replace the variables $\mathbf{X}_{i,J}$.

Proposition 4.17. *Under the assumptions of Proposition 4.16, let the empirical process defined on* $[0,1]^p \times \{1,\dots,m\}$ *by*

$$\begin{aligned}\mathbb{G}_n(\mathbf{z},k) &:= \sqrt{n}(G_n - G)(\mathbf{z},k)\\ &:= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\mathbb{1}\big(\mathbf{Z}_i \le \mathbf{z}, Y_i = k\big) - C_{\theta_0}(\mathbf{z})P_Y(k)\},\end{aligned}$$

and its bootstrapped empirical process

$$\mathbb{G}_n^*(\mathbf{z},k) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{1}\big(\mathbf{Z}_i^* \le \mathbf{z}, Y_i^* = k\big) - C_{\theta_0}(\mathbf{z})\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{1}(Y_i = k),$$

or equivalently $\mathbb{G}_n^* = \sqrt{n}(G_n^* - C_{\theta_0} \otimes P_{n,Y})$, $P_{n,Y}(k)$ *being the empirical proportion of* $\mathcal{S}_n$*-observations into* A_k*. Then, there exist two independent and identically distributed processes* $\mathbb{A}_G$ *and* $\mathbb{A}_G^{\perp}$ *such that* $\big(\mathbb{G}_n\,,\,\mathbb{G}_n^*\big)$ *converges to* $(\mathbb{A}_G\,,\,\mathbb{A}_G^{\perp})$ *weakly in* $\Big(\ell^\infty\big([0,1]^p \times \{1,\dots,m\}\big)\Big)^2$.

Remark 4.18. *The covariance function of* $\mathbb{A}_G$ *(or* $\mathbb{A}^{\perp}$*) is given by*

$$\begin{aligned}\mathbb{E}[\mathbb{A}_G(\mathbf{z},y)\mathbb{A}_G(\mathbf{z}',y')] &= \lim_n \mathbb{E}[\mathbb{G}_n(\mathbf{z},y)\mathbb{G}_n(\mathbf{z}',y')]\\ &= \mathbb{1}(y=y')\mathbb{P}(Y=y)C_{\theta_0}(\mathbf{z}\wedge\mathbf{z}') - \mathbb{P}(Y=y)\mathbb{P}(Y=y')C_{\theta_0}(\mathbf{z})C_{\theta_0}(\mathbf{z}').\end{aligned}$$

As a "toolbox", we will need the following lemma.

Lemma 4.19. *Let* $\hat{\theta}_0^b$ *and* $\hat{\theta}(A_k)$ *be the estimators based on the pseudo-sample* $(\hat{\mathbf{Z}}_{i,I|Y}, Y_i)_{i=1,\dots,n}$ *(and then on the sample* $(\mathbf{Z}_i, Y_i)_{i=1,\dots,n}$*) as*

$$\hat{\theta}_0^b := \arg\max_{\theta\in\Theta}\sum_{i=1}^{n}\log c_\theta(\hat{\mathbf{Z}}_{i,I|Y}),\ \textit{and}$$

$$\hat{\theta}(A_k) := \arg\max_{\theta\in\Theta}\sum_{i=1}^{n}\log c_\theta(\hat{\mathbf{Z}}_{i,I|Y}).\mathbb{1}(Y_i = k),\ k = 1,\dots,m.$$

We will assume they lie in the interior of Θ*. Set* $\Theta_{n,0} := \sqrt{n}\big(\hat{\theta}_0^b - \theta_0\big)$*, and, for* $k = 1,\dots,m$*,* $\Theta_{n,k} := \sqrt{n}\big(\hat{\theta}(A_k) - \theta_0\big)$*. Moreover, for any distribution* H *on* $[0,1]^p \times \{1,\dots,m\}$*, set*

$$\psi_{k,1}(H) := \int \frac{\partial l}{\partial\theta}\left(\theta_0, \left(\frac{\int \mathbb{1}\{z_q^1 \le z_q^2, y^1 = y^2\}dH(\mathbf{z}^1,y^1)}{\int \mathbb{1}\{y^1=k\}dH(\mathbf{z}^1,y^1)}\right)_{q=1,\dots,p}\right)\mathbb{1}\{y^2=k\}\,dH(\mathbf{z}^2,y^2),$$

$$\psi_{k,2}(H) := \int \frac{\partial^2 l}{\partial\theta^2}\left(\theta_0, \left(\frac{\int \mathbb{1}\{z_q^1 \le z_q^2, y^1 = y^2\}dH(\mathbf{z}^1,y^1)}{\int \mathbb{1}\{y^1=k\}\,dH(\mathbf{z}^1,y^1)}\right)_{q=1,\dots,p}\right)\mathbb{1}\{y^2=k\}\,dH(\mathbf{z}^2,y^2).$$

(i) For $k = 1,\dots,m$*,*

$$\Theta_{n,k} = -\frac{\sqrt{n}\psi_{k,1}(G_n)}{\psi_{k,2}(G_n)} + o_P(1).$$

(ii) For every discrete law P_Y *with values in* $\{1,\dots,m\}$*, the corresponding distribution* $\tilde{G} := C_{\theta_0} \otimes P_Y$ *satisfies* $\psi_{k,1}(\tilde{G}) = 0$*.*

(iii) $\psi_1 := (\psi_{1,1},\dots,\psi_{m,1})$ *is Hadamard-differentiable at every cdf* H*, and its differential is given by*

$$\dot{\psi}_{k,1}(H)(h) = \int \frac{\partial l}{\partial \theta}\left(\theta_0, \left(\frac{\int \mathbb{1}\{z_q^1 \le z_q^2, y^1 = y^2\} dH(\mathbf{z}^1, y^1)}{\int \mathbb{1}\{y^1 = k\} dH(\mathbf{z}^1, y^1)}\right)_{q=1,\dots,p}\right) \mathbb{1}\{y^2 = k\}\, dh(\mathbf{z}^2, y^2)$$
$$+ \sum_{j=1}^{p} \int \frac{\partial^2 l}{\partial\theta\, \partial z_j}\left(\theta_0, \left(\frac{\int \mathbb{1}\{z_q^1 \le z_q^2, y^1 = y^2\} dH(\mathbf{z}^1, y^1)}{\int \mathbb{1}\{y^1 = k\} dH(\mathbf{z}^1, y^1)}\right)_{q=1,\dots,p}\right) \mathbb{1}\{y^2 = k\}$$
$$\cdot \left(\frac{\int \mathbb{1}\{z_j^1 \le z_j^2, y^1 = y^2\} dh(\mathbf{z}^1, y^1)}{\int \mathbb{1}\{y^1 = k\} dH(\mathbf{z}^1, y^1)} - \frac{\int \mathbb{1}\{z_j^1 \le z_j^2, y^1 = y^2\} dH(\mathbf{z}^1, y^1) \int \mathbb{1}\{y^1 = k\} dh(\mathbf{z}^1, y^1)}{\left(\int \mathbb{1}\{y^1 = y^2\} dH(\mathbf{z}^1, y^1)\right)^2}\right) dH(\mathbf{z}^2, y^2)$$

(iv)

$$\Theta_{n,0} = -\frac{\sum_{k=1}^m \sqrt{n}\big(\psi_{k,1}(G_n)\big)}{\sum_{k=1}^m \psi_{k,2}(G_n)} + o_P(1) = \frac{\sum_{k=1}^m \psi_{k,2}(G_n)\Theta_{n,k}}{\sum_{k=1}^m \psi_{k,2}(G_n)} + o_P(1)$$

Proof : Note that $\hat{\mathbf{Z}}_{i,I|J}$ is an explicit measurable function of the sample $(\mathbf{Z}_{i,I|J})_{i=1,\dots,n}$. Indeed, for any $i = 1, \dots, n$ and $q = 1, \dots, p$,

$$\begin{aligned}
\hat{Z}_{i,q|Y} :=& \hat{F}_{n,q}(X_{i,q}|\mathbf{X}_J \in A_{Y_i,J}) \\
:=& \frac{\sum_{j=1}^n \mathbb{1}\{X_{j,q} \le X_{i,q}, \mathbf{X}_{j,J} \in A_{Y_i,J}\}}{\sum_{j=1}^n \mathbb{1}\{\mathbf{X}_{j,J} \in A_{k(\mathbf{X}_{i,J}),J}\}} \\
=& \frac{\sum_{j=1}^n \mathbb{1}\{F_q(X_{j,q}|\mathbf{X}_J \in A_{Y_j,J}) \le F_q(X_{i,q}|\mathbf{X}_J \in A_{Y_j,J}), Y_j = Y_i\}}{\sum_{j=1}^n \mathbb{1}\{Y_j = Y_i\}} \\
=& \frac{\sum_{j=1}^n \mathbb{1}\{F_q(X_{j,q}|\mathbf{X}_J \in A_{Y_j,J}) \le F_q(X_{i,q}|\mathbf{X}_J \in A_{Y_i,J}), Y_j = Y_i\}}{\sum_{j=1}^n \mathbb{1}\{Y_j = Y_i\}} \\
=& \frac{\sum_{j=1}^n \mathbb{1}\{Z_{j,q|Y} \le Z_{i,q|Y}, Y_j = Y_i\}}{\sum_{j=1}^n \mathbb{1}\{Y_j = Y_i\}}.
\end{aligned} \tag{4.32}$$

We deduce that $\hat{\theta}_0^b$ and $\hat{\theta}(A_k)$ are measurable functions of the unobservable random variables $\mathbf{Z}_{i,I|Y}$ and Y_i, for $i = 1, \dots, n$.

(i). Let $k \in \{1, \dots, m\}$. Applying successively the first order condition for the estimator $\hat{\theta}(A_k)$ and some Taylor series expansions, we have

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^n \frac{\partial l}{\partial \theta}(\hat{\theta}(A_k), \hat{\mathbf{Z}}_{i,I|J}) \mathbb{1}\{Y_i = k\} \\
&= B_n^{1,k} - B_n^{2,k}\big(\hat{\theta}(A_{k,J}) - \theta_0\big) + o_P\big(\hat{\theta}(A_{k,J}) - \theta_0\big), \text{ with}
\end{aligned}$$

$$B_n^{1,k} := \frac{1}{n}\sum_{i=1}^n \frac{\partial l}{\partial \theta}\big(\theta_0, \hat{\mathbf{Z}}_{i,I|J}\big)\mathbb{1}\{Y_i = k\} \text{ and } B_n^{2,k} := -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 l}{\partial \theta^2}\big(\theta_0, \hat{\mathbf{Z}}_{i,I|J}\big)\mathbb{1}\{Y_i = k\},$$

implying

$$\Theta_{n,k} := \sqrt{n}\big(\hat{\theta}(A_{k,J}) - \theta_0\big) = \frac{\sqrt{n}B_n^{1,k}}{B_n^{2,k}} + o_P\big(\Theta_{n,k}\big).$$

Now, invoking (4.32), let us compute the numerator of this expression:

$$\begin{aligned}
B_n^{1,k} &= \frac{1}{n}\sum_{i=1}^n \frac{\partial l}{\partial \theta}\left(\theta_0, \left(\frac{\sum_{j=1}^n \mathbb{1}\{Z_{j,q} \le Z_{i,q}, Y_q = k\}}{\sum_{j=1}^n \mathbb{1}\{Y_j = k\}}\right)_{q=1,\dots,p}\right)\mathbb{1}\{Y_i = k\} \\
&= \int \frac{\partial l}{\partial \theta}\left(\theta_0, \left(\frac{\int \mathbb{1}\{z_q^1 \le z_q^2, y^1 = k\} dG_n(\mathbf{z}^1, y^1)}{\int \mathbb{1}\{y^1 = k\} dG_n(\mathbf{z}^1, y^1)}\right)_{q=1,\dots,p}\right)\mathbb{1}\{y^2 = k\} dG_n(\mathbf{z}^2, y^2) \\
&= \psi_{k,1}(G_n).
\end{aligned}$$

In the same way, the denominator can be rewritten as

$$B_n^{2,k} = -\int \frac{\partial^2 l}{\partial \theta^2}\left(\theta_0, \left(\frac{\int \mathbb{1}\{z_q^1 \le z_q^2, y^1 = k\} dG_n(\mathbf{z}^1, y^1)}{\int \mathbb{1}\{y^1 = k\} dG_n(\mathbf{z}^1, y^1)}\right)_{q=1,\dots,p}\right) \mathbb{1}\{y^2 = k\} dG_n(\mathbf{z}^2, y^2)$$
$$= -\psi_{k,2}(G_n).$$

(ii). We now prove the second part of the lemma. Since $\tilde{G} = C_{\theta_0} \otimes F_Y$, we get

$$\begin{aligned}\psi_{k,1}(\tilde{G}) &:= \int \frac{\partial l}{\partial \theta}\left(\theta_0, \left(\frac{\int \mathbb{1}\{z_q^1 \le z_q^2, y^1 = k\} d\tilde{G}(\mathbf{z}^1, y^1)}{\int \mathbb{1}\{y^1 = k\} d\tilde{G}(\mathbf{z}^1, y^1)}\right)_{q=1,\dots,p}\right) \mathbb{1}\{y^2 = k\} d\tilde{G}(\mathbf{z}^2, y^2) \\ &= \int \frac{\partial l}{\partial \theta}\left(\theta_0, \left(\frac{\mathbb{P}\{Z_q^1 \le z_q^2, Y^1 = k\}}{\mathbb{P}\{Y^1 = k\}}\right)_{q=1,\dots,p}\right) \mathbb{1}\{y^2 = k\} d\tilde{G}(\mathbf{z}^2, y^2) \\ &= \int \frac{\partial l}{\partial \theta}\left(\theta_0, \left(\mathbb{P}\{Z_q^1 \le z_q^2\}\right)_{q=1,\dots,p}\right) dC_{\theta_0}(\mathbf{z}^2) \int \mathbb{1}\{y^2 = k\} dF_Y(y^2) \\ &= \mathbb{P}\{Y = k\} \int \frac{\partial l}{\partial \theta}\left(\theta_0, \mathbf{z}^2\right) dC_{\theta_0}(\mathbf{z}^2) = 0.\end{aligned}$$

(iii). We remark that the law G appears three times in $\psi_{k,1}$: two times in the log-density l and one time at the end of the main integral. By separating the effect of a change from H to $H + h$ in the main integral only (first term of the differential) and the effect of a change in l, and using the standard rule of differential calculus (l is differentiable), we obtain the second part of the given result.

(iv). As in the proof of (i), we apply successively the first order condition for $\hat{\theta}_{n,0}^b$ and some Taylor series expansion to get

$$0 = \frac{1}{n}\sum_{i=1}^n \frac{\partial l}{\partial \theta}(\hat{\theta}_0^b, \hat{\mathbf{Z}}_{i,I|Y}) = B_n^1 - \left(\hat{\theta}_0^b - \theta_0\right) B_n^2 + o_P\left(\hat{\theta}_0^b - \theta_0\right), \text{ with}$$

$$B_n^1 := \frac{1}{n}\sum_{i=1}^n \frac{\partial l}{\partial \theta}\left(\theta_0, \hat{\mathbf{Z}}_{i,I|Y}\right) = \sum_{k=1}^m B_n^{1,k} \text{ and } B_n^2 := -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 l}{\partial \theta^2}\left(\theta_0, \hat{\mathbf{Z}}_{i,I|Y}\right) = \sum_{k=1}^m B_n^{2,k}.$$

We deduce

$$\begin{aligned}\Theta_{n,0} := \sqrt{n}\left(\hat{\theta}_0^b - \theta_0\right) &= \frac{\sqrt{n} B_n^1}{B_n^2} + o_P\left(\Theta_{n,0}\right) = \frac{\sqrt{n}\sum_{k=1}^m B_n^{1,k}}{\sum_{k=1}^m B_n^{2,k}} + o_P\left(\Theta_{n,0}\right) \\ &= \frac{\sqrt{n}\sum_{k=1}^m \psi_{k,1}(G_n)}{\sum_{k=1}^m -\psi_{k,2}(G_n)} + o_P\left(\Theta_{n,0}\right) \\ &= \frac{\sum_{k=1}^m \psi_{k,2}(G_n)\Theta_{n,k}}{\sum_{k=1}^m \psi_{k,2}(G_n)} + o_P\left(\Theta_{n,0}\right). \ \square\end{aligned}$$

Lemma 4.20. *Let ℓ_n be defined by*

$$\ell_n := \sum_{i=1}^n \log\left(\frac{c_{\hat{\theta}_0^b}(\mathbf{Z}_i^*)}{c_{\theta_0}(\mathbf{Z}_i^*)}\right).$$

If there exists a random vector Θ_0 such that $\Theta_{n,0} \Longrightarrow \Theta_0$ under $\mathcal{P}_n$, then we have

$$\ell_n = \Theta_0^T \mathbb{W}^\perp - \frac{1}{2}\Theta_0^T I_0 \Theta_0 + o_P(1),$$

where $\mathbb{W}^\perp \sim \mathcal{N}(0, I_0)$ is independent of the sample $\left(\mathbf{Z}_{i,I|Y}, Y_i\right)_{i=1,\dots,n}$ and I_0 is the Fisher information matrix

$$I_0 := \mathbb{E}_{C_{\theta_0}}\left[\frac{\dot{c}_{\theta_0}^T(\mathbf{Z})\,\dot{c}_{\theta_0}(\mathbf{Z})}{c_{\theta_0}^2(\mathbf{Z})}\right].$$

Proof : By a Taylor expansion, we obtain

$$\begin{aligned}\ell_n &= \sum_{i=1}^{n}\{l\left(\hat{\theta}_0^b, \mathbf{Z}_i^*\right) - l\left(\theta_0, \mathbf{Z}_i^*\right)\} \\ &= \left(\hat{\theta}_0^b - \theta_0\right)^T \sum_{i=1}^{n} \frac{\partial l}{\partial \theta}\left(\theta_0, \mathbf{Z}_i^*\right) + \frac{1}{2}\left(\hat{\theta}_0^b - \theta_0\right)^T \sum_{i=1}^{n} \frac{\partial^2 l}{\partial \theta^2}\left(\theta_0, \mathbf{Z}_i^*\right)\left(\hat{\theta}_0^b - \theta_0\right) + R_n \\ &= \Theta_{n,0}^T \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial l}{\partial \theta}\left(\theta_0, Z_i^*\right)\right] + \frac{1}{2}\Theta_{n,0}^T \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l}{\partial \theta^2}\left(\theta_0, Z_i^*\right)\right] \Theta_{n,0} + R_n.\end{aligned}$$

First, we have

$$\begin{aligned}R_n &\leq Cst||\hat{\theta}_0^b - \theta_0||^3 \sup_{\theta | \|\theta-\theta_0\| \leq \|\hat{\theta}_0^b - \theta_0\|} ||\sum_{i=1}^{n} \frac{\partial^3 l}{\partial \theta^3}\left(\theta, Z_i^*\right)|| \\ &\leq Cst||\Theta_{n,0}||^3 \cdot \sup_{\theta | \|\theta-\theta_0\| \leq \|\hat{\theta}_0^b - \theta_0\|} ||\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^3 l}{\partial \theta^3}\left(\theta, Z_i^*\right)|| \cdot \frac{1}{\sqrt{n}} = O_P\left(\frac{1}{\sqrt{n}}\right),\end{aligned}$$

by Assumption (R). By the usual CLT, we know that $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \partial l/\partial \theta\left(\theta_0, Z_i^*\right) \longrightarrow \mathbb{W}^{\perp}$. $\mathbb{W}^{\perp}$ is independent of $\left(\mathbf{Z}_{i,I|Y}, Y_i\right)_{i=1,\ldots,n}$ as a limit of a sequence of variables that have the same property. Using the law of large numbers, we have also

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 l}{\partial \theta^2}\left(\theta_0, Z_i^*\right) = \frac{1}{n}\sum_{i=1}^{n} \frac{\ddot{c}_\theta}{c_\theta}(Z_i^*) - \frac{\dot{c}_\theta^T \dot{c}_\theta}{c_\theta^2}(Z_i^*) \Longrightarrow 0 - I_0 \ \square$$

4.7.2 Proof of Theorem 4.14

We first reason under $\mathcal{P}_n$ as in Theorem 1 in [57]. By Proposition 4.17, under $\mathcal{P}_n$, there exist two independent and identically distributed processes $\mathbb{A}_G$ and $\mathbb{A}_G^{\perp}$ such that

$$\sqrt{n}\Big(G_n - C_{\theta_0} \otimes P_Y \, , \, G_n^* - C_{\theta_0} \otimes P_{n,Y}\Big) \Longrightarrow (\mathbb{A}_G \, , \, \mathbb{A}_G^{\perp}),$$

weakly in $\Big(\ell^{\infty}\big([0,1]^p \times \{1,\ldots,m\}\big)\Big)^2$. By (iii) of Lemma 4.19, ψ_1 is Hadamard-differentiable and so, using the functional Delta-method, we deduce

$$\sqrt{n}\Big(\psi_1(G_n) - \psi_1(C_{\theta_0} \otimes P_Y) \, , \, \psi_1(G_n^*) - \psi_1(C_{\theta_0} \otimes P_{n,Y})\Big) \Longrightarrow \Big(\dot{\psi}_1(G)(\mathbb{A}_G) \, , \, \dot{\psi}_1(G)(\mathbb{A}_G^{\perp})\Big).$$

By (ii) of Lemma 4.19, $\psi_1(C_{\theta_0} \otimes P_Y) = \psi_1(C_{\theta_0} \otimes P_{n,Y}) = 0$, implying

$$\begin{aligned}&\sqrt{n}\Big(\psi_{1,1}(G_n), \ldots, \psi_{m,1}(G_n) \, , \, \psi_{1,1}(G_n^*), \ldots, \psi_{m,1}(G_n^*)\Big) \\ &\qquad \Longrightarrow \Big(\dot{\psi}_{1,1}(G)(\mathbb{A}_G), \ldots, \dot{\psi}_{m,1}(G)(\mathbb{A}_G) \, , \, \dot{\psi}_{1,1}(G)(\mathbb{A}_G^{\perp}), \ldots, \dot{\psi}_{m,1}(G)(\mathbb{A}_G^{\perp})\Big).\end{aligned}$$

By Slutsky's theorem, we have

$$\begin{aligned}&\sqrt{n}\Big(\frac{\psi_{1,1}(G_n)}{\psi_{1,2}(G_n)}, \ldots, \frac{\psi_{m,1}(G_n)}{\psi_{m,2}(G_n)} \, , \, \frac{\psi_{1,1}(G_n^*)}{\psi_{1,2}(G_n^*)}, \ldots, \frac{\psi_{m,1}(G_n^*)}{\psi_{m,2}(G_n^*)}\Big) \\ &\qquad \Longrightarrow \Big(\frac{\dot{\psi}_{1,1}(G)(\mathbb{A}_G)}{\psi_{1,2}(G)}, \ldots, \frac{\dot{\psi}_{m,1}(G)(\mathbb{A}_G)}{\psi_{m,2}(G)} \, , \, \frac{\dot{\psi}_{1,1}(G)(\mathbb{A}_G^{\perp})}{\psi_{1,2}(G)}, \ldots, \frac{\dot{\psi}_{m,1}(G)(\mathbb{A}_G^{\perp})}{\psi_{m,2}(G)}\Big).\end{aligned}$$

By (i) of Lemma 4.19, the latter convergence result implies

$$\begin{aligned}&\Big(\Theta_{n,1}, \ldots, \Theta_{n,m} \, , \, \Theta_{n,1}^*, \ldots, \Theta_{n,m}^*\Big) \\ &\qquad \Longrightarrow \Big(\frac{\dot{\psi}_{1,1}(G)(\mathbb{A}_G)}{-\psi_{1,2}(G)}, \ldots, \frac{\dot{\psi}_{m,1}(G)(\mathbb{A}_G)}{-\psi_{m,2}(G)} \, , \, \frac{\dot{\psi}_{1,1}(G)(\mathbb{A}_G^{\perp})}{-\psi_{1,2}(G)}, \ldots, \frac{\dot{\psi}_{m,1}(G)(\mathbb{A}_G^{\perp})}{-\psi_{m,2}(G)}\Big) \\ &\qquad =: \Big(\Theta_1, \ldots, \Theta_m \, , \, \Theta_1^{\perp}, \ldots, \Theta_m^{\perp}\Big).\end{aligned}$$

Moreover, $\Big(\Theta_1, \ldots, \Theta_m\Big)$ and $\Big(\Theta_1^\perp, \ldots, \Theta_m^\perp\Big)$ are independent and identically distributed under $\mathcal{P}_n$, by construction. Because of (iv) of Lemma 4.19, $\Theta_{n,0}$ can asymptotically be seen as a mean of the $\Theta_{n,k}$ and this provides

$$\Theta_{n,0} = \frac{\sum_{k=1}^m \psi_{k,2}(G_n)\Theta_{n,k}}{\sum_{k=1}^m \psi_{k,2}(G_n)} \Longrightarrow \frac{\sum_{k=1}^m \psi_{k,2}(G)\Theta_k}{\sum_{k=1}^m \psi_{k,2}(G)} =: \Theta_0.$$

Therefore, by the continuous mapping theorem, we deduce

$$\Big(\Theta_{n,0}, \ldots, \Theta_{n,m}\,,\, \Theta_{n,0}^*, \ldots, \Theta_{n,m}^*\Big) \Longrightarrow \Big(\Theta_0, \ldots, \Theta_m\,,\, \Theta_0^\perp, \ldots, \Theta_m^\perp\Big),$$

and we still have that $\Big(\Theta_0, \ldots, \Theta_m\Big)$ and $\Big(\Theta_0^\perp, \ldots, \Theta_m^\perp\Big)$ are independent and identically distributed under $\mathcal{P}_n$.

Now, we will work under $\mathcal{P}_n^*$ the probability measure over $\big([0,1]^p \times \{1, \ldots, m\}\big)^{\otimes 2n}$ whose density with respect to $\mathcal{P}_n$ is

$$\frac{d\mathcal{P}_n^*}{d\mathcal{P}_n}(\mathbf{z}_1, y_1, \ldots, \mathbf{z}_n, y_n, \mathbf{z}_1^*, y_1^*, \ldots, \mathbf{z}_n^*, y_n^*) = \prod_{i=1}^n \frac{c_{\hat{\theta}_0^b}(\mathbf{z}_i^*)}{c_{\theta_0}(\mathbf{z}_i^*)},$$

where $\hat{\theta}_0^b$ is the estimator of θ_0 when applied to the "sample" $(\mathbf{z}_1, y_1, \ldots, \mathbf{z}_n, y_n)$. We remark that

$$\frac{d\mathcal{P}_n^*}{d\mathcal{P}_n}(\mathbf{Z}_1, Y_1, \ldots, \mathbf{Z}_n, Y_n, \mathbf{Z}_1^*, Y_1^*, \ldots, \mathbf{Z}_n^*, Y_n^*) = \exp(\ell_n).$$

Since we have shown that $\Theta_{n,0} \Longrightarrow \Theta_0$ under $\mathcal{P}_n$, use Lemma 4.20 and obtain

$$\ell_n = \Theta_0^T \mathbb{W}^\perp - \frac{1}{2}\Theta_0^T I_0 \Theta_0 + o_P(1).$$

Therefore, under $\mathcal{P}_n$, we have

$$\left(\frac{d\mathcal{P}_n^*}{d\mathcal{P}_n}, \Theta_{n,0}, \ldots, \Theta_{n,m}, \Theta_{n,0}^*, \ldots, \Theta_{n,m}^*\right) \Longrightarrow \left(\zeta, \Theta_0, \ldots, \Theta_m, \Theta_0^\perp, \ldots, \Theta_m^\perp\right),$$

where $\zeta := \exp\left(\Theta_0^T \mathbb{W}^\perp - \Theta_0^T I_0 \Theta_0/2\right)$. Note that $\mathbb{E}[\zeta] = \mathbb{E}[\mathbb{E}[\zeta|\Theta_0]] = 1$ because Θ_0 and $\mathbb{W}^\perp$ are independent, and $\mathbb{W}^\perp \sim \mathcal{N}(0, I_0)$. This corresponds to condition (iii) of Theorem 3.10.5 of [139], and we deduce $\mathcal{P}_n^*$ is contiguous with respect to $\mathcal{P}_n$. We can then apply Le Cam's Third Lemma (Theorem 3.10.7 of [139]. We get that, under $\mathcal{P}_n^*$,

$$(\Theta_{n,0}, \ldots, \Theta_{n,m}, \Theta_{n,0}^*, \ldots, \Theta_{n,m}^*) \Longrightarrow \left(\tilde{\Theta}_0, \ldots, \tilde{\Theta}_m, \Theta_0^*, \ldots, \Theta_m^*\right),$$

where $\mathbb{E}[\chi(\tilde{\Theta}_{0:m}, \Theta_{0:m}^*)] = \mathbb{E}[\zeta\chi(\Theta_{0:m}, \Theta_{0:m}^\perp)]$ for any simple function χ. Choose w_1 and $w_2 \in \mathbb{R}^{m+1}$ and set $\Sigma := Var\big[\Theta_{0:m}\big]$. Then, we have

$$\begin{aligned}
&\mathbb{E}[\exp(iw_1^T\tilde{\Theta}_{0:m} + iw_2^T\Theta_{0:m}^*)] = \mathbb{E}[\zeta\exp(iw_1^T\Theta_{0:m} + iw_2^T\Theta_{0:m}^\perp)] \\
&= \mathbb{E}[\exp(\Theta_0^T\mathbb{W}^\perp - \Theta_0^T I_0\Theta_0/2 + iw_1^T\Theta_{0:m} + iw_2^T\Theta_{0:m}^\perp)] \\
&= \mathbb{E}\Big[\exp(iw_1^T\Theta_{0:m} - \Theta_0^T I_0\Theta_0/2)\,\mathbb{E}[\exp(\Theta_0^T\mathbb{W}^\perp + iw_2^T\Theta_{0:m}^\perp)\,|\,\Theta_{0:m}]\Big] \\
&= \mathbb{E}\left[\exp(iw_1^T\Theta_{0:m} - \Theta_0^T I_0\Theta_0/2)\exp\left(\frac{1}{2}\left(-w_2^T\Sigma w_2 + \Theta_0^T I_0\Theta_0 + 2iw_2\mathbb{E}[{\Theta_{0:m}^\perp}^T\mathbb{W}^\perp]\Theta_0\right)\right)\right] \\
&= \mathbb{E}\left[\exp\left(iw_1^T\Theta_{0:m} - w_2^T\Sigma w_2/2 + iw_2\mathbb{E}[{\Theta_{0:m}^\perp}^T\mathbb{W}^\perp]\Theta_0\right)\right] \\
&= \mathbb{E}\left[\exp\left(iw_1^T\Theta_{0:m} + iw_2\Theta_{0:m}^\perp + iw_2\mathbb{E}[{\Theta_{0:m}^\perp}^T\mathbb{W}^\perp]\Theta_0\right)\right].
\end{aligned}$$

Therefore, we have proven the following equality:

$$\left(\tilde{\Theta}_0, \ldots, \tilde{\Theta}_m, \Theta_0^*, \ldots, \Theta_m^*\right) \overset{\text{law}}{=} \left(\Theta_0, \ldots, \Theta_m, \Theta_0^\perp + a_0\Theta_0, \ldots, \Theta_m^\perp + a_m\Theta_0\right),$$

where $a_k = \mathbb{E}[{\Theta_k^\perp}^T \mathbb{W}^\perp]$. To finish the proof, it remains to show that $a_k = a_0$ for all $k \in \{1, \ldots, m\}$, i.e.

$$\mathbb{E}[{\Theta_0^\perp}^T \mathbb{W}^\perp] = \mathbb{E}[{\Theta_k^\perp}^T \mathbb{W}^\perp].$$

First, we know from the proof of Lemma 4.19 that $\Theta_{k,n} = -\dot{\psi}_{k,1}(G)(\mathbb{A}_G)/\psi_{k,2}(G) + o_P(1)$, $k = 1, \ldots, m$ and $\Theta_{0,n} = -\dot{\psi}_{0,1}(G)(\mathbb{A}_G)/\psi_{0,2}(G) + o_P(1)$, where

$$\psi_{0,1}(G) := \int \frac{\partial l}{\partial \theta}\left(\theta_0, \left(\frac{\int \mathbb{1}\{z_q^1 \le z_q^2, y^1 = y^2\} dG(\mathbf{z}^1, y^1)}{\int \mathbb{1}\{y^1 = y^2\} dG(\mathbf{z}^1, y^1)}\right)_{q=1,\ldots,p}\right) dG(\mathbf{z}^2, y^2),$$

and

$$\psi_{0,2}(G) := \int \frac{\partial^2 l}{\partial \theta^2}\left(\theta_0, \left(\frac{\int \mathbb{1}\{z_q^1 \le z_q^2, y^1 = y^2\} dG(\mathbf{z}^1, y^1)}{\int \mathbb{1}\{y^1 = y^2\} dG(\mathbf{z}^1, y^1)}\right)_{q=1,\ldots,p}\right) dG(\mathbf{z}^2, y^2).$$

This implies $\Theta_k = -\dot{\psi}_{k,1}(G)(\mathbb{A}_G)/\psi_{k,2}(G)$, $k = 0, \ldots, m$.

Actually, the reasoning is exactly the same when dealing with $\Theta_{k,n}^*$ and $\Theta_k^\perp$, $k = 0, \ldots, m$, replacing $\mathbb{A}_G$ by $\mathbb{A}_G^\perp$. We get

$$\Theta_k^\perp = -\frac{\dot{\psi}_{k,1}(G)(\mathbb{A}_G^\perp)}{\psi_{k,2}(G)}, \text{ and } \Theta_0^\perp = -\frac{\dot{\psi}_{0,1}(G)(\mathbb{A}_G^\perp)}{\psi_{0,2}(G)}.$$

Second, note that, when $k = 1, \ldots, m$,

$$\begin{aligned}
\psi_{k,2}(G) &:= \int \frac{\partial^2 l}{\partial \theta^2}\left(\theta_0, \left(\frac{\int \mathbb{1}\{z_q^1 \le z_q^2, y^1 = k\} dG(\mathbf{z}^1, y^1)}{\int \mathbb{1}\{y^1 = k\}\, dG(\mathbf{z}^1, y^1)}\right)_{q=1,\ldots,p}\right) \mathbb{1}\{y^2 = k\}\, dG(\mathbf{z}^2, y^2) \\
&= \mathbb{P}(Y = k) \int \frac{\partial^2 l}{\partial \theta^2}\left(\theta_0, \left(\int \mathbb{1}\{z_q^1 \le z_q^2\} dC_{\theta_0}(\mathbf{z}^1)\right)_{q=1,\ldots,p}\right) dC_{\theta_0}(\mathbf{z}^2) \\
&= \mathbb{P}(Y = k) \int \frac{\partial^2 l}{\partial \theta^2}(\theta_0, \mathbf{z})\, dC_{\theta_0}(\mathbf{z}^2) \\
&= \mathbb{P}(Y = k)\psi_{0,2}(G).
\end{aligned}$$

Third, let us calculate $\dot{\psi}_{k,1}(G)(h)$, $k = 0, 1, \ldots, m$. From Lemma 4.19, we have

$$\begin{aligned}
\dot{\psi}_{k,1}(G)(h) &= \int \frac{\partial l}{\partial \theta}\left(\theta_0, \mathbf{z}^2\right) \mathbb{1}\{y^2 = k\}\, dh(\mathbf{z}^2, y^2) + \sum_{j=1}^p \int \frac{\partial^2 l}{\partial \theta\, \partial z_j}\left(\theta_0, \mathbf{z}^2\right) \mathbb{1}\{y^2 = k\} \\
&\cdot \left(\frac{\int \mathbb{1}\{z_j^1 \le z_j^2, y^1 = y^2\} dh(\mathbf{z}^1, y^1)}{\int \mathbb{1}\{y^1 = y^2\} dG(\mathbf{z}^1, y^1)} - \frac{\int \mathbb{1}\{z_j^1 \le z_j^2, y^1 = y^2\} dG(\mathbf{z}^1, y^1) \int \mathbb{1}\{y^1 = y^2\} dh(\mathbf{z}^1, y^1)}{\left(\int \mathbb{1}\{y^1 = y^2\} dG(\mathbf{z}^1, y^1)\right)^2}\right) dG(\mathbf{z}^2, y^2).
\end{aligned}$$

for $k = 1, \ldots, m$. Since $G = C_{\theta_0} \otimes F_Y$, we can simplify the latter equalities:

$$\begin{aligned}
\dot{\psi}_{k,1}(G)(h) &= \int \frac{\partial l}{\partial \theta}\left(\theta_0, \mathbf{z}^2\right) \mathbb{1}\{y^2 = k\}\, dh(\mathbf{z}^2, y^2) + \mathbb{P}(Y = k) \sum_{j=1}^p \int \frac{\partial^2 l}{\partial \theta\, \partial z_j}\left(\theta_0, \mathbf{z}^2\right) \\
&\cdot \left(\frac{\int [\mathbb{1}\{z_j^1 \le z_j^2, y^1 = y^2\} - z_j^2 \mathbb{1}\{y^1 = y^2\}] dh(\mathbf{z}^1, y^1)}{\int \mathbb{1}\{y^1 = y^2\} dG(\mathbf{z}^1, y^1)}\right) dC_{\theta_0}(\mathbf{z}^2).
\end{aligned}$$

Since $\dot{\psi}_{0,1}(G)(h) = \sum_{k=1}^{m} \dot{\psi}_{k,1}(G)(h)$, we have

$$\dot{\psi}_{0,1}(G)(h) = \int \frac{\partial l}{\partial \theta}\left(\theta_0, \mathbf{z}^2\right)\, dh(\mathbf{z}^2, y^2) + \sum_{j=1}^{p} \int \frac{\partial^2 l}{\partial \theta\, \partial z_j}\left(\theta_0, \mathbf{z}^2\right) \cdot \left(\frac{\int [\mathbb{1}\{z_j^1 \le z_j^2, y^1 = y^2\} - z_j^2 \mathbb{1}\{y^1 = y^2\}] dh(\mathbf{z}^1, y^1)}{\int \mathbb{1}\{y^1 = y^2\} dG(\mathbf{z}^1, y^1)} \right) dC_{\theta_0}(\mathbf{z}^2) dF_Y(\mathbf{y}^2).$$

Then, we can rewrite $\dot{\psi}_{k,1}(G)(h) = M_1(h,k) + \mathbb{P}(Y = k) M_3(h,k)$ and $\dot{\psi}_{0,1}(G)(h) = M_2(h) + \sum_{k'=1}^{m} \mathbb{P}(Y = k') M_3(h,k')$, where

$$M_1(h,k) := \int \frac{\partial l}{\partial \theta}\left(\theta_0, \mathbf{z}^2\right) \mathbb{1}\{y^2 = k\}\, dh(\mathbf{z}^2, y^2), \quad M_2(h) := \int \frac{\partial l}{\partial \theta}\left(\theta_0, \mathbf{z}^2\right)\, dh(\mathbf{z}^2, y^2),$$
$$M_3(h,k) := \sum_{j=1}^{p} \int \frac{\partial^2 l}{\partial \theta\, \partial z_j}\left(\theta_0, \mathbf{z}^2\right) \left(\frac{\int [\mathbb{1}\{z_j^1 \le z_j^2, y^1 = k\} - z_j^2 \mathbb{1}\{y^1 = k\}] dh(\mathbf{z}^1, y^1)}{\int \mathbb{1}\{y^1 = k\} dG(\mathbf{z}^1, y^1)} \right) dC_{\theta_0}(\mathbf{z}^2).$$

Substituting h by $\mathbb{A}_G^{\perp}$, we get

$$\Theta_k^{\perp} = -\frac{\dot{\psi}_{k,1}(G)(\mathbb{A}_G^{\perp})}{\psi_{k,2}(G)} = -\frac{M_1(\mathbb{A}_G^{\perp}, k)}{\mathbb{P}(Y = k)\psi_{0,2}(G)} - \frac{M_3(\mathbb{A}_G^{\perp}, k)}{\psi_{0,2}(G)},$$
$$\Theta_0^{\perp} = -\frac{\dot{\psi}_{0,1}(G)(\mathbb{A}_G^{\perp})}{\psi_{0,2}(G)} = -\frac{M_2(\mathbb{A}_G^{\perp})}{\psi_{0,2}(G)} - \frac{\sum_{k'=1}^{m} \mathbb{P}(Y = k') M_3(\mathbb{A}_G^{\perp}, k')}{\psi_{0,2}(G)}.$$

Fourth, since $\mathbb{W}^{\perp}$ is the weak limit of $\sum_{i=1}^{n} \frac{\partial l}{\partial \theta}\left(\theta_0, \mathbf{Z}_i^*\right) / \sqrt{n}$ under $\mathcal{P}_n$, this implies $\mathbb{W}^{\perp} = \dot{\psi}_3(G)(\mathbb{A}_G^{\perp})$, with

$$\psi_3(G) = \int \frac{\partial l}{\partial \theta}\left(\theta_0, \mathbf{z}\right)\, dG(\mathbf{z}, y), \text{ and } \dot{\psi}_3(G)(h) = \int \frac{\partial l}{\partial \theta}\left(\theta_0, \mathbf{z}\right)\, dh(\mathbf{z}, y).$$

Finally, by (i) of the following Lemma 4.21, we have $\mathbb{E}[M_1(\mathbb{A}_G^{\perp}, k)^T \mathbb{W}^{\perp}] = \mathbb{P}(Y = k)\mathbb{E}[M_2(\mathbb{A}_G^{\perp})^T \mathbb{W}^{\perp}]$, By (ii) of the latter lemma, we have $\mathbb{E}[M_3(\mathbb{A}_G^{\perp}, k)^T \mathbb{W}^{\perp}] = \mathbb{E}[M_3(\mathbb{A}_G^{\perp}, k')^T \mathbb{W}^{\perp}]$ for all k and k'. Finally, we obtain $\mathbb{E}[{\Theta_0^{\perp}}^T \mathbb{W}^{\perp}] = \mathbb{E}[{\Theta_k^{\perp}}^T \mathbb{W}^{\perp}]$, which finishes the proof. □

Lemma 4.21. *Assume that $\overline{\mathcal{H}}_0^c$ is satisfied. Then,*

(i) For $k = 1, \ldots, m$,

$$\mathbb{E}\left[\int \frac{\partial l}{\partial \theta^T}\left(\theta_0, \mathbf{z}\right) \mathbb{1}\{y = k\}\, d\mathbb{A}_G^{\perp}(\mathbf{z}, y) \int \frac{\partial l}{\partial \theta}\left(\theta_0, \mathbf{z}'\right)\, d\mathbb{A}_G^{\perp}(\mathbf{z}', y') \right] = \mathbb{P}(Y = k)\, \mathbb{E}\left[\int \frac{\partial l}{\partial \theta^T}\left(\theta_0, \mathbf{z}\right)\, d\mathbb{A}_G^{\perp}(\mathbf{z}, y) \int \frac{\partial l}{\partial \theta}\left(\theta_0, \mathbf{z}'\right)\, d\mathbb{A}_G^{\perp}(\mathbf{z}', y') \right].$$

(ii) The expectations

$$\mathbb{E}\left[\int \frac{\partial^2 l}{\partial \theta^T\, \partial z_j}\left(\theta_0, \mathbf{z}^2\right) \left(\frac{\int [\mathbb{1}\{z_j^1 \le z_j^2, y^1 = k\} - z_j^2 \mathbb{1}\{y^1 = k\}] d\mathbb{A}_G^{\perp}(\mathbf{z}^1, y^1)}{\int \mathbb{1}\{y^1 = k\} dG(\mathbf{z}^1, y^1)} \right) dC_{\theta_0}(\mathbf{z}^2) \cdot \int \frac{\partial l}{\partial \theta}\left(\theta_0, \mathbf{z}^3\right)\, d\mathbb{A}_G^{\perp}(\mathbf{z}^3, y^3) \right]$$

do not depend on $k = 1, \ldots, m$.

Proof : (i) By simple calculations, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\int \frac{\partial l}{\partial \theta^T}(\theta_0,\mathbf{z})\,\mathbb{1}\{y=k\}\,d\mathbb{A}_G^{\perp}(\mathbf{z},y)\int \frac{\partial l}{\partial \theta}(\theta_0,\mathbf{z}')\;d\mathbb{A}_G^{\perp}(\mathbf{z}',y')\right]\\
&=\int \frac{\partial l}{\partial \theta^T}(\theta_0,\mathbf{z})\,\mathbb{1}\{y=k\}\frac{\partial l}{\partial \theta}(\theta_0,\mathbf{z}')\,d_{\mathbf{z},y,\,\mathbf{z}',y'}\left(\mathbb{E}\left[\mathbb{A}_G^{\perp}(\mathbf{z},y)\mathbb{A}_G^{\perp}(\mathbf{z}',y')\right]\right)\\
&=\int \frac{\partial l}{\partial \theta^T}(\theta_0,\mathbf{z})\,\mathbb{1}\{y=k\}\frac{\partial l}{\partial \theta}(\theta_0,\mathbf{z}')\,\{\delta_{y'=y}d\mathbb{P}(y)[dC_{\theta_0}(\mathbf{z})\delta_{\mathbf{z}'=\mathbf{z}}+dC_{\theta_0}(\mathbf{z}')\delta_{\mathbf{z}=\mathbf{z}'}]\\
&-dC_{\theta_0}(\mathbf{z})dC_{\theta_0}(\mathbf{z}')d\mathbb{P}(y)d\mathbb{P}(y')\}\\
&=2\mathbb{P}(Y=k)\int \frac{\partial l}{\partial \theta^T}(\theta_0,\mathbf{z})\frac{\partial l}{\partial \theta}(\theta_0,\mathbf{z})\,dC_{\theta_0}(\mathbf{z})-\mathbb{P}(Y=k)\int \frac{\partial l}{\partial \theta^T}(\theta_0,\mathbf{z})\,dC_{\theta_0}(\mathbf{z})\cdot\int \frac{\partial l}{\partial \theta}(\theta_0,\mathbf{z})\,dC_{\theta_0}(\mathbf{z}).
\end{aligned}$$

By summing up the latter identities w.r.t. $k=1,\ldots,m$, we prove (i).

(ii) For convenience, let us write $\phi_2(\mathbf{z}) := \partial^2 l/(\partial\theta^T\,\partial z_j)\,(\theta_0,\mathbf{z})$ and $\phi_3(\mathbf{z}) := \partial l(\theta_0,\mathbf{z})/\partial\theta^T$. We get the result if we prove that

$$A_{1,k} := \mathbb{E}\left[\int \phi_2(\mathbf{z}_2)\left(\frac{\int \mathbb{1}\{z_j^1\le z_j^2\}\mathbb{1}\{y^1=k\}d\mathbb{A}_G^{\perp}(\mathbf{z}^1,y^1)}{\int \mathbb{1}\{y^1=k\}dG(\mathbf{z}^1,y^1)}\right)dC_{\theta_0}(\mathbf{z}^2)\int \phi_3(\mathbf{z}_3)\,d\mathbb{A}_G^{\perp}(\mathbf{z}^3,y^3)\right] \text{ and}$$

$$A_{2,k} := \mathbb{E}\left[\int \phi_2(\mathbf{z}_2)\left(\frac{\int z_j^2\mathbb{1}\{y^1=k\}d\mathbb{A}_G^{\perp}(\mathbf{z}^1,y^1)}{\int \mathbb{1}\{y^1=k\}dG(\mathbf{z}^1,y^1)}\right)dC_{\theta_0}(\mathbf{z}^2)\int \phi_3(\mathbf{z}_3)\,d\mathbb{A}_G^{\perp}(\mathbf{z}^3,y^3)\right]$$

do not depend on k. We will do the task for $A_{1,k}$, $k=1,\ldots,m$, and the calculations will be similar for $A_{2,k}$. Note that

$$\begin{aligned}
A_{1,k}&=\frac{1}{\mathbb{P}(Y=k)}\int \phi_2(\mathbf{z}_2)\mathbb{1}\{z_j^1\le z_j^2\}\mathbb{1}\{y^1=k\}\phi_3(\mathbf{z}_3)\,dC_{\theta_0}(\mathbf{z}^2)d_{\mathbf{z}^1,y^1,\mathbf{z}^3,y^3}\mathbb{E}\left[\mathbb{A}_G^{\perp}(\mathbf{z}^1,y^1)\mathbb{A}_G^{\perp}(\mathbf{z}^3,y^3)\right]\\
&=\frac{1}{\mathbb{P}(Y=k)}\int \phi_2(\mathbf{z}_2)\mathbb{1}\{z_j^1\le z_j^2\}\mathbb{1}\{y^1=k\}\phi_3(\mathbf{z}_3)\,dC_{\theta_0}(\mathbf{z}^2)\\
&\left\{\delta_{y^3=y^1}d\mathbb{P}(y^1)[dC_{\theta_0}(\mathbf{z}^1)\delta_{\mathbf{z}^3=\mathbf{z}^1}+dC_{\theta_0}(\mathbf{z}^3)\delta_{\mathbf{z}^1=\mathbf{z}^3}]-C_{\theta_0}(\mathbf{z}^1)dC_{\theta_0}(\mathbf{z}^3)d\mathbb{P}(y^1)d\mathbb{P}(y^3)\right\}.
\end{aligned}$$

We deduce that

$$\begin{aligned}
A_{1,k}&=2\int \phi_2(\mathbf{z}_2)\mathbb{1}\{z_j^1\le z_j^2\}\phi_3(\mathbf{z}^1)\,dC_{\theta_0}(\mathbf{z}^1)dC_{\theta_0}(\mathbf{z}^2)\\
&\quad-\int \phi_2(\mathbf{z}_2)\mathbb{1}\{z_j^1\le z_j^2\}\phi_3(\mathbf{z}^3)\,dC_{\theta_0}(\mathbf{z}^1)dC_{\theta_0}(\mathbf{z}^2)dC_{\theta_0}(\mathbf{z}^3),
\end{aligned}$$

that does not depend on k. □

4.7.3 Proof of Proposition 4.16

As usual with the nonparametric bootstrap, we rewrite the bootstrapped empirical process by counting the number of times every observation of the initial sample is drawn:

$$d\overline{G}_n^* = \frac{1}{n}\sum_{i=1}^n M_{n,i}\delta_{(\mathbf{Z}_i^*,\mathbf{X}_{i,J})},$$

where $M_{n,i}$ denotes the number of times $(\mathbf{X}_{i,J}^*)$ has been redrawn in a n-size bootstrap resampling with replacement. It is well-known that $M_n := (M_{n,1},\ldots,M_{n,n})$ follows a multinomial distribution $\mathcal{M}(n,n^{-1},\ldots,n^{-1})$: its mean is n and the associated probabilities are $1/n,\ldots,1/n$. In other words,

$$\overline{\mathbb{G}}_n^*(\mathbf{z},\mathbf{x}_J)=\frac{1}{\sqrt{n}}\sum_{i=1}^n M_{n,i}\{\mathbb{1}((\mathbf{Z}_i^*,\mathbf{X}_{i,J})\le(\mathbf{z},\mathbf{x}_J))-C_{\theta_0}(\mathbf{z})F_{n,J}(\mathbf{x}_J)\}.$$

We can remove the dependence between the random components $M_{n,i}$, $i = 1, \ldots, n$ by a "Poissonization" procedure. We mimic van der Vaart and Wellner [139], p.346: instead of drawing n times the initial observations, this is done N_n times, where N_n follows a Poisson distribution with mean n and N_n is independent of the initial sample. Then, the n variables $M_{N_n,1}, \ldots, M_{N_n,n}$ are i.i.d. Poisson random variables with mean one. And we can build the new process as

$$\tilde{\mathbb{G}}_n^*(\mathbf{z}, \mathbf{x}_J) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} M_{N_n,i} \{ \mathbb{1}_{(\mathbf{Z}_i^*, \mathbf{X}_{i,J}) \leq (\mathbf{z}, \mathbf{x}_J)} - C_{\theta_0}(\mathbf{z}) F_{n,J}(\mathbf{x}_J) \}.$$

Actually, the distance between $\overline{\mathbb{G}}_n^*$ and $\tilde{\mathbb{G}}_n^*$ is negligible. Indeed, for every $(\mathbf{z}, \mathbf{x}_J)$,

$$\Delta_n(\mathbf{z}, \mathbf{x}_J) := (\tilde{\mathbb{G}}_n^* - \overline{\mathbb{G}}_n^*)(\mathbf{z}, \mathbf{x}_J) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (M_{N_n,i} - M_{n,i}) \{ \mathbb{1}_{(\mathbf{Z}_i^*, \mathbf{X}_{i,J}) \leq (\mathbf{z}, \mathbf{x}_J)} - C_{\theta_0}(\mathbf{z}) F_{n,J}(\mathbf{x}_J) \}$$

is centered. Moreover, by independence between the observations and by the resampling scheme, we have

$$\begin{aligned}
\mathbb{E}[\|\Delta_n\|_\infty^2] = \mathbb{E}[\sup_{\mathbf{z}, \mathbf{x}_J} \Delta_n^2(\mathbf{z}, \mathbf{x}_J)] &\leq \frac{1}{n^2} \sum_{i,j=1}^{n} \mathbb{E}[|(M_{N_n,i} - M_{n,i})(M_{N_n,j} - M_{n,j})|] \\
&\leq \frac{1}{n^2} \sum_{i,j=1}^{n} \mathbb{E}[(M_{N_n,i} - M_{n,i})^2]^{1/2} \mathbb{E}[(M_{N_n,j} - M_{n,j})^2]^{1/2} \\
&\leq \mathbb{E}[(M_{N_n,1} - M_{n,1})^2],
\end{aligned}$$

because the sequence $(M_{N_n,i} - M_{n,i})_{i=1,\ldots,n}$ is exchangeable. Given $N_n = k$, the i-th variable $|M_{N_n,i} - M_{n,i}|$ is binomial with the parameters $(|k-n|, 1/n)$, i.e.

$$P(|M_{k,i} - M_{n,i}| = l) = C_{|k-n|}^l \frac{1}{n^l} \left(1 - \frac{1}{n}\right)^{|k-n|-l}, \; l = 0, \ldots, |k-n|.$$

Therefore, we obtain

$$\mathbb{E}[(M_{N_n,i} - M_{n,i})^2] = \sum_{k=0}^{\infty} \exp(-n) \frac{n^k}{k!} \left\{ \frac{|k-n|}{n} (1 - \frac{1}{n}) + \left(\frac{|k-n|}{n} \right)^2 \right\}.$$

Simple calculations provide

$$\sum_{k=0}^{\infty} \exp(-n) \frac{n^k}{k!} \frac{|k-n|}{n} = \frac{2n^n}{n!} \exp(-n) \sim \left(\frac{2}{\pi n} \right)^{1/2},$$

by Stirling's formula, and

$$\sum_{k=0}^{\infty} \exp(-n) \frac{n^k}{k!} \left(\frac{k-n}{n} \right)^2 = \frac{\exp(-n)}{n^2} \sum_{k=0}^{\infty} \frac{n^k}{k!} (k(k-1) + k(1-2n) + n^2) = \frac{1}{n}.$$

We deduce $\mathbb{E}[(M_{N_n,i} - M_{n,i})^2] = O(n^{-1/2})$ and $\mathbb{P}(\|\Delta_n\|_\infty > \epsilon) \longrightarrow 0$, when n tends to the infinity, given almost all sequences $\mathcal{S}_n := (\mathbf{Z}_i, \mathbf{X}_{i,J})_{i=1,\ldots,n}$. This means that we can safely replace $\overline{\mathbb{G}}_n^*$ by $\tilde{\mathbb{G}}_n^*$, and the theorem follows if we prove the weak convergence of $(\overline{\mathbb{G}}_n, \tilde{\mathbb{G}}_n^*)$.

Note that we can rewrite

$$\begin{aligned}
\tilde{\mathbb{G}}_n^*(\mathbf{z}, \mathbf{x}_J) = &\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (M_{N_n,i} - 1) \{ \mathbb{1}(\mathbf{Z}_i^*, X_{i,J}) \leq (\mathbf{z}, \mathbf{x}_J) - C_{\theta_0}(\mathbf{z}) F_J(\mathbf{x}_J) \} \\
+ &\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ \mathbb{1}(\mathbf{Z}_i^*, X_{i,J}) \leq (\mathbf{z}, \mathbf{x}_J) - C_{\theta_0}(\mathbf{z}) F_J(\mathbf{x}_J) \} - C_{\theta_0}(\mathbf{z}) \sqrt{n} \left(F_{n,J} - F_J \right)(\mathbf{x}_J) \\
+ &\left(1 - \frac{1}{n} \sum_{i=1}^{n} M_{N_n,i} \right) C_{\theta_0}(\mathbf{z}) \sqrt{n} \left(F_{n,J} - F_J \right)(\mathbf{x}_J) \\
:= &\; \tilde{\mathbb{G}}_{n,1}^*(\mathbf{z}, \mathbf{x}_J) + \tilde{\mathbb{G}}_{n,2}^*(\mathbf{z}, \mathbf{x}_J) - \mathbb{G}_{n,3}(\mathbf{z}, \mathbf{x}_J) + R_n(\mathbf{z}, \mathbf{x}_J).
\end{aligned}$$

Obviously, the last remaining term is $o_P(1)$ uniformly w.r.t. $(\mathbf{z}, \mathbf{x}_J)$, and it can be forgotten. Moreover, since the variables $(M_{N_n,i}-1)_{i=1,\dots,n}$ are i.i.d., centered with variance one and independent of the data, we can invoke some multiplier bootstrap results. Consider we live in the space $\mathcal{W} := [0,1]^p \times [0,1]^p \times \mathbb{R}^{d-p}$ that is related to our observations $W_i := (\mathbf{Z}_i, \mathbf{Z}_i^*, \mathbf{X}_{i,J})$, $i = 1, \dots, n$. The true distribution of W_i under the null is P_W, whose cdf is $C_{\theta_0} \otimes C_{\theta_0} \otimes F_J$. Applying Corollary 2.9.3. in [139], the sequence of processes

$$(\mathbb{W}_n, \mathbb{W}_n^*) := \left(n^{-1/2} \sum_{i=1}^n (\delta_{W_i} - P_W), n^{-1/2} \sum_{i=1}^n (M_{N_n,i} - 1)(\delta_{W_i} - P_W) \right)$$

converges weakly in $\ell^\infty(\mathcal{F}) \times \ell^\infty(\mathcal{F})$ to a vector of independent Gaussian processes, where $\mathcal{F}$ denotes any Donsker class of measurable functions from $\mathcal{W}$ to $\mathbb{R}$.

Now, let us consider the class $\mathcal{F}$ of functions

$$f_{\mathbf{z}_0, \mathbf{z}_0', \mathbf{x}_{J,0}} : (\mathbf{z}, \mathbf{z}', \mathbf{x}_J) \mapsto \mathbb{1}(\mathbf{z} \leq \mathbf{z}_0, \mathbf{z}' \leq \mathbf{z}_0', \mathbf{x}_J \leq \mathbf{x}_{J,0}),$$

for any triplet $(\mathbf{z}_0, \mathbf{z}_0', \mathbf{x}_{J,0})$ in $[0,1]^p \times [0,1]^p \times \overline{\mathbb{R}}^{d-p}$. Note that $\mathcal{F}$ is Donsker, that $\tilde{\mathbb{G}}_{n,1}^*(\mathbf{z}, \mathbf{x}_J) = \mathbb{W}_n^* f_{1,\mathbf{z},\mathbf{x}_J}$, $\tilde{\mathbb{G}}_{n,2}^*(\mathbf{z}, \mathbf{x}_J) = \mathbb{W}_n f_{1,\mathbf{z},\mathbf{x}_J}$ and that $\tilde{\mathbb{G}}_{n,3}^*(\mathbf{z}, \mathbf{x}_J) = C_{\theta_0}(\mathbf{z}) \mathbb{W}_n f_{1,1,\mathbf{x}_J}$. By the permanence of the Donsker property (see Section 2.10 in [139], and the continuity of C_{θ_0}, the process $\tilde{\mathbb{G}}_n^*$ converges in $\ell^\infty([0,1]^p \times \mathbb{R}^{d-p})$ to a gaussian process $\overline{\mathbb{A}}^\perp$. Obviously, $\overline{\mathbb{G}}_n$ tends in distribution in $\ell^\infty([0,1]^p \times \mathbb{R}^{d-p})$ to a Gaussian process $\overline{\mathbb{A}}$, whose covariance function is given by

$$\mathbb{E}\left[\overline{\mathbb{G}}_n(\mathbf{z}, \mathbf{x}_J) \overline{\mathbb{G}}_n(\mathbf{z}', \mathbf{x}_J')\right] = C_{\theta_0}(\mathbf{z} \wedge \mathbf{z}') F_J(\mathbf{x}_J \wedge \mathbf{x}_J') - C_{\theta_0}(\mathbf{z}) F_J(\mathbf{x}_J) C_{\theta_0}(\mathbf{z}') F_J(\mathbf{x}_J'),$$

for every $\mathbf{z}, \mathbf{z}', \mathbf{x}_J, \mathbf{x}_J')$. By some standard calculations, we check that $\mathbb{E}[\tilde{\mathbb{G}}_n^*(\mathbf{z}, \mathbf{x}_J) \tilde{\mathbb{G}}_n^*(\mathbf{z}', \mathbf{x}_J')] = \mathbb{E}[\overline{\mathbb{G}}_n(\mathbf{z}, \mathbf{x}_J) \overline{\mathbb{G}}_n(\mathbf{z}', \mathbf{x}_J')$ for every couples $(\mathbf{z}, \mathbf{x}_J)$ and $(\mathbf{z}', \mathbf{x}_J')$, implying that $\overline{\mathbb{A}}$ and $\overline{\mathbb{A}}^\perp$ have the same covariance functions. Moreover, the two limiting processes $\overline{\mathbb{A}}$ and $\overline{\mathbb{A}}^\perp$ are uncorrelated because

$$\mathbb{E}[\overline{\mathbb{G}}_n(\mathbf{z}, \mathbf{x}_J) \tilde{\mathbb{G}}_n^*(\mathbf{z}', \mathbf{x}_J')] = \mathbb{E}[\overline{\mathbb{G}}_n(\mathbf{z}, \mathbf{x}_J) \mathbb{E}[\tilde{\mathbb{G}}_n^*(\mathbf{z}', \mathbf{x}_J') | \mathcal{S}_n]] = 0,$$

for every couples $(\mathbf{z}, \mathbf{x}_J)$ and $(\mathbf{z}', \mathbf{x}_J')$. Therefore, the $\overline{\mathbb{A}}$ and $\overline{\mathbb{A}}^\perp$ are two independent versions of the same Gaussian process.

Remark 4.22. *If there were no resampling of the observations $\mathbf{X}_{i,J}$ at the first level, this would no longer be true. Indeed, the corresponding bootstrapped process would be given by*

$$\mathbb{G}_n^{**}(\mathbf{z}, \mathbf{x}_J) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{1}(\mathbf{X}_{i,J} \leq \mathbf{x}_J) \left\{ \mathbb{1}(\mathbf{Z}_i^* \leq \mathbf{z}) - C_{\theta_0}(\mathbf{z}) \right\},$$

implying

$$\mathbb{E}\left[\overline{\mathbb{G}}_n^{**}(\mathbf{z}, \mathbf{x}_J) \overline{\mathbb{G}}_n^{**}(\mathbf{z}', \mathbf{x}_J')\right] = F_J(\mathbf{x}_J \wedge \mathbf{x}_J') [C_{\theta_0}(\mathbf{z} \wedge \mathbf{z}') - C_{\theta_0}(\mathbf{z}) C_{\theta_0}(\mathbf{z}')],$$

that is different of $\mathbb{E}\left[\overline{\mathbb{G}}_n(\mathbf{z}, \mathbf{x}_J) \overline{\mathbb{G}}_n(\mathbf{z}', \mathbf{x}_J')\right]$.

To conclude, we apply Corollary 1.4.5. in [139]: for every bounded nonnegative Lipschitz function h and $\tilde{h}$,

$$\begin{aligned} \mathbb{E}[h(\overline{\mathbb{G}}_n) \tilde{h}(\tilde{\mathbb{G}}_n^*)] - \mathbb{E}[h(\overline{\mathbb{A}}) \tilde{h}(\overline{\mathbb{A}}^\perp)] &= \mathbb{E}[h(\overline{\mathbb{G}}_n) \left(\mathbb{E}[\tilde{h}(\tilde{\mathbb{G}}_n^*) | \mathcal{S}_n] - \mathbb{E}[\tilde{h}(\overline{\mathbb{A}}^\perp)] \right)] \\ &+ \mathbb{E}[\left(h(\overline{\mathbb{G}}_n) - \mathbb{E}[h(\overline{\mathbb{A}})]\right)] \mathbb{E}[\tilde{h}(\overline{\mathbb{A}}^\perp)]. \end{aligned}$$

The first (resp. second) term tends to zero by the weak convergence of $\tilde{\mathbb{G}}_n^*$ (resp. $\overline{\mathbb{G}}_n$). This concludes the proof. □

Chapter 5

About kernel-based estimation of conditional Kendall's tau: finite-distance bounds and asymptotic behavior

Abstract

We study nonparametric estimators of conditional Kendall's tau, a measure of concordance between two random variables given some covariates. We prove non-asymptotic bounds with explicit constants, that hold with high probabilities. We provide "direct proofs" of the consistency and the asymptotic law of conditional Kendall's tau. A simulation study evaluates the numerical performance of such nonparametric estimators.

Keywords: Conditional dependence measures, kernel smoothing, conditional Kendall's tau.

Based on [40]: Derumigny, A., & Fermanian, J. D., About kernel-based estimation of conditional Kendall's tau: finite-distance bounds and asymptotic behavior. *ArXiv preprint*, arXiv:1810.06234, 2018.

5.1 Introduction

In the field of dependence modeling, it is common to work with dependence measures. Contrary to usual linear correlations, most of them have the advantage of being defined without any condition on moments, and of being invariant to changes in the underlying marginal distributions. Such summaries of information are very popular and can be explicitly written as functionals of the underlying copulas: Kendall's tau, Spearman's rho, Blomqvist's coefficient... See Nelsen [106] for an introduction. In particular, for more than a century (Spearman (1904), Kendall (1938)), Kendall's tau has become a popular dependence measure in $[-1, 1]$. It quantifies the positive or negative dependence between two random variables X_1 and X_2. Denoting by $C_{1,2}$ the unique underlying copula of (X_1, X_2) that are assumed to be

continuous, their Kendall's tau can be directly defined as

$$\begin{aligned}\tau_{1,2} &:= 4\int_{[0,1]^2} C_{1,2}(u_1,u_2)\, C_{1,2}(du_1,du_2) - 1 \\ &= \mathbb{P}\big((X_{1,1}-X_{2,1})(X_{1,2}-X_{2,2})>0\big) - \mathbb{P}\big((X_{1,1}-X_{2,1})(X_{1,2}-X_{2,2})<0\big),\end{aligned} \tag{5.1}$$

where $(X_{i,1},X_{i,2})_{i=1,2}$ are two independent versions of $\mathbf{X}:=(X_1,X_2)$. This measure is then interpreted as the probability of observing a *concordant pair* minus the probability of observing a *discordant pair*. See [86] for an historical perspective on Kendall's tau. Its inference is discussed in many textbooks (see [71] or [94], e.g.). Its links with copulas and other dependence measures can be found in [106] or [75].

Similar dependence measure can be introduced in a conditional setup, when a p-dimensional covariate $\mathbf{Z}$ is available. When thousands of papers refer to Kendall's tau, only a few of them have considered conditional Kendall's tau (as defined below) until now. The goal is now to model the dependence between the two components X_1 and X_2, given the vector of covariates $\mathbf{Z}$. Logically, we can invoke the conditional copula $C_{1,2|\mathbf{Z}=\mathbf{z}}$ of (X_1,X_2) given $\mathbf{Z}=\mathbf{z}$ for any point $\mathbf{z}\in\mathbb{R}^p$ (see Patton [111, 112]), and the corresponding conditional Kendall's tau would be simply defined as

$$\begin{aligned}\tau_{1,2|\mathbf{Z}=\mathbf{z}} &:= 4\int_{[0,1]^2} C_{1,2|\mathbf{Z}=\mathbf{z}}(u_1,u_2)\, C_{1,2|\mathbf{Z}=\mathbf{z}}(du_1,du_2) - 1 \\ &= \mathbb{P}\big((X_{1,1}-X_{2,1})(X_{1,2}-X_{2,2})>0\big|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}\big) \\ &\qquad - \mathbb{P}\big((X_{1,1}-X_{2,1})(X_{1,2}-X_{2,2})<0\big|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}\big),\end{aligned}$$

where $(X_{i,1},X_{i,2},\mathbf{Z}_i)_{i=1,2}$ are two independent versions of $(X_1,X_2,\mathbf{Z})$. As above, this is the probability of observing a *concordant pair* minus the probability of observing a *discordant pair*, conditionally on $\mathbf{Z}_1$ and $\mathbf{Z}_2$ being both equal to $\mathbf{z}$. Note that, as conditional copulas themselves, conditional Kendall's taus are invariant w.r.t. increasing transformations of the conditional margins X_1 and X_2, given $\mathbf{Z}$. Of course, if $\mathbf{Z}$ is independent of (X_1,X_2) then, for every $\mathbf{z}\in\mathbb{R}^p$, the conditional Kendall's tau $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ is equal to the (unconditional) Kendall's tau $\tau_{1,2}$.

Conditional Kendall's tau, and more generally conditional dependence measures, are of interest per se because they allow to summarize the evolution of the dependence between X_1 and X_2, when the covariate $\mathbf{Z}$ is changing. Surprisingly, their nonparametric estimates have been introduced in the literature only a few years ago ([63],[141],[52]) and their properties have not yet been fully studied in depth. Indeed, until now and to the best of our knowledge, the theoretical properties of nonparametric conditional Kendall's tau estimates have been obtained "in passing" in the literature, as a sub-product of the weak-convergence of conditional copula processes ([141]) or as intermediate quantities that will be "plugged-in" ([50]). Therefore, such properties have been stated under too demanding assumptions. In particular, some assumptions were related to the estimation of conditional margins, while this is not required (Kendall's tau are based on ranks). In this paper, we will directly study nonparametric estimates $\hat{\tau}_{1,2|\mathbf{z}}$ without relying on the theory/inference of copulas. Therefore, we will state their main usual properties of statistical estimates: exponential bounds in probability, consistency, asymptotic normality.

Our $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ has not to be confused with so-called "conditional Kendall's tau" in the case of truncated data ([136], [101]), in the case of semi-competing risk models ([89], [73]), or for other partial information schemes ([27], [81], among others). Indeed, particularly in biostatistics or reliability, the inference of dependence models under truncation/censoring can be led by considering some types of conditional Kendall's tau, given some algebraic relationships among the underlying random variables. This would

induce conditioning by subsets. At the opposite, we will consider point-wise conditioning events only in this paper, under a nonparametric point-of-view. Nonetheless, such point-wise events can be found in the literature, in parametric or semi-parametric frameworks, as for the identifiability of frailty distributions in bivariate proportional models ([107], [100]). Other related papers are [8] or [97], that are dealing with extreme co-movements (bivariate extreme-value theory). There, the the tail conditioning events of Kendall's tau have probabilities that go to zero with the sample size.

In Section 5.2, different kernel-based estimators of the conditional Kendall's tau are proposed. In Section 5.3, the theoretical properties of the latter estimators are proved, first with finite-distance bounds and then under an asymptotic point-of-view. A short simulation study is provided in Section 5.4. Proofs are postponed into the appendix.

5.2 Definition of several kernel-based estimators of $\tau_{1,2|\mathbf{z}}$

Let $(X_{i,1}, X_{i,2}, \mathbf{Z}_i)$, $i = 1, \ldots, n$ be an i.i.d. sample distributed as $(X_1, X_2, \mathbf{Z})$, and $n \geq 2$. Assuming continuous underlying distributions, there are several equivalent ways of defining the conditional Kendall's tau:

$$\begin{aligned}
\tau_{1,2|\mathbf{Z}=\mathbf{z}} &= 4\,\mathbb{P}\big(X_{1,1} > X_{2,1}, X_{1,2} > X_{2,2}\big|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) - 1 \\
&= 1 - 4\,\mathbb{P}\big(X_{1,1} > X_{2,1}, X_{1,2} < X_{2,2}\big|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) \\
&= \mathbb{P}\big((X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2}) > 0\big|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) \\
&\qquad - \mathbb{P}\big((X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2}) < 0\big|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big).
\end{aligned}$$

Motivated by each of the latter expressions, we introduce several kernel-based estimators of $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$:

$$\begin{aligned}
\hat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}} &:= 4\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,n}(\mathbf{z})w_{j,n}(\mathbf{z})\mathbb{1}\big\{X_{i,1} < X_{j,1}, X_{i,2} < X_{j,2}\big\} - 1, \\
\hat{\tau}^{(2)}_{1,2|\mathbf{Z}=\mathbf{z}} &:= \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,n}(\mathbf{z})w_{j,n}(\mathbf{z})\Big(\mathbb{1}\big\{(X_{i,1} - X_{j,1}).(X_{i,2} - X_{j,2}) > 0\big\} \\
&\qquad\qquad - \mathbb{1}\big\{(X_{i,1} - X_{j,1}).(X_{i,2} - X_{j,2}) < 0\big\}\Big), \\
\hat{\tau}^{(3)}_{1,2|\mathbf{Z}=\mathbf{z}} &:= 1 - 4\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,n}(\mathbf{z})w_{j,n}(\mathbf{z})\mathbb{1}\big\{X_{i,1} < X_{j,1}, X_{i,2} > X_{j,2}\big\},
\end{aligned}$$

where $\mathbb{1}$ denotes the indicator function, $w_{i,n}$ is a sequence of weights given by

$$w_{i,n}(\mathbf{z}) = \frac{K_h(\mathbf{Z}_i - \mathbf{z})}{\sum_{j=1}^{n} K_h(\mathbf{Z}_j - \mathbf{z})}, \tag{5.2}$$

with $K_h(\cdot) := h^{-p}K(\cdot/h)$ for some kernel K on $\mathbb{R}^p$, and $h = h(n)$ denotes a usual bandwidth sequence that tends to zero when $n \to \infty$. In this paper, we have chosen usual Nadaraya-Watson weights. Obviously, there are alternatives (local linear, Priestley-Chao, Gasser-Müller, etc., weight), that would lead to different theoretical results.

The estimators $\hat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}}$, $\hat{\tau}^{(2)}_{1,2|\mathbf{Z}=\mathbf{z}}$ and $\hat{\tau}^{(3)}_{1,2|\mathbf{Z}=\mathbf{z}}$ look similar, but they are nevertheless different, as shown in Proposition 5.1. These differences are due to the fact that all the $\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}}$, $k = 1, 2, 3$ are affine transformations of a double-indexed sum, on every pair (i, j), including the diagonal terms where $i = j$. The treatment of these diagonal terms is different for each of the three estimators defined above. Indeed,

setting $s_n := \sum_{i=1}^n w_{i,n}^2(\mathbf{z})$, it can be easily proved that $\hat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}}$ takes values in the interval $[-1\,,\,1-2s_n]$, $\hat{\tau}^{(2)}_{1,2|\mathbf{Z}=\mathbf{z}}$ in $[-1+s_n\,,\,1-s_n]$, and $\hat{\tau}^{(3)}_{1,2|\mathbf{Z}=\mathbf{z}}$ in $[-1+2s_n\,,\,1]$. Moreover, there exists a direct relationship between these estimators, given by the following proposition.

Proposition 5.1. *Almost surely,* $\hat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}} + s_n = \hat{\tau}^{(2)}_{1,2|\mathbf{Z}=\mathbf{z}} = \hat{\tau}^{(3)}_{1,2|\mathbf{Z}=\mathbf{z}} - s_n$*, where* $s_n := \sum_{i=1}^n w_{i,n}^2(\mathbf{z})$.

This proposition is proved in Section 5.5.1. As a consequence, we can rescale easily the previous estimators so that the new estimator will take values in the whole interval $[-1,1]$. This would yield

$$\tilde{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} := \frac{\hat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}}}{1-s_n} + \frac{s_n}{1-s_n} = \frac{\hat{\tau}^{(2)}_{1,2|\mathbf{Z}=\mathbf{z}}}{1-s_n} = \frac{\hat{\tau}^{(3)}_{1,2|\mathbf{Z}=\mathbf{z}}}{1-s_n} - \frac{s_n}{1-s_n}.$$

Note that none of the latter estimators depends on any estimation of conditional marginal distributions. In other words, we only have to choose conveniently the weights $w_{i,n}$ to obtain an estimator of the conditional Kendall's tau. This is coherent with the fact that conditional Kendall's taus are invariant with respect to conditional marginal distributions. Moreover, note that, in the definition of our estimators, the inequalities are strict (there are no terms corresponding to the cases $i=j$). This is inline with the definition of (conditional) Kendall's tau itself through concordant/discordant pairs of observations.

The definition of $\hat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}}$ can be motivated as follows. For $j=1,2$, let $\hat{F}_{j|\mathbf{Z}}(\cdot|\mathbf{Z}=\mathbf{z})$ be an estimator of the conditional cdf of X_j given $\mathbf{Z}=\mathbf{z}$. Then, a usual estimator of the conditional copula of X_1 and X_2 given $\mathbf{Z}=\mathbf{z}$ is

$$\hat{C}_{1,2|\mathbf{Z}}(u_1,u_2|\mathbf{Z}=\mathbf{z}) := \sum_{i=1}^n w_{i,n}(\mathbf{z}) \mathbb{1}\big\{\hat{F}_{1|\mathbf{Z}}(X_{i,1}|\mathbf{Z}=\mathbf{z}) \le u_1\,,\,\hat{F}_{2|\mathbf{Z}}(X_{i,2}|\mathbf{Z}=\mathbf{z}) \le u_2\big\}.$$

See [141] or [52], e.g. The latter estimator of the conditional copula can be plugged into (5.1) to define an estimator of the conditional Kendall's tau itself:

$$\begin{aligned}\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} &:= 4\int \hat{C}_{1,2|\mathbf{Z}}(u_1,u_2|\mathbf{Z}=\mathbf{z})\,\hat{C}_{1,2|\mathbf{Z}}(du_1,du_2|\mathbf{Z}=\mathbf{z}) - 1 \\ &= 4\sum_{j=1}^n w_{j,n}(\mathbf{z})\hat{C}_{1,2|\mathbf{Z}}\big(\hat{F}_{1|\mathbf{Z}}(X_{j,1}|\mathbf{Z}=\mathbf{z}), \hat{F}_{2|\mathbf{Z}}(X_{j,2}|\mathbf{Z}=\mathbf{z})\big|\mathbf{Z}=\mathbf{z}\big) - 1.\end{aligned} \tag{5.3}$$

Since the functions $\hat{F}_{j|\mathbf{Z}}(\cdot|\mathbf{Z}=\mathbf{z})$ are non-decreasing, this reduces to

$$\begin{aligned}\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} &= 4\sum_{i=1}^n\sum_{j=1}^n w_{i,n}(\mathbf{z})w_{j,n}(\mathbf{z})\mathbb{1}\big\{X_{i,1}\le X_{j,1}, X_{i,2}\le X_{j,2}\big\} - 1 \\ &= 4\sum_{i=1}^n\sum_{j=1}^n w_{i,n}(\mathbf{z})w_{j,n}(\mathbf{z})\mathbb{1}\big\{X_{i,1}< X_{j,1}, X_{i,2}< X_{j,2}\big\} - 1 + o_P(1) = \hat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}} + o_P(1).\end{aligned}$$

Veraverbeke et al. [141], Subsection 3.2, introduced their estimator of $\tau_{1,2|\mathbf{z}}$ by (5.3). By the functional Delta-Method, they deduced its asymptotic normality as a sub-product of the weak convergence of the process $\sqrt{nh}\big(\hat{C}_{1,2|\mathbf{Z}}(\cdot,\cdot|z) - C_{1,2|\mathbf{Z}}(\cdot,\cdot|z)\big)$ when $\mathbf{Z}$ is univariate. In our case, we will obtain more and stronger theoretical properties of $\hat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}}$ under weaker conditions by a more direct analysis based on ranks. In particular, we will not require any regularity condition on the conditional marginal distributions, contrary to [141]. Indeed, in the latter paper, it is required that $F_{j|\mathbf{Z}}(\cdot|\mathbf{Z}=\mathbf{z})$ has to be two times continuously differentiable (assumption $\tilde{R}3$) and its inverse has to be continuous (assumption $R1$). This is not satisfied for some simple univariate cdf as $F_j(t) = t\mathbb{1}(t\in[0,1])/2 + \mathbb{1}(t\in(1,2])/2 + t\mathbb{1}(t\in(2,4])/4 + \mathbb{1}(t>4)$, for instance. Note that We could similarly justify $\hat{\tau}^{(3)}_{1,2|\mathbf{Z}=\mathbf{z}}$ in a similar way by considering conditional survival copulas.

Let us define g_1, g_2, g_3 by

$$g_1(\mathbf{X}_i, \mathbf{X}_j) := 4\mathbb{1}\big\{X_{i,1} < X_{j,1}, X_{i,2} < X_{j,2}\big\} - 1,$$
$$g_2(\mathbf{X}_i, \mathbf{X}_j) := \mathbb{1}\big\{(X_{i,1} - X_{j,1}) \times (X_{i,2} - X_{j,2}) > 0\big\} - \mathbb{1}\big\{(X_{i,1} - X_{j,1}) \times (X_{i,2} - X_{j,2}) < 0\big\},$$
$$g_3(\mathbf{X}_i, \mathbf{X}_j) := 1 - 4\mathbb{1}\big\{X_{i,1} < X_{j,1}, X_{i,2} > X_{j,2}\big\},$$

where for $i = 1, \ldots, n$, we set $\mathbf{X}_i := (X_{i,1}, X_{i,2})$. Clearly, $\hat{\tau}^{(k)}_{1,2|\mathbf{z}}$ is a smoothed estimator of $\mathbb{E}[g_k(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}]$, $k = 1, 2, 3$. The choice of the bandwidth h can be done in a data-driven way following the general conditional U-statistics framework detailed in Dony and Mason [45, Section 2]. Indeed, for any $k \in \{1, 2, 3\}$ and $\mathbf{z} \in \mathcal{Z}$, denote by $\hat{\tau}^{(h,k)}_{-(i,j),\,1,2|\mathbf{Z}=\mathbf{z}}$ the estimator $\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}}$ that is made with the smoothing parameter h and our dataset where the i-th and j-th observations have been removed. As a consequence, the random function $\hat{\tau}^{(h,k)}_{-(i,j),\,1,2|\mathbf{Z}=\cdot}$ is independent of $\big((\mathbf{X}_i, \mathbf{Z}_i), (\mathbf{X}_j, \mathbf{Z}_j)\big)$. As usual with kernel methods, the bandwidth $\hat{h}$ has to be chosen. It would be tempting to propose h as the minimizer of the cross-validation criteria

$$CV_{DM}(h) := \frac{2}{n(n-1)} \sum_{i,j=1}^{n} \Big(g_k(\mathbf{X}_i, \mathbf{X}_j) - \hat{\tau}^{(h,k)}_{-(i,j),\,1,2|\mathbf{Z}=(\mathbf{Z}_i+\mathbf{Z}_j)/2}\Big)^2 K_h(\mathbf{Z}_i - \mathbf{Z}_j),$$

for $k = 1, 2, 3$ or for $\tilde{\tau}_{1,2|\mathbf{Z}=\cdot}$. The latter criterion would be a "naively localized" version of the usual cross-validation method. Unfortunately, we observe that the function $h \mapsto CV_{DM}(h)$ is most often decreasing in the range of realistic bandwidth values. If we remove the weight $K_h(\mathbf{Z}_i - \mathbf{Z}_j)$, then there is no reason why $g_k(\mathbf{X}_i, \mathbf{X}_j)$ should be equal on average to $\hat{\tau}^{(k)}_{-(i,j),\,1,2|\mathbf{Z}=(\mathbf{Z}_i+\mathbf{Z}_j)/2}$, and we are not interested in the prediction of concordance/discordance pairs for which the Z_i and Z_j are far apart. Therefore, a modification of this criteria is necessary. We propose to separate the choice of h for the terms $g_k(\mathbf{X}_i, \mathbf{X}_j) - \hat{\tau}^{(h,k)}_{-(i,j),\,1,2|\mathbf{Z}=(\mathbf{Z}_i+\mathbf{Z}_j)/2}$ and the selection of the "convenient pairs" of observations (i, j). This leads to the new criterion

$$CV_{\tilde{h}}(h) := \frac{2}{n(n-1)} \sum_{i,j=1}^{n} \Big(g_k(\mathbf{X}_i, \mathbf{X}_j) - \hat{\tau}^{(h,k)}_{-(i,j),\,1,2|\mathbf{Z}=(\mathbf{Z}_i+\mathbf{Z}_j)/2}\Big)^2 \tilde{K}_{\tilde{h}}(\mathbf{Z}_i - \mathbf{Z}_j), \tag{5.4}$$

with a potentially different kernel $\tilde{K}$ and a new fixed tuning parameter $\tilde{h}$. Even if more complex procedures are possible, we suggest to simply choose $\tilde{K}(\mathbf{z}) := \mathbb{1}\{|\mathbf{z}|_\infty \leq 1\}$ and to calibrate $\tilde{h}$ so that only a fraction of the pairs (i, j) have non-zero weight. In practice, set $\tilde{h}$ as the empirical quantile of $\big(\{|\mathbf{Z}_i - \mathbf{Z}_j|_\infty : 1 \leq i < j \neq n\}$ of order $2N_{pairs}/(n(n-1))$, where N_{pairs} is the number of pairs we want to keep.

Note that such dependence measures are of interest for the purpose of estimating (conditional or unconditional) copula models too. Indeed, several popular parametric families of copulas have a simple one-to-one mapping between their parameter and the associated Kendall's tau (or Spearman's rho): Gaussian, Student with a fixed degree of freedom, Clayton, Gumbel and Frank copulas, etc. Then, assume for instance that the conditional copula $C_{1,2|\mathbf{Z}=\mathbf{z}}$ belongs is a Gaussian copula with a parameter $\rho(\mathbf{z})$. Then, by estimating its conditional Kendall's tau $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$, we get an estimate of the corresponding parameter $\rho(\mathbf{z})$, and finally of the conditional copula itself. See [119], e.g.

5.3 Theoretical results

5.3.1 Finite distance bounds

Hereafter, we will consider the behavior of conditional Kendall's tau estimates given $\mathbf{Z} = \mathbf{z}$ belongs to some fixed open subset $\mathcal{Z}$ in $\mathbb{R}^p$. For the moment, let us state an instrumental result that is of interest per

se. Let $\hat{f}_{\mathbf{Z}}(\mathbf{z}) := n^{-1}\sum_{j=1}^{n} K_h(\mathbf{Z}_j - \mathbf{z})$ be the usual kernel estimator of the density $f_{\mathbf{Z}}$ of the conditioning variable $\mathbf{Z}$. Note that the estimators $\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}}$, $k = 1, \ldots, 3$ are well-behaved only whenever $\hat{f}_{\mathbf{Z}}(\mathbf{z}) > 0$. Denote the joint density of $(\mathbf{X}, \mathbf{Z})$ by $f_{\mathbf{X},\mathbf{Z}}$. In our study, we need some usual conditions of regularity.

Assumption 5.3.1. *The kernel K is bounded, and set $\|K\|_\infty =: C_K$. It is symmetrical and satisfies $\int K = 1$, $\int |K| < \infty$. This kernel is of order α for some integer $\alpha > 1$: for all $j = 1, \ldots, \alpha - 1$ and every indices $i_1, \ldots, i_j$ in $\{1, \ldots, p\}$, $\int K(\mathbf{u}) u_{i_1} \ldots u_{i_j}\, d\mathbf{u} = 0$. Moreover, $\mathbb{E}[K_h(\mathbf{Z} - \mathbf{z})] > 0$ for every $\mathbf{z} \in \mathcal{Z}$ and $h > 0$. Set $\tilde{K}(\cdot) := K^2(\cdot)/\int K^2$ and $\|\tilde{K}\|_\infty =: C_{\tilde{K}}$.*

Assumption 5.3.2. *$f_{\mathbf{Z}}$ is α-times continuously differentiable on $\mathcal{Z}$ and there exists a constant $C_{K,\alpha} > 0$ s.t., for all $\mathbf{z} \in \mathcal{Z}$,*

$$\int |K|(\mathbf{u}) \sum_{i_1,\ldots,i_\alpha=1}^{p} |u_{i_1} \ldots u_{i_\alpha}| \sup_{t\in[0,1]} \Big| \frac{\partial^\alpha f_{\mathbf{Z}}}{\partial z_{i_1} \ldots \partial z_{i_\alpha}}(\mathbf{z} + th\mathbf{u}) \Big|\, d\mathbf{u} \le C_{K,\alpha}.$$

Moreover, $C_{\tilde{K},2}$ denotes a similar constant replacing K by $\tilde{K}$ and α by two.

Assumption 5.3.3. *There exist two positive constants $f_{\mathbf{Z},min}$ and $f_{\mathbf{Z},max}$ such that, for every $\mathbf{z} \in \mathcal{Z}$, $f_{\mathbf{Z},min} \le f_{\mathbf{Z}}(\mathbf{z}) \le f_{\mathbf{Z},max}$.*

Proposition 5.2. *Under Assumptions 5.3.1-5.3.3 and if $C_{K,\alpha} h^\alpha/\alpha! < f_{\mathbf{Z},min}$, for any $\mathbf{z} \in \mathcal{Z}$, the estimator $\hat{f}_{\mathbf{Z}}(\mathbf{z})$ is strictly positive with a probability larger than*

$$1 - 2\exp\Big(- nh^p \big(f_{\mathbf{Z},min} - C_{K,\alpha} h^\alpha/\alpha!\big)^2 \,/\, \big(2 f_{\mathbf{Z},max} \int K^2 + (2/3) C_K (f_{\mathbf{Z},min} - C_{K,\alpha} h^\alpha/\alpha!)\big)\Big).$$

The latter proposition is proved in Section 5.5.2. It guarantees that our estimators $\hat{\tau}^{(k)}_{1,2|\mathbf{z}}$, $k = 1, \ldots, 3$, are well-behaved with a probability close to one.

The next regularity assumption is necessary to explicitly control the bias of $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}$.

Assumption 5.3.4. *For every $\mathbf{x} \in \mathbb{R}^2$, $\mathbf{z} \mapsto f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}, \mathbf{z})$ is differentiable on $\mathcal{Z}$ almost everywhere up to the order α. For every $0 \le k \le \alpha$ and every $1 \le i_1, \ldots, i_\alpha \le p$, let*

$$\mathcal{H}_{k,\vec{\iota}}(\mathbf{u}, \mathbf{v}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{z}) := \sup_{t\in[0,1]} \left| \frac{\partial^k f_{\mathbf{X},\mathbf{Z}}}{\partial z_{i_1} \ldots \partial z_{i_k}} \Big(\mathbf{x}_1, \mathbf{z} + th\mathbf{u}\Big) \frac{\partial^{\alpha-k} f_{\mathbf{X},\mathbf{Z}}}{\partial z_{i_{k+1}} \ldots \partial z_{i_\alpha}} \Big(\mathbf{x}_2, \mathbf{z} + th\mathbf{v}\Big) \right|,$$

denoting $\vec{\iota} = (i_1, \ldots, i_\alpha)$. Assume that $\mathcal{H}_{k,\vec{\iota}}(\mathbf{u}, \mathbf{v}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{z})$ is integrable and there exists a finite constant $C_{\mathbf{XZ},\alpha} > 0$ such that, for every $\mathbf{z} \in \mathcal{Z}$ and every $h < 1$,

$$\int |K|(\mathbf{u}) |K|(\mathbf{v}) \sum_{k=0}^{\alpha} \binom{\alpha}{k} \sum_{i_1,\ldots,i_\alpha=1}^{p} \mathcal{H}_{k,\vec{\iota}}(\mathbf{u}, \mathbf{v}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{z}) |u_{i_1} \ldots u_{i_k} v_{i_{k+1}} \ldots v_{i_\alpha}|\, d\mathbf{u}\, d\mathbf{v}\, d\mathbf{x}_1\, d\mathbf{x}_2$$

is less than $C_{\mathbf{XZ},\alpha}$.

The next three propositions state pointwise and inform exponential inequalities for the estimators $\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}}$, when $k = 1, 2, 3$. They are proved in Section 5.5.3. We will denote $c_1 := c_3 := 4$ and $c_2 := 2$.

Proposition 5.3 (Exponential bound with explicit constants)**.** *Under Assumptions 5.3.1-5.3.4, for every $t > 0$ such that $C_{K,\alpha} h^\alpha/\alpha! + t \le f_{\mathbf{Z},min}/2$ and every $t' > 0$, if $C_{\tilde{K},2} h^2 < f_{\mathbf{z}}(\mathbf{z})$, we have*

$$\begin{aligned}
&\mathbb{P}\left(|\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}}| > \frac{c_k}{f^2_{\mathbf{z}}(\mathbf{z})} \Big(\frac{C_{\mathbf{XZ},\alpha} h^\alpha}{\alpha!} + \frac{3 f_{\mathbf{z}}(\mathbf{z}) \int K^2}{2nh^p} + t' \Big) \times \left(1 + \frac{16 f^2_{\mathbf{Z}}(\mathbf{z})}{f^3_{\mathbf{Z},min}} \Big(\frac{C_{K,\alpha} h^\alpha}{\alpha!} + t \Big) \right) \right) \\
&\le 2\exp\Big(- \frac{nh^p t^2}{2 f_{\mathbf{Z},max} \int K^2 + (2/3) C_K t} \Big) + 2\exp\Big(- \frac{(n-1) h^{2p} t'^2}{4 f^2_{\mathbf{Z},max} (\int K^2)^2 + (8/3) C^2_K t'} \Big) \\
&+ 2\exp\Big(- \frac{nh^p (f_{\mathbf{z}}(\mathbf{z}) - C_{\tilde{K},2} h^2)^2}{8 f_{\mathbf{Z},max} \int \tilde{K}^2 + 4 C_{\tilde{K}} (f_{\mathbf{z}}(\mathbf{z}) - C_{\tilde{K},2} h^2)/3} \Big),
\end{aligned}$$

for any $\mathbf{z} \in \mathcal{Z}$ and every $k = 1, 2, 3$.

Alternatively, we could obtain a better rate of approximation, at the price of managing unknown (universal) constants instead of explicit constants.

Proposition 5.4 (Alternative exponential bound without explicit constants)**.** *Under Assumptions 5.3.1-5.3.4, for every* $t>0$ *such that* $C_{K,\alpha}h^\alpha/\alpha!+t \leq f_{\mathbf{Z},min}/2$ *and every* $t'>0$ *s.t.* $t' \leq 2h^p(\int K^2)^3 f^3_{\mathbf{Z},max}/C_K^4$*, there exist some universal constants* C_2 *and* α_2 *s.t.*

$$\begin{aligned}
&\mathbb{P}\left(|\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}}| > \frac{c_k}{f^2_{\mathbf{z}}(\mathbf{z})}\Big(\frac{C_{\mathbf{XZ},\alpha}h^\alpha}{\alpha!} + \frac{3f_{\mathbf{z}}(\mathbf{z})\int K^2}{2nh^p} + t'\Big) \times \left(1 + \frac{16 f^2_{\mathbf{Z}}(\mathbf{z})}{f^3_{\mathbf{Z},min}}\Big(\frac{C_{K,\alpha}h^\alpha}{\alpha!} + t\Big)\right)\right) \\
&\leq 2\exp\Big(-\frac{nh^p t^2}{2f_{\mathbf{Z},max}\int K^2 + (2/3)C_K t}\Big) + 2\exp\left(-\frac{nh^p(f_{\mathbf{z}}(\mathbf{z}) - C_{\tilde{K},2}h^2)^2}{8f_{\mathbf{Z},max}\int \tilde{K}^2 + 4C_{\tilde{K}}(f_{\mathbf{z}}(\mathbf{z}) - C_{\tilde{K},2}h^2)/3}\right) \\
&+ 2\exp\Big(\frac{nh^p t^2}{32\int K^2(\int |K|)^2 f^3_{\mathbf{Z},max} + 8C_K\int |K| f_{\mathbf{Z},max} t/3}\Big) + C_2\exp\left(-\frac{\alpha_2 nh^p t'}{8f_{\mathbf{Z},max}(\int K^2)}\right),
\end{aligned}$$

for any $\mathbf{z}\in\mathcal{Z}$ *and every* $k=1,2,3$*, if* $C_{\tilde{K},2}h^2 < f_{\mathbf{Z}}(\mathbf{z})$ *and* $6h^p\big(\int |K|\big)^2 f_{\mathbf{z},max} < \int K^2$*.*

Remark 5.5. *In Propositions 5.2, 5.3 and 5.4, when the support of* K *is included in* $[-c,c]^p$ *for some* $c>0$*,* $f_{\mathbf{Z},max}$ *can be replaced by a local bound* $\sup_{\tilde{\mathbf{z}}\in\mathcal{V}(\mathbf{z},\epsilon)} f_{\mathbf{Z}}(\tilde{\mathbf{z}})$*, denoting by* $\mathcal{V}(\mathbf{z},\epsilon)$ *a closed ball of center* $\mathbf{z}$ *and any radius* $\epsilon>0$*, when* $h\,c<\epsilon$*. Moreover, if it is not guaranteed that* $\mathbb{E}[K_h(\mathbf{Z}-\mathbf{z})]$ *is positive, the results above apply, replacing* $2/3$ *by* $4/3$ *in the denominators of inequality in Proposition 5.3.*

As a corollary, the two latter result yield the weak consistency of $\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}}$ for every $\mathbf{z}\in\mathcal{Z}$, when $nh^{2p}\to\infty$ (choose the constants t and $t'\sim h^p$ sufficiently small, in Proposition 5.4, e.g.).

It is possible to obtain uniform bounds, by slightly strengthening our assumptions. Note that this next result will be true if n is sufficiently large, when Proposition 5.4 was true for every n.

Assumption 5.3.5. *The kernel* K *is Lipschitz on* $(\mathcal{Z},\|\cdot\|_\infty)$*, with a constant* λ_K *and* $\mathcal{Z}$ *is a subset of an hypercube in* $\mathbb{R}^p$ *whose volume is denoted by* $\mathcal{V}$*. Moreover,* K *and* K^2 *are regular in the sense of [66] or [47].*

Proposition 5.6 (Uniform exponential bound)**.** *Under the assumptions 5.3.1-5.3.5, there exist some constants* L_K *and* $\overline{C}_K$ *(resp.* $L_{\tilde{K}}$ *and* $\overline{C}_{\tilde{K}}$*) that depend only on the VC characteristics of* K *(resp.* $\tilde{K}$*), s.t., for every* $\mu\in(0,1)$ *such that* $\mu f_{\mathbf{z},min} < C_{\mathbf{XZ},\alpha}h^\alpha/\alpha! + b_K\int K^2 f_{\mathbf{Z},max}/C_K$*, if* $f_{\mathbf{Z},max} < \tilde{C}_{\mathbf{XZ},2}h^2/2 + b_{\tilde{K}}\int \tilde{K}^2 f_{\mathbf{Z},max}/C_{\tilde{K}}$*,*

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{Z}}|\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}}| > \frac{c_k}{f^2_{\mathbf{z},min}(1-\mu)^2}\left(\frac{C_{\mathbf{XZ},\alpha}h^\alpha}{\alpha!} + \frac{3f_{\mathbf{z},max}\int K^2}{2nh^p} + t\right)\right) \\
&\leq \quad L_K\exp\big(-C_{f,K}nh^p\big(\mu f_{\mathbf{z},min} - \frac{C_{\mathbf{XZ},\alpha}h^\alpha}{\alpha!}\big)^2\big) \\
&+ \quad C_2 D\exp\left(-\frac{\alpha_2 nthp^p}{8f_{\mathbf{Z},max}(\int K^2)}\right) + L_{\tilde{K}}\exp\big(-C_{f,\tilde{K}}nh^p(f_{\mathbf{z},max} - \tilde{C}_{\mathbf{XZ},2}h^2)^2/4\big) \\
&+ \quad 2\exp\big(-\frac{A_2 nh^p t^2 C_K^{-4}}{16^2 A_1^2\int K^2 f^3_{\mathbf{z},max}(\int |K|)^2}\big) + 2\exp(-\frac{A_2 nh^p t}{16C_K^2 A_1}),
\end{aligned}$$

for n *sufficiently large,* $k=1,2,3$*, and for every* $t>0$ *s.t.*

$$t \leq 2h^p(\int K^2)^3 f^3_{\mathbf{Z},max}/C_K^4,$$

$$-16A_1C_K^2\frac{A_{\overline{g}}\int K^2 f^3_{\mathbf{z},max}(\int |K|)^2}{n^{1/2}h^{p/2}}\ln(h^p\int K^2 f^3_{\mathbf{z},max}(\int |K|)^2) < t, \text{ and}$$

$$nh^p t \geq \big(\int K^2\big) f_{\mathbf{z},max} M_2(p+\beta)^{3/2}\log\Big(\frac{4C_K^2}{h^p f_{\mathbf{z},max}\int K^2}\Big),\ \beta = \max\big(0,\frac{\log D}{\log n}\big),\ D := \lceil \mathcal{V}(\frac{4C_K\lambda_K}{h})^p\rceil,$$

for some universal constants C_2,α_2,M_2,A_1,A_2 *and a constant* $A_{\overline{g}}$ *that depends on* K *and* $f_{\mathbf{z},max}$*.*

We have denoted $C_{f,K} := \log(1 + b_K/(4L_K))/(L_K b_K f_{\mathbf{z},max} \int K^2)$, for an arbitrarily chosen positive constant $b_K \geq \overline{C}_K$. Similarly, $C_{f,\tilde{K}} := \log(1 + b_{\tilde{K}}/(4L_{\tilde{K}}))/(L_{\tilde{K}} b_{\tilde{K}} f_{\mathbf{z},max} \int \tilde{K}^2)$, $b_{\tilde{K}} \geq \overline{C}_{\tilde{K}}$.

5.3.2 Asymptotic behavior

The previous exponential inequalities are not optimal to prove asymptotic results. Indeed, they directly or indirectly rely on upper bounds, as in Hoeffding or Bernstein-type inequalities. In the case of kernel estimates, this implies the necessary condition $nh^{2p} \to \infty$, at least. By a direct approach, it is possible to state the consistency of $\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}}$, $k = 1, 2, 3$, and then of $\tilde{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}$, under the weaker condition $nh^p \to \infty$.

Proposition 5.7 (Consistency)**.** *Under Assumption 5.3.1, if $nh_n^p \to \infty$, $\lim K(\mathbf{t})|\mathbf{t}|^p = 0$ when $|\mathbf{t}| \to \infty$, $f_{\mathbf{Z}}$ and $\mathbf{z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$ are continuous on $\mathcal{Z}$, then $\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}}$ tends to $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ in probability, when $n \to \infty$ for any $k = 1, 2, 3$.*

This property is proved in Section 5.5.6. Moreover, Proposition 5.6 does not allow to state the strong uniform consistency of $\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}}$ because the threshold t has to be of order h^p at most. Here again, a direct approach is possible.

Proposition 5.8 (Uniform consistency)**.** *Under Assumption 5.3.1, assume that $nh_n^{2p}/\log n \to \infty$, $\lim K(\mathbf{t})|\mathbf{t}|^p = 0$ when $|\mathbf{t}| \to \infty$, K is Lipschitz, $f_{\mathbf{Z}}$ and $\mathbf{z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$ are continuous on a bounded set $\mathcal{Z}$, and there exists a lower bound $f_{\mathbf{Z},\min}$ s.t. $f_{\mathbf{Z},\min} \leq f_{\mathbf{Z}}(\mathbf{z})$ for any $\mathbf{z} \in \mathcal{Z}$. Then $\sup_{\mathbf{z}\in\mathcal{Z}} \left|\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}}\right| \to 0$ almost surely, when $n \to \infty$ for any $k = 1, 2, 3$.*

This property is proved in Section 5.5.7. To derive the asymptotic law of this estimator, we will assume:

Assumption 5.3.6. *(i) $nh_n^p \to \infty$ and $nh_n^{p+2\alpha} \to 0$; (ii) $K(\cdot)$ is compactly supported.*

Proposition 5.9 (Joint asymptotic normality at different points)**.** *Let $\mathbf{z}'_1, \ldots, \mathbf{z}'_{n'}$ be fixed points in a set $\mathcal{Z} \subset \mathbb{R}^p$. Assume 5.3.1, 5.3.4, 5.3.6, that the $\mathbf{z}'_i$ are distinct and that $f_{\mathbf{Z}}$ and $\mathbf{z} \mapsto f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}, \mathbf{z})$ are continuous on $\mathcal{Z}$, for every $\mathbf{x}$. Then, as $n \to \infty$,*

$$(nh_n^p)^{1/2} \left(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i} - \tau_{1,2|\mathbf{Z}=\mathbf{z}'_i}\right)_{i=1,\ldots,n'} \xrightarrow{D} \mathcal{N}(0, \mathbb{H}^{(k)}),\ k = 1, 2, 3,$$

where $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}$ denotes any of the estimators $\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}}$, $k = 1, 2, 3$ or $\tilde{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}$, and $\mathbb{H}$ is the $n' \times n'$ diagonal real matrix defined by

$$[\mathbb{H}^{(k)}]_{i,j} = \frac{4 \int K^2 \mathbb{1}_{\{i=j\}}}{f_{\mathbf{Z}}(\mathbf{z}'_i)} \left\{ \mathbb{E}[g_k(\mathbf{X}_1, \mathbf{X}) g_k(\mathbf{X}_2, \mathbf{X}) | \mathbf{Z} = \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}'_i] - \tau^2_{1,2|\mathbf{Z}=\mathbf{z}'_i} \right\},$$

for every $1 \leq i, j \leq n'$, and $(\mathbf{X}, \mathbf{Z})$, $(\mathbf{X}_1, \mathbf{Z}_1)$, $(\mathbf{X}_2, \mathbf{Z}_2)$ are independent versions.

This proposition is proved in Section 5.5.8.

Remark 5.10. *The latter results will provide some simple tests of the constancy of the function $\mathbf{z} \mapsto \tau_{1,2|\mathbf{z}}$, and then of the constancy of the associated conditional copula itself. This would test the famous "simplifying assumption" ("$\mathcal{H}_0 : C_{1,2|\mathbf{Z}=\mathbf{z}}$ does not depend on the choice of $\mathbf{z}$"), a key assumption for vine modeling in particular: see [5] or [68] for a discussion, [38] for a review and a presentation of formal tests for this hypothesis.*

5.4 Simulation study

In this simulation study, we draw i.i.d. random samples $(X_{i,1}, X_{i,2}, Z_i)$, $i = 1, \ldots, n$, with univariate explanatory variables ($p = 1$). We consider two settings, that correspond to bounded and/or unbounded explanatory variables respectively:

1. $\mathcal{Z} =]0,1[$ and the law of Z is uniform on $]0,1[$. Conditionally on $Z = z$, $X_1|Z = z$ and $X_2|Z = z$ both follow a Gaussian distribution $\mathcal{N}(z, 1)$. Their associated conditional copula is Gaussian and their conditional Kendall's tau is given by $\tau_{1,2|Z=z} = 2z - 1$.

2. $\mathcal{Z} = \mathbb{R}$ and the law of Z is $\mathcal{N}(0, 1)$. Conditionally on $Z = z$, $X_1|Z = z$ and $X_2|Z = z$ both follow a Gaussian distribution $\mathcal{N}(\Phi(z), 1)$, where $\Phi(\cdot)$ is the cdf of the $\mathbf{Z}$. Their associated conditional copula is Gaussian and their conditional Kendall's tau is given by $\tau_{1,2|Z=z} = 2\Phi(z) - 1$.

These simple frameworks allow us to compare the numerical properties of our different estimators in different parts of the space, in particular when Z is close to zero or one, i.e. when the conditional Kendall's tau is close to -1 or to 1. We compute the different estimators $\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}}$ for $k = 1, 2, 3$, and the symmetrically rescaled version $\tilde{\tau}_{1,2|z}$. The bandwidth h is chosen as proportional to the usual "rule-of-thumb" for kernel density estimation, i.e. $h = \alpha_h \hat{\sigma}(Z) n^{-1/5}$ with $\alpha_h \in \{0.5, 0.75, 1, 1.5, 2\}$ and $n \in \{100, 500, 1000, 2000\}$. For each setting, we consider three local measures of goodness-of-fit: for a given z and for any Kendall's tau estimate (say $\hat{\tau}_{1,2|Z=z}$), let

- the (local) bias: $Bias(z) := \mathbb{E}[\hat{\tau}_{1,2|Z=z}] - \tau_{1,2|Z=z}$,
- the (local) standard deviation: $Sd(z) := \mathbb{E}\Big[\big(\hat{\tau}_{1,2|Z=z} - \mathbb{E}[\hat{\tau}_{1,2|Z=z}]\big)^2\Big]^{1/2}$,
- the (local) mean square-error: $MSE(z) := \mathbb{E}\Big[\big(\hat{\tau}_{1,2|Z=z} - \tau_{1,2|Z=z}\big)^2\Big]$.

We also consider their integrated version w.r.t the usual Lebesgue measure on the whole support of z, respectively denoted by $IBias$, ISd and $IMSE$. Some results concerning these integrated measures are given in Table 5.1 (resp. Table 5.2) for Setting 1 (resp. Setting 2), and for different choices of α_h and n. For the sake of effective calculations of these measures, all the theoretical previous expectations are replaced by their empirical counterparts based on 500 simulations.

For every n, the best results seem to be obtained with $\alpha_h = 1.5$ and the fourth (rescaled) estimator, particularly in terms of bias. This is not so surprising, because the estimators $\hat{\tau}^{(k)}$, $k = 1, 2, 3$, do not have the right support at a finite distance. Note that this comparative advantage of $\tilde{\tau}$ in terms of bias decreases with n, as expected. In terms of integrated variance, all the considered estimators behave more or less similarly, particularly when $n \geq 500$.

To illustrate our results for Setting 1 (resp. Setting 2), the functions $z \mapsto Bias(z)$, $Sd(z)$ and $MSE(z)$ have been plotted on Figures 5.1-5.2 (resp. Figures 5.3-5.4), both with our empirically optimal choice $\alpha_h = 1.5$. We can note that, considering the bias, the estimator $\tilde{\tau}$ behaves similarly as $\hat{\tau}^{(1)}$ when the true τ is close to -1, and similarly as $\hat{\tau}^{(3)}$ when the true Kendall's tau is close to 1. But globally, the best pointwise estimator is clearly obtained with the rescaled version $\tilde{\tau}_{1,2|\mathbf{Z}=\cdot}$, after a quick inspection of MSE levels, and even if the differences between our four estimators weaken for large sample sizes. The comparative advantage of $\tilde{\tau}_{1,2|\mathbf{z}}$ more clearly appears with Setting 2 than with Setting 1. Indeed, in the former case, the support of $\mathbf{Z}$'s distribution is the whole line. Then $\hat{f}_{\mathbf{Z}}$ does not suffer any more from the boundary bias phenomenon, contrary to what happened with Setting 1. As a consequence, the biases induced by the definitions of $\hat{\tau}^{(k)}_{1,2|\mathbf{z}}$, $k = 1, 3$, appear more strinkingly in Figure 5.3, for instance: when z

is close to (-1) (resp. 1), the biases of $\hat{\tau}^{(1)}_{1,2|z}$ (resp. $\hat{\tau}^{(3)}_{1,2|z}$) and $\tilde{\tau}_{1,2|z}$ are close, when the bias $\hat{\tau}^{(3)}_{1,2|z}$ (resp. $\hat{\tau}^{(1)}_{1,2|z}$) is a lot larger. Since the squared biases are here significantly larger than the variances in the tails, $\tilde{\tau}_{1,2|z}$ provides the best estimator globally considering "both sides" together. But even in the center of $\mathbf{Z}$'s distribution, the latter estimator behaves very well.

In Setting 2 where there is no boundary problem, we also try to estimate the conditional Kendall's tau using our cross-validation criterion (5.4), with $N_{pairs} = 1000$. More precisely, denoting by h^{CV} the minimizer of the cross-validation criterion, we try different choices $h = \alpha_h \times h^{CV}$ with $\alpha_h \in \{0.5, 0.75, 1, 1.5, 2\}$. The results in terms of integrated bias, standard deviation and MSE are given in Table 5.3. We do not find any substantial improvements compared to the previous Table 5.2, where the bandwidth was chosen "roughly". In Table 5.4, we compare the average h^{CV} with the previous choice of h. The expectation of h^{CV} is always higher than the "rule-of-thumb" h^{ref}, but the difference between both decreases when the sample size n increases. The standard deviation of h^{CV} is quite high for low values of n, but decreases as a function of n. This may be seen as quite surprising given the fact that the number of pairs N_{pairs} used in the computation of the criterion stays constant. Nevertheless, when the sample size increases, the selected pairs are better in the sense that the differences $|Z_i - Z_j|$ can become smaller as more replications of Z_i are available.

5.5 Proofs

For convenience, we recall Berk's (1970) inequality (see Theorem A in Serfling [123, p.201]). Note that, if $m = 1$, this reduces to Bernstein's inequality.

Lemma 5.11. *Let $m, n > 0$, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ i.i.d. random vectors with values in a measurable space $\mathcal{X}$ and $g : \mathcal{X}^m \to [a, b]$ be a symmetric real bounded function. Set $\theta := \mathbb{E}[g(\mathbf{X}_1, \ldots, \mathbf{X}_m)]$ and $\sigma^2 := Var[g(\mathbf{X}_1, \ldots, \mathbf{X}_m)]$. Then, for any $t > 0$ and $n \geq m$,*

$$\mathbb{P}\left(\binom{n}{m}^{-1} \sum_c g(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_m}) - \theta \geq t\right) \leq \exp\left(-\frac{[n/m]t^2}{2\sigma^2 + (2/3)(b-\theta)t}\right),$$

where $\sum_c$ denotes summation over all subgroups of m distinct integers $(i_1, \ldots, i_m)$ of $\{1, \ldots n\}$.

5.5.1 Proof of Proposition 5.1

Since there are no ties a.s.,

$$\begin{aligned} 1 + \hat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}} &= 4\sum_{i=1}^n \sum_{j=1}^n w_{i,n}(\mathbf{z}) w_{j,n}(\mathbf{z}) \Big(\mathbb{1}\{X_{i,1} < X_{j,1}\} - \mathbb{1}\{X_{i,1} < X_{j,1}, X_{i,2} > X_{j,2}\} \Big) \\ &= 4\sum_{i=1}^n \sum_{j=1}^n w_{i,n}(\mathbf{z}) w_{j,n}(\mathbf{z}) \mathbb{1}\{X_{i,1} < X_{j,1}\} + \hat{\tau}^{(3)}_{1,2|\mathbf{Z}=\mathbf{z}} - 1. \end{aligned}$$

But

$$\begin{aligned} 1 &= \sum_{i=1}^n \sum_{j=1}^n w_{i,n}(\mathbf{z}) w_{j,n}(\mathbf{z}) = \sum_{i=1}^n \sum_{j=1}^n w_{i,n}(\mathbf{z}) w_{j,n}(\mathbf{z}) \Big(\mathbb{1}\{X_{i,1} \leq X_{j,1}\} + \mathbb{1}\{X_{i,1} > X_{j,1}\} \Big) \\ &= 2\sum_{i=1}^n \sum_{j=1}^n w_{i,n}(\mathbf{z}) w_{j,n}(\mathbf{z}) \mathbb{1}\{X_{i,1} < X_{j,1}\} + \sum_{i=1}^n w_{i,n}^2, \end{aligned}$$

implying that

$$1 + \hat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}} = 4\left(\frac{1 - s_n}{2}\right) + \hat{\tau}^{(3)}_{1,2|\mathbf{Z}=\mathbf{z}} - 1,$$

		$n = 100$			$n = 500$			$n = 1000$			$n = 2000$		
		IBias	ISd	IMSE	IBias	ISd	IMSE	IBias	ISd	IMSE	IBias	ISd	IMSE
$\alpha_h = 0.5$	$\hat{\tau}^{(1)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-133	197	66.5	-34.5	84.9	9.86	-18.2	61.6	4.85	-10.9	46	2.65
	$\hat{\tau}^{(2)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-12.9	187	43.7	-4.08	84.4	8.58	-0.9	61.5	4.49	-1.07	46	2.53
	$\hat{\tau}^{(3)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	107	190	56.6	26.4	84.5	9.26	16.4	61.5	4.76	8.8	46	2.6
	$\tilde{\tau}_{1,2\|\mathbf{Z}=\mathbf{z}}$	**-0.91**	213	48.2	**-1.18**	86.9	8.55	**0.733**	62.4	4.46	**-0.149**	46.4	2.5
$\alpha_h = 0.75$	$\hat{\tau}^{(1)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-88	150	35.8	-26.3	68	6.32	-13.9	50.7	3.33	-7.98	37.6	1.8
	$\hat{\tau}^{(2)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-10.4	145	26.3	-5.97	67.9	5.6	-2.33	50.6	3.12	-1.39	37.5	1.74
	$\hat{\tau}^{(3)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	67.2	146	30.6	14.3	67.9	5.75	9.2	50.6	3.19	5.2	37.5	1.76
	$\tilde{\tau}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-2.06	157	26.7	-3.99	69.2	5.49	-1.21	51.2	3.05	-0.76	37.8	1.69
$\alpha_h = 1$	$\hat{\tau}^{(1)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-67.8	123	24.5	-19.2	58.7	4.8	-11	43.1	2.52	-6.34	33	1.44
	$\hat{\tau}^{(2)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-9.99	121	19	-3.95	58.6	4.39	-2.35	43.1	2.39	-1.39	33	1.4
	$\hat{\tau}^{(3)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	47.8	122	20.9	11.3	58.7	4.47	6.34	43.1	2.41	3.57	33	1.41
	$\tilde{\tau}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-3.48	128	18.1	-2.34	59.5	4.18	-1.46	43.4	2.29	-0.897	33.2	1.35
$\alpha_h = 1.5$	$\hat{\tau}^{(1)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-44.6	101	17.5	-15.9	50.4	4.12	-9.7	35.9	2.13	-5.52	27.6	1.28
	$\hat{\tau}^{(2)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-5.81	100	14.9	-5.68	50.3	3.84	-3.84	35.9	2.02	-2.18	27.6	1.24
	$\hat{\tau}^{(3)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	33	101	15.5	4.58	50.3	3.77	2.01	35.9	1.99	1.15	27.6	1.23
	$\tilde{\tau}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-1.09	104	**13.4**	-4.55	50.8	**3.57**	-3.19	36.1	**1.9**	-1.83	27.7	**1.18**
$\alpha_h = 2$	$\hat{\tau}^{(1)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-37.8	91.4	17.3	-11.8	43.8	4.14	-7.2	31.2	2.35	-5.97	23.7	1.43
	$\hat{\tau}^{(2)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-8.03	91.4	15.4	-3.93	43.8	3.94	-2.75	31.2	2.28	-3.44	23.7	1.39
	$\hat{\tau}^{(3)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	21.7	91.7	15.4	3.91	43.8	3.87	1.7	31.2	2.24	-0.912	23.7	1.37
	$\tilde{\tau}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-4.5	94.2	13.5	-3.01	44.1	3.62	-2.24	31.3	2.12	-3.16	23.8	1.32

Table 5.1: Results of the simulation in Setting 1. All values have been multiplied by 1000. Bold values indicate optimal choices for the chosen measure of performance.

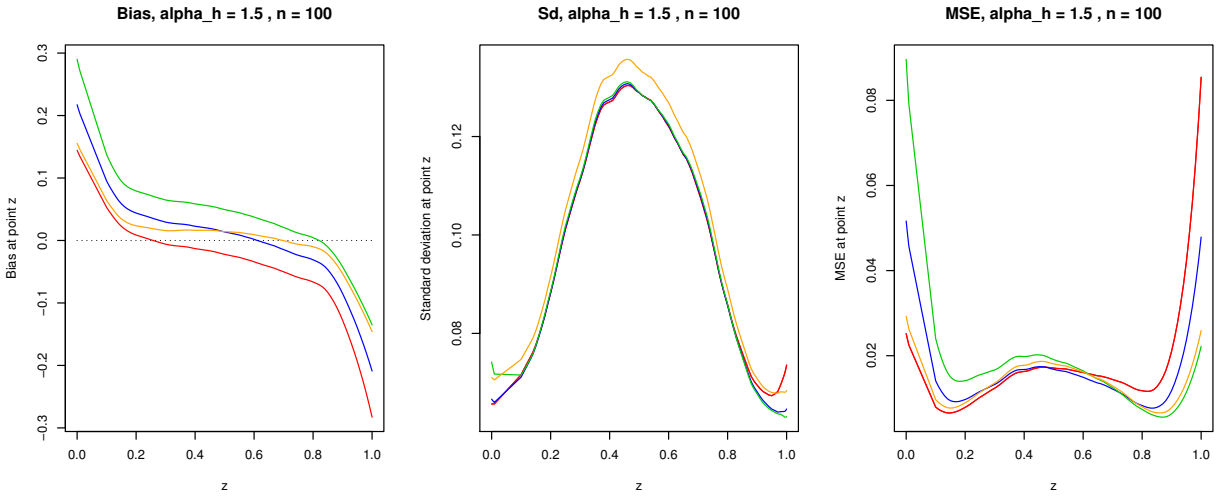


Figure 5.1: Local bias, standard deviation and MSE for the estimators $\hat{\tau}^{(1)}$ (red) , $\hat{\tau}^{(2)}$ (blue), $\hat{\tau}^{(3)}$ (green), $\tilde{\tau}$ (orange), with $n = 100$ and $\alpha_h = 1.5$ in Setting 1. The dotted line on the first figure is the reference at 0.

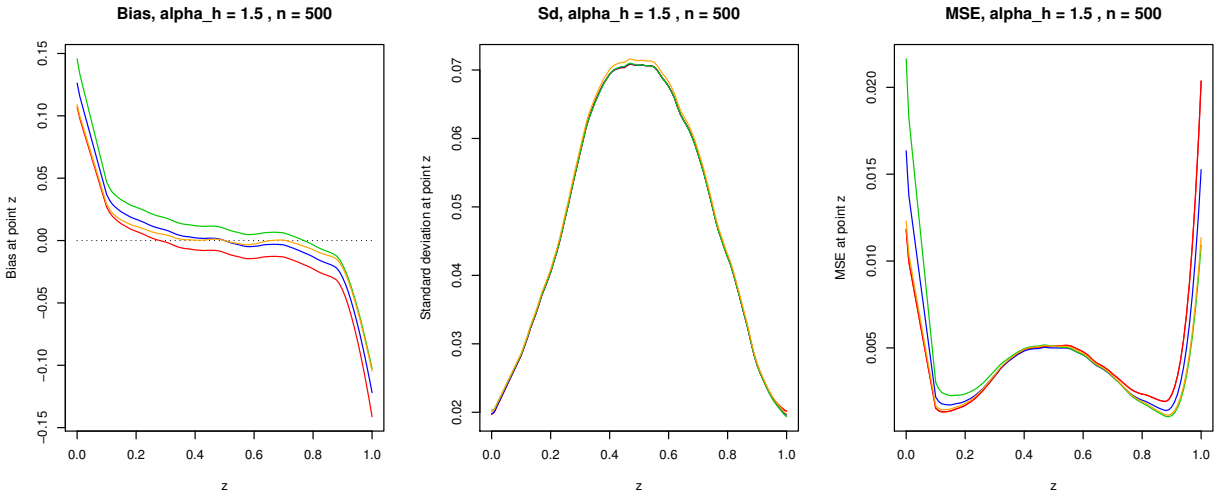


Figure 5.2: Local bias, standard deviation and MSE for the estimators $\hat{\tau}^{(1)}$ (red) , $\hat{\tau}^{(2)}$ (blue), $\hat{\tau}^{(3)}$ (green), $\tilde{\tau}$ (orange), with $n = 500$ and $\alpha_h = 1.5$ in Setting 1. The dotted line on the first figure is the reference at 0.

		$n = 100$			$n = 500$			$n = 1000$			$n = 2000$		
		IBias	ISd	IMSE	IBias	ISd	IMSE	IBias	ISd	IMSE	IBias	ISd	IMSE
$\alpha_h = 0.5$	$\hat{\tau}^{(1)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-207	227	180	-54.1	83.9	16.9	-29.6	55.3	5.81	-16.9	38.9	2.49
	$\hat{\tau}^{(2)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	**1.15**	207	97	**0.845**	80.5	10.8	0.557	54.4	4.35	**0.145**	38.6	2.04
	$\hat{\tau}^{(3)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	210	228	181	55.7	83.2	16.4	30.7	55.4	5.9	17.2	38.9	2.5
	$\hat{\tau}^{(4)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	1.4	225	51.9	0.987	81.4	6.86	**0.456**	55	3.22	0.175	38.9	1.66
$\alpha_h = 0.75$	$\hat{\tau}^{(1)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-144	175	98.6	-33.3	60.6	7.5	-19.8	41.9	3.12	-10.6	30.5	1.42
	$\hat{\tau}^{(2)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-2.33	163	56.2	1.73	59.4	5.56	-0.0619	41.7	2.51	0.665	30.4	1.24
	$\hat{\tau}^{(3)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	140	176	99.2	36.8	60.7	7.73	19.7	42.1	3.12	11.9	30.5	1.45
	$\hat{\tau}^{(4)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-3.15	170	30.3	1.69	60.2	3.85	-0.093	42.1	1.95	0.645	30.5	1.05
$\alpha_h = 1$	$\hat{\tau}^{(1)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-99.8	143	57.7	-24.9	50.9	5.06	-13.5	36.6	2.28	-6.92	26.6	1.09
	$\hat{\tau}^{(2)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	1.17	132	34.6	0.903	50.4	4.02	1.16	36.5	1.97	1.46	26.6	0.994
	$\hat{\tau}^{(3)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	102	139	54.4	26.7	51	5.13	15.8	36.6	2.33	9.83	26.6	1.11
	$\hat{\tau}^{(4)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	2.51	138	20.1	0.897	50.9	2.89	1.16	36.7	1.56	1.48	26.7	0.847
$\alpha_h = 1.5$	$\hat{\tau}^{(1)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-59.1	104	28.1	-14.7	42.3	3.87	-7.56	29.7	1.86	-4.17	21.8	0.932
	$\hat{\tau}^{(2)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	4.34	99.7	21.4	2.05	42.1	3.48	2.07	29.6	1.75	1.35	21.8	0.899
	$\hat{\tau}^{(3)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	67.8	103	29.6	18.8	42.3	3.96	11.7	29.6	1.92	6.87	21.8	0.957
	$\hat{\tau}^{(4)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	3.34	103	**13.4**	2.08	42.5	**2.6**	2.08	29.7	**1.39**	1.35	21.8	**0.755**
$\alpha_h = 2$	$\hat{\tau}^{(1)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-37.2	88.2	23.9	-9.57	38.2	4.6	-3.75	26.2	2.34	-1.09	19.8	1.32
	$\hat{\tau}^{(2)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	8.17	85.9	21.2	2.69	38	4.45	3.32	26.1	2.3	2.99	19.8	1.32
	$\hat{\tau}^{(3)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	53.5	87.4	25.3	14.9	38.1	4.74	10.4	26.2	2.41	7.08	19.8	1.36
	$\hat{\tau}^{(4)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	8.47	88.5	15	2.69	38.4	3.59	3.33	26.3	1.93	3	19.9	1.15

Table 5.2: Results of the simulation in Setting 2. All values have been multiplied by 1000. Bold values indicate optimal choices for the chosen measure of performance.

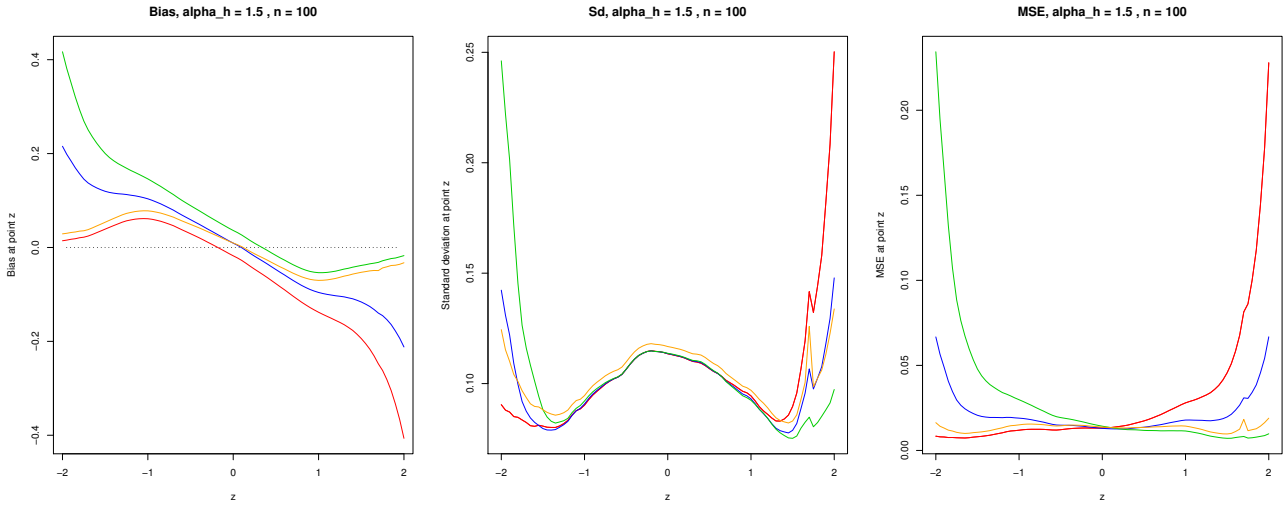


Figure 5.3: Local bias, standard deviation and MSE for the estimators $\hat{\tau}^{(1)}$ (red) , $\hat{\tau}^{(2)}$ (blue), $\hat{\tau}^{(3)}$ (green), $\tilde{\tau}$ (orange), with $n = 100$ and $\alpha_h = 1.5$ in Setting 2. The dotted line on the first figure is the reference at 0.

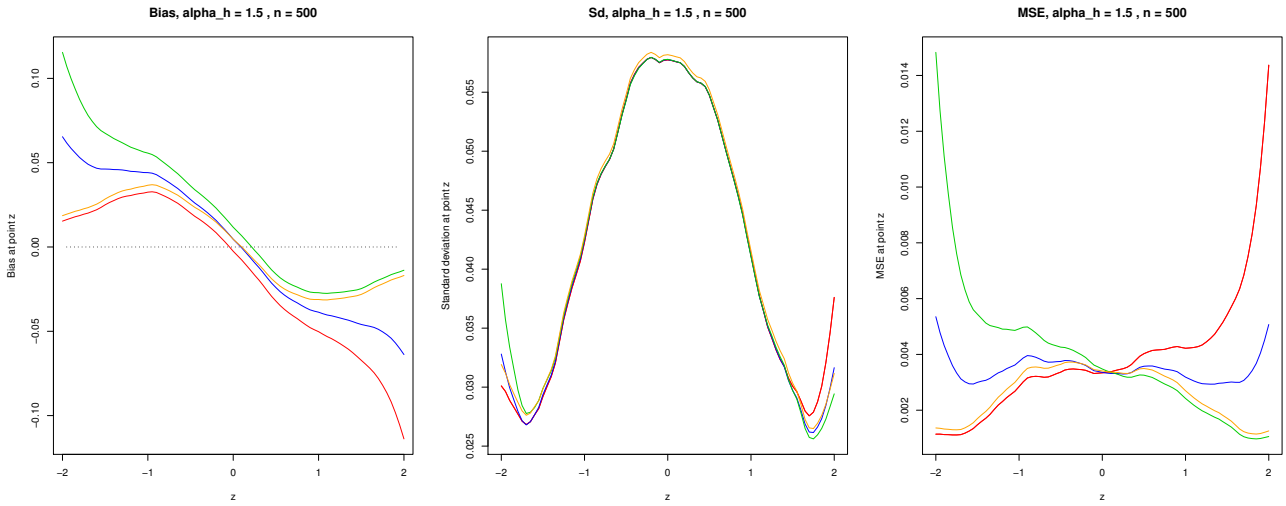


Figure 5.4: Local bias, standard deviation and MSE for the estimators $\hat{\tau}^{(1)}$ (red) , $\hat{\tau}^{(2)}$ (blue), $\hat{\tau}^{(3)}$ (green), $\tilde{\tau}$ (orange), with $n = 500$ and $\alpha_h = 1.5$ in Setting 2. The dotted line on the first figure is the reference at 0.

		$n = 100$			$n = 500$			$n = 1000$			$n = 2000$		
		IBias	ISd	IMSE	IBias	ISd	IMSE	IBias	ISd	IMSE	IBias	ISd	IMSE
$\alpha_h = 0.5$	$\hat{\tau}^{(1)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-111	154	66.2	-36.9	66.8	9.01	-22.4	48.2	4.06	-12.9	36.1	2.04
	$\hat{\tau}^{(2)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	**0.0488**	137	36.3	**0.236**	64.2	6.45	**0.546**	46.8	3.14	**1.29**	35.7	1.78
	$\hat{\tau}^{(3)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	111	151	60.6	37.4	66.3	8.88	23.5	47.2	4.07	15.5	36.2	2.18
	$\hat{\tau}^{(4)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	1.38	132	18.3	0.27	64.5	4.49	0.61	46.8	2.36	1.29	35.6	1.49
$\alpha_h = 0.75$	$\hat{\tau}^{(1)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-67.4	117	35.7	-23.3	52.1	5.27	-13.9	37.8	2.4	-7.6	29	1.3
	$\hat{\tau}^{(2)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	4.32	108	23.5	0.809	50.7	4.21	1.03	37.2	2.07	1.78	28.8	1.21
	$\hat{\tau}^{(3)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	76.1	119	35.4	24.9	51.6	5.12	16	37.6	2.49	11.2	29.1	1.39
	$\hat{\tau}^{(4)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	4.98	106	**13.3**	0.86	51.6	3.13	1.03	37.5	1.63	1.81	28.9	1.02
$\alpha_h = 1$	$\hat{\tau}^{(1)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-43	101	28	-15.8	45.7	4.44	-9.51	33.1	2.04	-4.68	25.1	1.07
	$\hat{\tau}^{(2)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	7.87	93.1	22.4	2.01	44.8	3.91	1.57	32.7	1.87	2.29	24.9	1.03
	$\hat{\tau}^{(3)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	58.8	97.6	27.2	19.8	45.3	4.41	12.7	32.9	2.1	9.27	25.1	1.14
	$\hat{\tau}^{(4)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	8.51	98	15.7	2.05	46	**3.01**	1.57	33.1	**1.5**	2.33	25.1	**0.871**
$\alpha_h = 1.5$	$\hat{\tau}^{(1)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	-16.1	95.6	41.7	-6.36	43	6.35	-4.04	30.6	2.87	-1.11	22.1	1.34
	$\hat{\tau}^{(2)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	14.9	**92.6**	40.4	5.08	42.6	6.2	3.17	30.4	2.83	3.47	22	1.34
	$\hat{\tau}^{(3)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	46	92.8	42.2	16.5	42.6	6.45	10.4	30.4	2.94	8.06	22.1	1.4
	$\hat{\tau}^{(4)}_{1,2\|\mathbf{Z}=\mathbf{z}}$	15.6	100	35.2	5.11	44	5.31	3.17	31	2.45	3.5	22.4	1.17

Table 5.3: Results of the simulation in Setting 2 using $h = \alpha_h \times h^{CV}$ where h^{CV} has been chosen by cross-validation. All values have been multiplied by 1000. Bold values indicate optimal choices for the chosen measure of performance.

n	100	500	1000	2000
$\mathbb{E}[h^{CV}]$	0.77	0.43	0.34	0.27
$Sd[h^{CV}]$	0.17	0.091	0.060	0.057
$h^{ref} = n^{-1/5}$	0.40	0.29	0.25	0.22

Table 5.4: Expectation and standard deviation of the bandwidth selected by cross-validation as a function of the sample size n, and comparison with bandwidth h^{ref} chosen by the rule-of-thumb.

and then $\hat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}} = \hat{\tau}^{(3)}_{1,2|\mathbf{Z}=\mathbf{z}} - 2s_n$. Moreover,

$$\begin{aligned}\hat{\tau}^{(2)}_{1,2|\mathbf{Z}=\mathbf{z}} &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,n}(\mathbf{z})w_{j,n}(\mathbf{z})\Big(\mathbb{1}\big\{X_{i,1}>X_{j,1}, X_{i,2}>X_{j,2}\big\} + \mathbb{1}\big\{X_{i,1}<X_{j,1}, X_{i,2}<X_{j,2}\big\}\\ &\quad - \mathbb{1}\big\{X_{i,1}>X_{j,1}, X_{i,2}<X_{j,2}\big\} - \mathbb{1}\big\{X_{i,1}<X_{j,1}, X_{i,2}>X_{j,2}\big\}\\ &= 2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,n}(\mathbf{z})w_{j,n}(\mathbf{z})\Big(\mathbb{1}\big\{X_{i,1}>X_{j,1}, X_{i,2}>X_{j,2}\big\} - \mathbb{1}\big\{X_{i,1}>X_{j,1}, X_{i,2}<X_{j,2}\big\}\Big)\\ &= \frac{1}{2}\big(\hat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}}+1\big) + \frac{1}{2}\big(\hat{\tau}^{(3)}_{1,2|\mathbf{Z}=\mathbf{z}}-1\big) = \frac{\hat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}} + \hat{\tau}^{(3)}_{1,2|\mathbf{Z}=\mathbf{z}}}{2} = \hat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}} + s_n = \hat{\tau}^{(3)}_{1,2|\mathbf{Z}=\mathbf{z}} - s_n. \ \square\end{aligned}$$

5.5.2 Proof of Proposition 5.2

Lemma 5.12. *Under Assumptions 5.3.1, 5.3.2 and 5.3.3, we have for any* $t>0$*,*

$$\mathbb{P}\bigg(\big|\hat{f}_{\mathbf{Z}}(\mathbf{z}) - f_{\mathbf{Z}}(\mathbf{z})\big| \geq \frac{C_{K,\alpha}h^{\alpha}}{\alpha!} + t\bigg) \leq 2\exp\bigg(-\frac{nh^p t^2}{2f_{\mathbf{Z},max}\int K^2 + (2/3)C_K t}\bigg).$$

This Lemma is proved below. If, for some $\epsilon>0$, we have $C_{K,\alpha}h^{\alpha}/\alpha! + t \leq f_{\mathbf{Z},min} - \epsilon$, then $\hat{f}(\mathbf{z}) \geq \epsilon > 0$ with a probability larger than $1 - 2\exp\big(-nh^p t^2/(2f_{\mathbf{Z},max}\int K^2 + (2/3)C_K t)\big)$. So, we should choose the largest t as possible, which yields Proposition 5.2.

It remains to prove Lemma 5.12. Use the usual decomposition between a stochastic component and a bias: $\hat{f}_{\mathbf{Z}}(\mathbf{z}) - f_{\mathbf{Z}}(\mathbf{z}) = \big(\hat{f}_{\mathbf{Z}}(\mathbf{z}) - \mathbb{E}[\hat{f}_{\mathbf{Z}}(\mathbf{z})]\big) + \big(\mathbb{E}[\hat{f}_{\mathbf{Z}}(\mathbf{z})] - f_{\mathbf{Z}}(\mathbf{z})\big)$. We first bound the bias from above.

$$\mathbb{E}[\hat{f}_{\mathbf{Z}}(\mathbf{z})] - f_{\mathbf{Z}}(\mathbf{z}) = \int_{\mathbb{R}^p} K(\mathbf{u})\Big(f_{\mathbf{Z}}\big(\mathbf{z}+h\mathbf{u}\big) - f_{\mathbf{Z}}(\mathbf{z})\Big)d\mathbf{u}.$$

Set $\phi_{\mathbf{z},\mathbf{u}}(t) := f_{\mathbf{Z}}\big(\mathbf{z}+th\mathbf{u}\big)$ for $t\in[0,1]$. This function has at least the same regularity as $f_{\mathbf{Z}}$, so it is α-differentiable. By a Taylor-Lagrange expansion, we get

$$\int_{\mathbb{R}^p} K(\mathbf{u})\Big(f_{\mathbf{Z}}\big(\mathbf{z}+h\mathbf{u}\big) - f_{\mathbf{Z}}(\mathbf{z})\Big)d\mathbf{u} = \int_{\mathbb{R}^p} K(\mathbf{u})\bigg(\sum_{i=1}^{\alpha-1}\frac{1}{i!}\phi^{(i)}_{\mathbf{z},\mathbf{u}}(0) + \frac{1}{\alpha!}\phi^{(\alpha)}_{\mathbf{z},\mathbf{u}}(t_{\mathbf{z},\mathbf{u}})\bigg)d\mathbf{u},$$

for some real number $t_{\mathbf{z},\mathbf{u}}\in(0,1)$. By Assumption 5.3.1 and for every $i<\alpha$, $\int_{\mathbb{R}^p} K(\mathbf{u})\phi^{(i)}_{\mathbf{z},\mathbf{u}}(0)\,d\mathbf{u} = 0$. Therefore,

$$\begin{aligned}\Big|\mathbb{E}[\hat{f}_{\mathbf{Z}}(\mathbf{z})] - f_{\mathbf{Z}}(\mathbf{z})\Big| &= \Big|\int_{\mathbb{R}^p} K(\mathbf{u})\frac{1}{\alpha!}\phi^{(\alpha)}_{\mathbf{z},\mathbf{u}}(t_{\mathbf{z},\mathbf{u}})d\mathbf{u}\Big|\\ &= \frac{1}{\alpha!}\Big|\int_{\mathbb{R}^p} K(\mathbf{u})\sum_{i_1,\ldots,i_\alpha=1}^{p} h^{\alpha}u_{i_1}\ldots u_{i_\alpha}\frac{\partial^{\alpha}f_{\mathbf{Z}}}{\partial z_{i_1}\ldots\partial z_{i_\alpha}}\big(\mathbf{z}+t_{\mathbf{z},\mathbf{u}}h\mathbf{u}\big)d\mathbf{u}\Big| \leq \frac{C_{K,\alpha}}{\alpha!}h^{\alpha}.\end{aligned}$$

Second, the stochastic component may be written as

$$\hat{f}_{\mathbf{Z}}(\mathbf{z}) - \mathbb{E}[\hat{f}_{\mathbf{Z}}(\mathbf{z})] = n^{-1}\sum_{i=1}^{n} K_h(\mathbf{Z}_i-\mathbf{z}) - \mathbb{E}\Big[n^{-1}\sum_{i=1}^{n}K_h(\mathbf{Z}_i-\mathbf{z})\Big] = n^{-1}\sum_{i=1}^{n}\big(g_{\mathbf{z}}(\mathbf{Z}_i) - \mathbb{E}[g(\mathbf{Z}_i)]\big),$$

where $g(\mathbf{Z}_i) := K_h(\mathbf{Z}_i-\mathbf{z})$. Apply Lemma 5.11 with $m=1$ and the latter $g(\mathbf{Z}_i)$. Here, we have $b=-a=h^{-p}C_K$, $\theta=\mathbb{E}[g(\mathbf{Z}_1)]\geq 0$ and $\big|Var[g(\mathbf{Z}_1)]\big| \leq h^{-p}f_{\mathbf{Z},max}\int K^2$, and we get

$$\mathbb{P}\left(\Big|\frac{1}{n}\sum_{i=1}^{n}K_h(\mathbf{Z}_i-\mathbf{z}) - \mathbb{E}[K_h(\mathbf{Z}_i-\mathbf{z})]\Big| \geq t\right) \leq 2\exp\bigg(-\frac{nt^2}{2h^{-p}f_{\mathbf{Z},max}\int K^2 + (2/3)h^{-p}C_K t}\bigg). \ \square$$

5.5.3 Proof of Proposition 5.3

We show the result for $k = 1$. The two other cases can be proven in the same way.

Consider the decomposition

$$\begin{aligned}\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}} &= 4\sum_{1\le i,j\le n} w_{i,n}(\mathbf{z})w_{j,n}(\mathbf{z})\mathbb{1}\{\mathbf{X}_i < \mathbf{X}_j\} - 4\mathbb{P}\big(\mathbf{X}_1 < \mathbf{X}_2 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) \\ &= \frac{4}{n^2\hat{f}^2_{\mathbf{Z}}(\mathbf{z})}\sum_{1\le i,j\le n} K_h(\mathbf{Z}_i - \mathbf{z})K_h(\mathbf{Z}_j - \mathbf{z})\Big(\mathbb{1}\{\mathbf{X}_i < \mathbf{X}_j\} - \mathbb{P}\big(\mathbf{X}_1 < \mathbf{X}_2 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big)\Big) \\ &=: \frac{4}{\hat{f}^2_{\mathbf{Z}}(\mathbf{z})}\sum_{1\le i,j\le n} S_{i,j}(\mathbf{z}).\end{aligned}$$

Therefore, for any positive numbers x and $\lambda(\mathbf{z})$, we have

$$\begin{aligned}\mathbb{P}(|\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}}| > x) &\le \mathbb{P}\Big(\frac{1}{\hat{f}^2_{\mathbf{Z}}(\mathbf{z})} > \frac{1+\lambda(\mathbf{z})}{f^2_{\mathbf{Z}}(\mathbf{z})}\Big) + \mathbb{P}\Big(\frac{4(1+\lambda(\mathbf{z}))}{f^2_{\mathbf{Z}}(\mathbf{z})} \times |\sum_{1\le i,j\le n} S_{i,j}(\mathbf{z})| > x\Big) \\ &\le \mathbb{P}\Big(|\frac{1}{\hat{f}^2_{\mathbf{Z}}(\mathbf{z})} - \frac{1}{f^2_{\mathbf{Z}}(\mathbf{z})}| > \frac{\lambda(\mathbf{z})}{f^2_{\mathbf{Z}}(\mathbf{z})}\Big) + \mathbb{P}\Big(\frac{4(1+\lambda(\mathbf{z}))}{f^2_{\mathbf{Z}}(\mathbf{z})} \times |\sum_{1\le i,j\le n} S_{i,j}(\mathbf{z})| > x\Big).\end{aligned}$$

For any t s.t. $C_{K,\alpha}h^\alpha/\alpha! + t < f_{\mathbf{Z},min}/2$, set

$$\lambda(\mathbf{z}) = \frac{16 f^2_{\mathbf{z}}(\mathbf{z})}{f^3_{\mathbf{Z},min}}\Big(\frac{C_{K,\alpha}h^\alpha}{\alpha!} + t\Big).$$

Then, this yields

$$\begin{aligned}\mathbb{P}\Big(|\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}}| > x\Big) &\le \mathbb{P}\Big(|\frac{1}{\hat{f}^2_{\mathbf{Z}}(\mathbf{z})} - \frac{1}{f^2_{\mathbf{Z}}(\mathbf{z})}| > \frac{16}{f^3_{\mathbf{Z},min}}\big(\frac{C_{K,\alpha}h^\alpha}{\alpha!} + t\big)\Big) \\ &+ \mathbb{P}\Big(|\sum_{1\le i,j\le n} S_{i,j}(\mathbf{z})| > \frac{f^2_{\mathbf{z}}(\mathbf{z})x}{4(1+\lambda(\mathbf{z}))}\Big).\end{aligned}$$

By setting

$$x = \frac{4}{f^2_{\mathbf{z}}(\mathbf{z})}\Big(\frac{C_{\mathbf{XZ},\alpha}h^\alpha}{\alpha!} + \frac{3f_{\mathbf{z}}(\mathbf{z})\int K^2}{2nh^p} + t'\Big)\Big(1 + \frac{16 f^2_{\mathbf{Z}}(\mathbf{z})}{f^3_{\mathbf{Z},min}}\Big(\frac{C_{K,\alpha}h^\alpha}{\alpha!} + t\Big)\Big),$$

and applying the next two lemmas 5.13 and 5.14, we get the result. □

Lemma 5.13. *Under Assumptions 5.3.1-5.3.3 and if* $C_{K,\alpha}h^\alpha/\alpha! + t < f_{\mathbf{Z},min}/2$ *for some* $t > 0$*,*

$$\mathbb{P}\Big(|\frac{1}{\hat{f}^2_{\mathbf{Z}}(\mathbf{z})} - \frac{1}{f^2_{\mathbf{Z}}(\mathbf{z})}| > \frac{16}{f^3_{\mathbf{Z},min}}\Big(\frac{C_{K,\alpha}h^\alpha}{\alpha!} + t\Big)\Big) \le 2\exp\Big(-\frac{nh^p t^2}{2f_{\mathbf{Z},max}\int K^2 + (2/3)C_K t}\Big),$$

and $\hat{f}_{\mathbf{Z}}(\mathbf{z})$ *is strictly positive on these events.*

Proof : Applying the mean value inequality to the function $x \mapsto 1/x^2$, we get the inequality $\Big|1/\hat{f}^2_{\mathbf{Z}}(\mathbf{z}) - 1/f^2_{\mathbf{Z}}(\mathbf{z})\Big| \le 2\big|\hat{f}_{\mathbf{Z}}(\mathbf{z}) - f_{\mathbf{Z}}(\mathbf{z})\big|/f^{*3}_{\mathbf{Z}}$, where $f^*_{\mathbf{Z}}$ lies between $\hat{f}_{\mathbf{Z}}(\mathbf{z})$ and $f_{\mathbf{Z}}(\mathbf{z})$. Denote by $\mathcal{E}$ the event $\mathcal{E} := \big\{|\hat{f}_{\mathbf{Z}}(\mathbf{z}) - f_{\mathbf{Z}}(\mathbf{z})| \le C_{K,\alpha}h^\alpha/\alpha! + t\big\}$. By Lemma 5.12, we obtain

$$\mathbb{P}(\mathcal{E}) \ge 1 - 2\exp\Big(-\frac{nh^p t^2}{2f_{\mathbf{Z},max}\int K^2 + (2/3)C_K t}\Big). \tag{5.5}$$

Therefore, on this event $\mathcal{E}$, $\big|\hat{f}_{\mathbf{Z}}(\mathbf{z}) - f_{\mathbf{Z}}(\mathbf{z})\big| \le f_{\mathbf{Z},min}/2$, so that $f_{\mathbf{Z},min}/2 \le \hat{f}_{\mathbf{Z}}(\mathbf{z})$. We have also $f_{\mathbf{Z},min}/2 \le f_{\mathbf{Z}}(\mathbf{z})$ and then $f_{\mathbf{Z},min}/2 \le f^*_{\mathbf{Z}}$. Combining the previous inequalities, we finally get

$$\Big|\frac{1}{\hat{f}^2_{\mathbf{Z}}(\mathbf{z})} - \frac{1}{f^2_{\mathbf{Z}}(\mathbf{z})}\Big| \le \frac{16}{f^3_{\mathbf{Z},min}}\big|\hat{f}_{\mathbf{Z}}(\mathbf{z}) - f_{\mathbf{Z}}(\mathbf{z})\big| \le \frac{16}{f^3_{\mathbf{Z},min}}\Big(\frac{C_{K,\alpha}h^\alpha}{\alpha!} + t\Big),$$

on $\mathcal{E}$. But since

$$\mathbb{P}\Big(|\frac{1}{\hat{f}^2_{\mathbf{Z}}(\mathbf{z})}-\frac{1}{f^2_{\mathbf{Z}}(\mathbf{z})}|>\frac{16}{f^3_{\mathbf{Z},min}}\Big(\frac{C_{K,\alpha}h^\alpha}{\alpha!}+t\Big)\Big)\leq\mathbb{P}(\mathcal{E}^c),$$

we deduce the result. □

Lemma 5.14. *Under Assumptions 5.3.1-5.3.4, if $C_{\tilde{K},2}h^2<f_{\mathbf{z}}(\mathbf{z})$, we have for any $t>0$*

$$\mathbb{P}\Big(\Big|\sum_{1\leq i,j\leq n}S_{i,j}(\mathbf{z})\Big|>\frac{C_{\mathbf{XZ},\alpha}h^\alpha}{\alpha!}+\frac{3f_{\mathbf{z}}(\mathbf{z})\int K^2}{2nh^p}+t\Big)\leq 2\exp\Big(-\frac{(n-1)h^{2p}t^2}{4f^2_{\mathbf{Z},max}(\int K^2)^2+(8/3)C^2_Kt}\Big)$$
$$+\quad 2\exp\Big(-\frac{nh^p(f_{\mathbf{z}}(\mathbf{z})-C_{\tilde{K},2}h^2)^2}{8f_{\mathbf{Z},max}\int\tilde{K}^2+4C_{\tilde{K}}(f_{\mathbf{z}}(\mathbf{z})-C_{\tilde{K},2}h^2)/3}\Big).$$

Proof : Note that $\sum_{1\leq i,j\leq n}S_{i,j}(\mathbf{z})=\sum_{1\leq i\neq j\leq n}\big(S_{i,j}(\mathbf{z})-\mathbb{E}[S_{i,j}(\mathbf{z})]\big)+n(n-1)\mathbb{E}[S_{1,2}(\mathbf{z})]+\sum_{i=1}^nS_{i,i}(\mathbf{z})$. The "diagonal term" $\sum_{i=1}^nS_{i,i}(\mathbf{z})=-\mathbb{P}\big(\mathbf{X}_1<\mathbf{X}_2\big|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}\big)\sum_{i=1}^nK_h^2(\mathbf{Z}_i-\mathbf{z})/n^2$ is negative and negligible. It will be denoted by $-\Delta_n(\mathbf{z})<0$. Note that $\tilde{K}(\cdot):=K^2(\cdot)/\int K^2$ is a two-order kernel. Then, $\tilde{f}_{\mathbf{z}}(\mathbf{z}):=\sum_{i=1}^n\tilde{K}_h(\mathbf{Z}_i-\mathbf{z})/n$ is a consistent estimator of $f_{\mathbf{Z}}(\mathbf{z})$. Therefore, due to Lemma 5.12 and with obvious notations, we have for every $\varepsilon>0$

$$\mathbb{P}\Big(\big|\tilde{f}_{\mathbf{Z}}(\mathbf{z})-f_{\mathbf{Z}}(\mathbf{z})\big|\geq\frac{C_{\tilde{K},2}h^2}{2}+\varepsilon\Big)\leq 2\exp\Big(-\frac{nh^p\varepsilon^2}{2f_{\mathbf{Z},max}\int\tilde{K}^2+(2/3)C_{\tilde{K}}\varepsilon}\Big).$$

This implies

$$\mathbb{P}\Big(|\frac{\int K^2}{n^2h^p}\sum_{i=1}^n\tilde{K}_h(\mathbf{Z}_i-\mathbf{z})-\frac{f_{\mathbf{Z}}(\mathbf{z})\int K^2}{nh^p}|\geq\Big(\frac{\int K^2}{nh^p}\Big)\Big(\frac{C_{\tilde{K},2}h^2}{2}+\varepsilon\Big)\Big)$$
$$\leq\quad 2\exp\Big(-\frac{nh^p\varepsilon^2}{2f_{\mathbf{Z},max}\int\tilde{K}^2+(2/3)C_{\tilde{K}}\varepsilon}\Big).$$

By choosing ε s.t. $C_{\tilde{K},2}h^2/2+\varepsilon=f_{\mathbf{z}}(\mathbf{z})/2$, Δ_n will be smaller than $3f_{\mathbf{z}}(\mathbf{z})\int K^2/(2nh^p)$ with a probability that is larger than

$$1-2\exp\Big(-\frac{nh^p\varepsilon^2}{2f_{\mathbf{Z},max}\int\tilde{K}^2+(2/3)C_{\tilde{K}}\varepsilon}\Big).\tag{5.6}$$

Now, let us deal with the main term, that is decomposed as a stochastic component and a bias component. First, let us deal with the bias. Simple calculations provide, if $i\neq j$,

$$\mathbb{E}[S_{i,j}(\mathbf{z})]=n^{-2}\mathbb{E}\Big[K_h(\mathbf{Z}_i-\mathbf{z})K_h(\mathbf{Z}_j-\mathbf{z})\Big(\mathbb{1}\big\{\mathbf{X}_i<\mathbf{X}_j\big\}-\mathbb{P}\big(\mathbf{X}_i<\mathbf{X}_j\big|\mathbf{Z}_i=\mathbf{Z}_j=\mathbf{z}\big)\Big)\Big]$$
$$=n^{-2}\int_{\mathbb{R}^{2p+2}}K_h(\mathbf{z}_1-\mathbf{z})K_h(\mathbf{z}_2-\mathbf{z})\Big(\mathbb{1}\big\{\mathbf{x}_1<\mathbf{x}_2\big\}-\mathbb{P}\big(\mathbf{X}_i<\mathbf{X}_j\big|\mathbf{Z}_i=\mathbf{Z}_j=\mathbf{z}\big)\Big)$$
$$\times f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1,\mathbf{z}_1)\,f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_2,\mathbf{z}_2)\,d\mathbf{x}_1\,d\mathbf{z}_1\,d\mathbf{x}_2\,d\mathbf{z}_2$$
$$=n^{-2}\int_{\mathbb{R}^{2p+2}}K(\mathbf{u})K(\mathbf{v})\Big(\mathbb{1}\big\{\mathbf{x}_1<\mathbf{x}_2\big\}-\mathbb{P}\big(\mathbf{X}_i<\mathbf{X}_j\big|\mathbf{Z}_i=\mathbf{Z}_j=\mathbf{z}\big)\Big)$$
$$\times\Big(f_{\mathbf{X},\mathbf{Z}}\Big(\mathbf{x}_1,\mathbf{z}+h\mathbf{u}\Big)\,f_{\mathbf{X},\mathbf{Z}}\Big(\mathbf{x}_2,\mathbf{z}+h\mathbf{v}\Big)-f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1,\mathbf{z})\,f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_2,\mathbf{z})\Big)\,d\mathbf{x}_1\,d\mathbf{u}\,d\mathbf{x}_2\,d\mathbf{v},$$

because, for every $\mathbf{z}$,

$$0=\int_{\mathbb{R}^4}\Big(\mathbb{1}\big\{\mathbf{x}_1<\mathbf{x}_2\big\}-\mathbb{P}\big(\mathbf{X}_1<\mathbf{X}_2\big|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}\big)\Big)\,f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1,\mathbf{z})f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_2,\mathbf{z})\,d\mathbf{x}_1\,d\mathbf{x}_2.$$

Apply the Taylor-Lagrange formula to the function $\phi_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u},\mathbf{v}}(t):=f_{\mathbf{X},\mathbf{Z}}\big(\mathbf{x}_1,\mathbf{z}+th\mathbf{u}\big)\,f_{\mathbf{X},\mathbf{Z}}\big(\mathbf{x}_2,\mathbf{z}+th\mathbf{v}\big)$. With obvious notation, this yields

$$\mathbb{E}[S_{i,j}(\mathbf{z})]=n^{-2}\int K(\mathbf{u})K(\mathbf{v})\Big(\mathbb{1}\big\{\mathbf{x}_1<\mathbf{x}_2\big\}-\mathbb{P}\big(\mathbf{X}_i<\mathbf{X}_j\big|\mathbf{Z}_i=\mathbf{Z}_j=\mathbf{z}\big)\Big)$$

$$\times \left(\sum_{k=1}^{\alpha-1} \frac{1}{k!} \phi^{(k)}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u},\mathbf{v}}(0) + \frac{1}{\alpha!} \phi^{(\alpha)}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u},\mathbf{v}}(t_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u},\mathbf{v}}) \right) d\mathbf{x}_1\, d\mathbf{u}\, d\mathbf{x}_2\, d\mathbf{v}$$

$$= \int \frac{K(\mathbf{u})K(\mathbf{v})}{n^2 \alpha!} \Big(\mathbb{1}\big\{\mathbf{x}_1 < \mathbf{x}_2\big\} - \mathbb{P}\big(\mathbf{X}_i < \mathbf{X}_j \big| \mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}\big) \Big) \phi^{(\alpha)}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u},\mathbf{v}}(t_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u},\mathbf{v}}) d\mathbf{x}_1\, d\mathbf{u}\, d\mathbf{x}_2\, d\mathbf{v}.$$

Since $\phi^{(\alpha)}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u},\mathbf{v}}(t)$ is equal to

$$\sum_{k=0}^{\alpha} \binom{\alpha}{k} \sum_{i_1,\ldots,i_\alpha=1}^{p} h^\alpha u_{i_1} \ldots u_{i_k} v_{i_{k+1}} \ldots v_{i_\alpha} \frac{\partial^k f_{\mathbf{X},\mathbf{Z}}}{\partial z_{i_1} \ldots \partial z_{i_k}} \Big(\mathbf{x}_1, \mathbf{z} + th\mathbf{u} \Big) \frac{\partial^{\alpha-k} f_{\mathbf{X},\mathbf{Z}}}{\partial z_{i_{k+1}} \ldots \partial z_{i_\alpha}} \Big(\mathbf{x}_2, \mathbf{z} + th\mathbf{v} \Big),$$

using Assumption 5.3.4, we get

$$\big| \mathbb{E}[S_{1,2}(\mathbf{z})] \big| \leq C_{\mathbf{XZ},\alpha} h^\alpha / (n^2 \alpha!). \tag{5.7}$$

Second, the stochastic component will be bounded from above. Indeed,

$$\sum_{1 \leq i \neq j \leq n} (S_{i,j}(\mathbf{z}) - \mathbb{E}[S_{i,j}(\mathbf{z})]) = \frac{1}{n^2} \sum_{1 \leq i \neq j \leq n} g_{\mathbf{z}}\big((\mathbf{X}_i, \mathbf{Z}_i)\,, (\mathbf{X}_j, \mathbf{Z}_j)\big),$$

with the function $g_{\mathbf{z}}$ defined by

$$g_{\mathbf{z}}\big((\mathbf{X}_i, \mathbf{Z}_i), (\mathbf{X}_j, \mathbf{Z}_j)\big) := K_h(\mathbf{Z}_i - \mathbf{z}) K_h(\mathbf{Z}_j - \mathbf{z}) \Big(\mathbb{1}\big\{\mathbf{X}_i < \mathbf{X}_j\big\} - \mathbb{P}\big(\mathbf{X}_i < \mathbf{X}_j \big| \mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}\big) \Big)$$
$$- \mathbb{E}\Big[K_h(\mathbf{Z}_i - \mathbf{z}) K_h(\mathbf{Z}_j - \mathbf{z}) \Big(\mathbb{1}\big\{\mathbf{X}_i < \mathbf{X}_j\big\} - \mathbb{P}\big(\mathbf{X}_i < \mathbf{X}_j \big| \mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}\big) \Big) \Big].$$

The symmetrized version of g is $\tilde{g}_{i,j} = \Big(g_{\mathbf{z}}\big((\mathbf{X}_i, \mathbf{Z}_i)\,, (\mathbf{X}_j, \mathbf{Z}_j)\big) + g_{\mathbf{z}}\big((\mathbf{X}_j, \mathbf{Z}_j)\,, (\mathbf{X}_i, \mathbf{Z}_i)\big) \Big)/2$. We can now apply Lemma 5.11 to the sum of the $\tilde{g}_{i,j}$. With its notation, $\theta = \mathbb{E}\big[\tilde{g}_{i,j}\big] = 0$. Moreover,

$$\begin{aligned}
&\Big| Var\Big[g_{\mathbf{z}}\big((\mathbf{X}_i, \mathbf{Z}_i), (\mathbf{X}_j, \mathbf{Z}_j)\big) \Big] \Big| \\
&\leq \int K_h^2(\mathbf{z}_1 - \mathbf{z}) K_h^2(\mathbf{z}_2 - \mathbf{z}) \Big(\mathbb{1}\big\{\mathbf{x}_1 < \mathbf{x}_2\big\} - \mathbb{P}\big(\mathbf{X}_i < \mathbf{X}_j \big| \mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}\big) \Big)^2 \\
&\times f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1, \mathbf{z}_1) f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_2, \mathbf{z}_2)\, d\mathbf{x}_1\, d\mathbf{x}_2\, d\mathbf{z}_1\, d\mathbf{z}_2 \\
&\leq \int \frac{K^2(\mathbf{t}_1) K^2(\mathbf{t}_2)}{h^{2p}} f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1, \mathbf{z} - h\mathbf{t}_1) f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_2, \mathbf{z} - h\mathbf{t}_2)\, d\mathbf{x}_1\, d\mathbf{x}_2\, d\mathbf{t}_1\, d\mathbf{t}_2 \\
&\leq h^{-2p} f^2_{\mathbf{Z},max} \Big(\int K^2 \Big)^2,
\end{aligned}$$

and the same upper bound applies for $\tilde{g}_{i,j}$ (invoke Cauchy-Schwarz inequality). Here, we choose $b = -a = 2C_K^2 h^{-2p}$. This yields

$$\mathbb{P}\Big(\frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \tilde{g}_{i,j} > t \Big) \leq \exp\left(- \frac{[n/2]t^2}{2h^{-2p} f^2_{\mathbf{Z},max} (\int K^2)^2 + (4/3) C_K^2 h^{-2p} t} \right) \tag{5.8}$$

Then, for every $t > 0$, we obtain

$$\begin{aligned}
&\mathbb{P}\Big(\big| \sum_{1 \leq i \neq j \leq n} \big(S_{i,j}(\mathbf{z}) - \mathbb{E}[S_{i,j}(\mathbf{z})]\big) \big| \geq t \Big) \\
&\leq \mathbb{P}\Big(\frac{1}{n^2} \big| \sum_{1 \leq i \neq j \leq n} g_{\mathbf{z}}\big((\mathbf{X}_i, \mathbf{Z}_i)\,, (\mathbf{X}_j, \mathbf{Z}_j)\big) \big| \geq t \Big) \\
&\leq \mathbb{P}\Big(\frac{(n-1)}{n} \times \frac{2}{n(n-1)} \big| \sum_{1 \leq i < j \leq n} \tilde{g}_{i,j} \big| \geq t \Big) \\
&\leq 2 \exp\left(- \frac{[n/2]t^2}{2h^{-2p} f^2_{\mathbf{Z},max} (\int K^2)^2 + (4/3) C_K^2 h^{-2p} t} \right).
\end{aligned}$$

The latter inequality, (5.6) and (5.7) yield the result. □

5.5.4 Proof of Proposition 5.4

Alternatively, we can apply Theorem 1 in Major [99] instead of the Bernstein-type inequality that has been used in the proof of Proposition 5.3. With the notations of this proof, this will yield the following lemma, that straightforwardly implies the result.

Lemma 5.15. *Under Assumptions 5.3.1-5.3.4 and when* $t \leq 2h^p(\int K^2)^3 f^3_{\mathbf{Z},max}/C_K^4$, $6h^p f_{\mathbf{Z},max}\big(\int |K|\big)^2 < \int K^2$ *and* $C_{\tilde{K},2}h^2 < f_{\mathbf{z}}(\mathbf{z})$*, we have*

$$\begin{aligned}\mathbb{P}\Big(\Big|\sum_{1\leq i,j\leq n} S_{i,j}(\mathbf{z})\Big| > \frac{C_{\mathbf{XZ},\alpha}h^\alpha}{\alpha!} + \frac{3f_{\mathbf{z}}(\mathbf{z})\int K^2}{2nh^p} + t\Big) &\leq C_2\exp\Big(-\frac{\alpha_2 nh^p t}{8f_{\mathbf{Z},max}(\int K^2)}\Big)\\ &+ 2\exp\Big(-\frac{nh^p(f_{\mathbf{z}}(\mathbf{z}) - C_{\tilde{K},2}h^2)^2}{8f_{\mathbf{Z},max}\int \tilde{K}^2 + 4C_{\tilde{K}}(f_{\mathbf{z}}(\mathbf{z}) - C_{\tilde{K},2}h^2)/3}\Big)\\ &+ 2\exp\Big(\frac{nh^p t^2}{32\int K^2(\int |K|)^2 f^3_{\mathbf{Z},max} + 8C_K\int |K| f_{\mathbf{Z},max}t/3}\Big)\end{aligned}$$

for some universal positive constants C_2 *and* α_2*.*

Proof : We lead exactly the same reasoning and the same notations as in Lemma 5.14, until (5.8). Now, with the same notations, introduce $\overline{g}_i := \mathbb{E}[\tilde{g}_{i,j}|\mathbf{X}_i, \mathbf{Z}_i]$ and consider $\xi_{i,j} := \tilde{g}_{i,j} - \overline{g}_i - \overline{g}_j$. Then, $\xi_{i,j}$ is a degenerate (symmetrical) U-statistics because $\mathbb{E}[\xi_{i,j}|\mathbf{X}_i, \mathbf{Z}_i] = \mathbb{E}[\xi_{i,j}|\mathbf{X}_j, \mathbf{Z}_j] = 0$, when $i \neq j$. Actually $\xi_{i,j} =: \xi_{\mathbf{z}}(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{X}_j, \mathbf{Z}_j)$ for some function $\xi_{\mathbf{z}}$ and set

$$\ell_{\mathbf{z}} : (\mathbf{x}_1, \mathbf{z}_1, \mathbf{x}_2, \mathbf{z}_2) \mapsto \frac{h^{2p}}{4C_K^2}\xi_{\mathbf{z}}\big((\mathbf{x}_1, \mathbf{z}_1)\,,\,(\mathbf{x}_2, \mathbf{z}_2)\big), \tag{5.9}$$

for a fixed $\mathbf{z}$ and a fixed h. This yields $\|\ell_{\mathbf{z}}\|_\infty \leq 1$ and, by usual changes of variables, we obtain

$$\begin{aligned}&\int \ell^2_{\mathbf{z}}(\mathbf{x}_1, \mathbf{z}_1, \mathbf{x}_2, \mathbf{z}_2)\, f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1, \mathbf{z}_1) f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_2, \mathbf{z}_2)\, d\mathbf{x}_1\, d\mathbf{x}_2\, d\mathbf{z}_1\, d\mathbf{z}_2\\ &\quad\leq\quad 3h^{2p}\frac{(\int K^2 f_{\mathbf{z},max})^2}{(4C_K^2)^2} + 6h^{3p}\frac{\int K^2 f_{\mathbf{z},max}(\int |K| f_{\mathbf{z},max})^2}{(4C_K^2)^2} \leq \sigma^2, \text{ with}\end{aligned}$$

$$\sigma := h^p C_\sigma,\ C_\sigma := \frac{\int K^2 f_{\mathbf{z},max}}{2C_K^2}, \tag{5.10}$$

because $6h^p\int K^2 f_{\mathbf{z},max}(\int |K| f_{\mathbf{z},max})^2 \leq (\int K^2 f_{\mathbf{z},max})^2$. With the notations of [99], this implies $D = 1$, $m = 1$ and L is arbitrarily small. Therefore, Theorem 2 in [99] yields

$$\mathbb{P}\Big(\frac{1}{2n}|\sum_{i\neq j}\ell_{\mathbf{z}}(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{X}_j, \mathbf{Z}_j)| > x\Big) \leq C_2\exp\Big(-\frac{\alpha_2 x}{\sigma}\Big), \tag{5.11}$$

for some universal constants C_2 and α_2 when $x \leq n\sigma^3$. By setting $t/2 = 4C_K^2 x/(nh^{2p})$ and applying Lemma 5.11, this provides

$$\begin{aligned}\mathbb{P}\Big(|\sum_{1\leq i\neq j\leq n}\big(S_{i,j}(\mathbf{z}) - \mathbb{E}[S_{i,j}(\mathbf{z})]\big)| \geq t\Big) &\leq \mathbb{P}\Big(\frac{1}{n^2}|\sum_{1\leq i\neq j\leq n}\xi_{ij}| \geq t/2\Big) + \mathbb{P}\Big(|\frac{1}{n}\sum_{i=1}^n \overline{g}_i| \geq t/4\Big)\\ &\leq\quad C_2\exp\Big(-\frac{\alpha_2 nth^p}{8f_{\mathbf{Z},max}(\int K^2)}\Big)\\ &+\quad 2\exp\Big(\frac{nh^p t^2}{32\int K^2(\int |K|)^2 f^3_{\mathbf{Z},max} + 8/3C_K\int |K| f_{\mathbf{Z},max}t}\Big)\end{aligned}$$

when $t \leq 2h^p(\int K^2)^3 f^3_{\mathbf{Z},max}/C_K^4$. The latter inequality, (5.6) and (5.7) conclude the proof. □

5.5.5 Proof of Proposition 5.6

For $k=1$, we follow the paths of the proof of Proposition 5.4. Since $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}-\tau_{1,2|\mathbf{Z}=\mathbf{z}} = 4\sum_{1\le i,j\le n} S_{i,j}(\mathbf{z})/\hat{f}^2_{\mathbf{Z}}(\mathbf{z})$, we prove the result if we bound from above $1/\hat{f}^2_{\mathbf{Z}}(\mathbf{z})$ and $\big|\sum_{1\le i,j\le n} S_{i,j}(\mathbf{z})\big|$ uniformly w.r.t. $\mathbf{z}\in\mathcal{Z}$. To be specific, for any positive constant $\mu<1$, if $|\hat{f}_{\mathbf{Z}}(\mathbf{z})-f_{\mathbf{Z}}(\mathbf{z})|\le \mu f_{\mathbf{z},min}$, then $1/\hat{f}^2_{\mathbf{z},max}(\mathbf{z}) \le f^{-2}_{\mathbf{z},min}(1-\mu)^{-2}$. We deduce

$$\begin{aligned}\mathbb{P}(\sup_{\mathbf{z}\in\mathcal{Z}}|\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}-\tau_{1,2|\mathbf{Z}=\mathbf{z}}|>x) &\le \mathbb{P}\big(\|\hat{f}_{\mathbf{Z}}-f_{\mathbf{Z}}\|_\infty > \mu f_{\mathbf{z},min}\big)\\ &+ \mathbb{P}\Big(\frac{4}{f^2_{\mathbf{Z},min}(1-\mu)^2}\sup_{\mathbf{z}\in\mathcal{Z}}\big|\sum_{1\le i,j\le n} S_{i,j}(\mathbf{z})\big|>x\Big).\end{aligned}$$

First invoke the uniform exponential inequality, as stated in [116], Proposition 9: for every $\varepsilon < b_K\int K^2 f_{\mathbf{Z},max}/C_K$,

$$\mathbb{P}\big(\|\hat{f}_{\mathbf{Z}}-f_{\mathbf{Z}}\|_\infty > \varepsilon + \frac{C_{\mathbf{XZ},\alpha}h^\alpha}{\alpha!}\big) \le \mathbb{P}\big(\|\hat{f}_{\mathbf{Z}}-\mathbb{E}[\hat{f}_{\mathbf{Z}}]\|_\infty>\varepsilon\big) \le L_K\exp\big(-C_{f,K}nh^p\varepsilon^2\big), \tag{5.12}$$

for n sufficiently large. Then, apply Lemma 5.16, by setting (x,ε) so that

$$x=\frac{4}{f^2_{\mathbf{z},min}(1-\mu)^2}\Big(\frac{C_{\mathbf{XZ},\alpha}h^\alpha}{\alpha!}+\frac{3f_{\mathbf{z},max}\int K^2}{2nh^p}+t\Big) \text{ and } \varepsilon+\frac{C_{\mathbf{XZ},\alpha}h^\alpha}{\alpha!}=\mu f_{\mathbf{z},min}.\ \square$$

Lemma 5.16. *Under the assumptions of Proposition 5.6, we have*

$$\begin{aligned}&\mathbb{P}\Big(\sup_{\mathbf{z}\in\mathcal{Z}}\Big|\sum_{1\le i,j\le n}S_{i,j}(\mathbf{z})\Big|>\frac{C_{\mathbf{XZ},\alpha}h^\alpha}{\alpha!}+\frac{3f_{\mathbf{z},max}\int K^2}{2nh^p}+t\Big)\\ &\le\ C_2D\exp\Big(-\frac{\alpha_2 nth^p}{8f_{\mathbf{Z},max}(\int K^2)}\Big)+L_{\tilde{K}}\exp\big(-C_{f,\tilde{K}}nh^p(f_{\mathbf{z},max}-\tilde{C}_{\mathbf{XZ},2}h^2)^2/4\big)\\ &+\ 2\exp\big(-\frac{A_2nh^pt^2C_K^{-4}}{16^2A_1^2\int K^2 f^3_{\mathbf{z},max}(\int|K|)^2}\big)+2\exp(-\frac{A_2nh^pt}{16C_K^2A_1}),\end{aligned}$$

when $t\le 2h^p(\int K^2)^3 f^3_{\mathbf{Z},max}/C_K^4$,

$$-16A_1C_K^2\frac{A_{\overline{g}}\int K^2 f^3_{\mathbf{z},max}(\int|K|)^2}{n^{1/2}h^{p/2}}\ln(h^p\int K^2 f^3_{\mathbf{z},max}(\int|K|)^2)<t,\ \textit{and}$$

$$nh^pt\ge(\int K^2)f_{\mathbf{z},max}M_2(p+\beta)^{3/2}\log\Big(\frac{4C_K^2}{h^pf_{\mathbf{z},max}\int K^2}\Big),\ \beta=\max\big(0,\frac{\log D}{\log n}\big),\ D:=\lceil\mathcal{V}(\frac{4C_K\lambda_K}{h})^p\rceil,$$

for some universal constants C_2,α_2,M_2,A_1,A_2 *and a constant* $A_{\overline{g}}$ *that depends on* K *and* $f_{\mathbf{z},max}$.

Proof : We will use the arguments and notations of the proof of Lemmas 5.14 and 5.15. We still invoke the decomposition $\sum_{1\le i,j\le n}S_{i,j}(\mathbf{z})=\sum_{1\le i\ne j\le n}\big(S_{i,j}(\mathbf{z})-\mathbb{E}[S_{i,j}(\mathbf{z})]\big)+n(n-1)\mathbb{E}[S_{1,2}(\mathbf{z})]+\sum_{i=1}^n S_{i,i}(\mathbf{z})$.

First let us find a uniform bound for the "diagonal term" $\Delta_n(\mathbf{z})=\sum_{i=1}^n S_{i,i}(\mathbf{z})=\int K^2\tilde{f}_{\mathbf{z}}(\mathbf{z})/(nh^p)$. As in (5.12), for every $\varepsilon<b_{\tilde{K}}\int\tilde{K}^2 f_{\mathbf{Z},max}/C_{\tilde{K}}$,

$$\mathbb{P}\big(\|\tilde{f}_{\mathbf{Z}}-f_{\mathbf{Z}}\|_\infty>\varepsilon+\frac{\tilde{C}_{\mathbf{XZ},2}h^2}{2}\big)\le L_{\tilde{K}}\exp\big(-C_{f,\tilde{K}}nh^p\varepsilon^2\big),$$

for n sufficiently large. This implies

$$\begin{aligned}&\mathbb{P}\Big(\sup_{\mathbf{z}\in\mathcal{Z}}|\frac{\int K^2}{n^2h^p}\sum_{i=1}^n\tilde{K}_h(\mathbf{Z}_i-\mathbf{z})-\frac{f_{\mathbf{Z}}(\mathbf{z})\int K^2}{nh^p}|\ge\Big(\frac{\int K^2}{nh^p}\Big)\big(\varepsilon+\frac{\tilde{C}_{\mathbf{XZ},2}h^2}{2}\big)\Big)\\ &\le\ L_{\tilde{K}}\exp\big(-C_{f,\tilde{K}}nh^p\varepsilon^2\big).\end{aligned}$$

Choose ε s.t. $\tilde{C}_{\mathbf{XZ},2}h^2/2+\varepsilon = f_{\mathbf{z},max}/2$. Then, $\sup_{\mathbf{z}}|\Delta_n(\mathbf{z})|$ will be smaller than $3f_{\mathbf{z},max}\int K^2/(2nh^p)$ with a probability that is larger than

$$1 - L_{\bar{K}}\exp\big(-C_{f,\bar{K}}nh^p\varepsilon^2\big). \tag{5.13}$$

Moreover, it is easy to see that

$$\sup_{\mathbf{z}\in\mathcal{Z}}\big|\mathbb{E}[S_{1,2}(\mathbf{z})]\big| \leq C_{\mathbf{XZ},\alpha}h^\alpha/(n^2\alpha!). \tag{5.14}$$

With the same notations as in the proof of Lemma 5.15, the stochastic component will be driven by

$$\sum_{1\leq i\neq j\leq n}(S_{i,j}(\mathbf{z})-\mathbb{E}[S_{i,j}(\mathbf{z})]) = \frac{1}{n^2}\sum_{1\leq i\neq j\leq n} g_{\mathbf{z}}\big((\mathbf{X}_i,\mathbf{Z}_i),(\mathbf{X}_j,\mathbf{Z}_j)\big)$$
$$= \frac{1}{n^2}\sum_{1\leq i\neq j\leq n}\tilde{g}_{i,j} = \frac{1}{n^2}\sum_{1\leq i\neq j\leq n}\xi_{i,j} + \frac{2(n-1)}{n^2}\sum_{i=1}^n\overline{g}_i.$$

Now apply Theorem 1 in [99], by recalling (5.9) and considering the family $\mathcal{F} := \Big\{\ell_{\mathbf{z}},\, \mathbf{z}\in\mathcal{Z}\Big\}$, for a fixed bandwidth h. The constant σ has the same value as in (5.10). It is easy to check that the latter class of functions is L^2 dense (see [99]). Set $\varepsilon\in(0,1)$. Since K is λ_K-Lipschitz, every function $\ell_{\mathbf{z}}\in\mathcal{F}$ can be approximated in L^2 by a function $\ell_{\mathbf{z}_j}\in\mathcal{F}$, for some $j\in\{1,\ldots,m\}$ s.t. $\int|\ell_{\mathbf{z}}-\ell_{\mathbf{z}_i}|^2d\nu\leq\varepsilon^2$, for any probability measure ν. Indeed, $\int|\ell_{\mathbf{z}}-\ell_{\mathbf{z}_i}|^2d\nu\leq 64\lambda_K^2\|\mathbf{z}-\mathbf{z}_j\|_\infty^2C_K^2h^{-2}$ that is less than ε^2, if we cover $\mathcal{Z}$ by a grid of m points $(\mathbf{z}_j)$ in $\mathcal{Z}$ s.t. $\|\mathbf{z}-\mathbf{z}_j\|_\infty\leq\varepsilon h/(8C_K\lambda_K) := \varepsilon\delta$. This can be done with $m\leq\varepsilon^{-p}\lceil\prod_{k=1}^p\big((b_k-a_k)/\delta\big)\rceil = \varepsilon^{-p}\lceil\mathcal{V}\delta^{-p}\rceil$ points. Then, with the notations of [99], $L=p$ and $D=\mathcal{V}(8C_K\lambda_K/h)^p$. As above, this yields

$$\mathbb{P}\Big(\sup_{\mathbf{z}\in\mathcal{Z}}\frac{1}{n^2}\Big|\sum_{1\leq i\neq j\leq n}\xi_{\mathbf{Z}}(\mathbf{X}_i,\mathbf{Z}_i,\mathbf{X}_j,\mathbf{Z}_j)\,,\,(\mathbf{X}_j,\mathbf{Z}_j)\big)\Big|>t\Big)\leq C_2D\exp\bigg(-\frac{\alpha_2nh^pt}{8(\int K^2)f_{\mathbf{Z},max}}\bigg), \tag{5.15}$$

when $t\leq 2h^p(\int K^2)^3f_{\mathbf{Z},max}^3/C_K^4$.

It remains to bound $\mathbb{P}(\sup_{\mathbf{z}\in\mathcal{Z}}|n^{-1}\sum_{i=1}^n\overline{g}_i|>t/4)$. Consider the family of functions

$$\mathcal{F} := \{(\mathbf{x}_1,\mathbf{z}_1)\in\mathbb{R}\times\mathcal{Z}\mapsto\frac{h^p}{4C_K^2}\mathbb{E}[g_{\mathbf{z}}(\mathbf{x}_1,\mathbf{z}_1,\mathbf{X},\mathbf{Z})],\,\mathbf{z}\in\mathcal{Z}\}.$$

This family of functions is bounded is one and its variance is less than $\overline{\sigma}^2 := h^p\int K^2f_{\mathbf{z},max}^3\big(\int|K|\big)^2$. Apply Propositions 9 and 10 in [50] that is coming from [47]: for some universal constants A_1 and A_2, some constant $A_{\overline{g}}$ that depends on K and $f_{\mathbf{z},max}$ (see Proposition 1 in [47]) and for every $x>0$,

$$\mathbb{P}\Big(\sup_{\mathbf{z}\in\mathcal{Z}}\frac{h^p}{4C_K^2}\Big|\sum_{i=1}^n\mathbb{E}[g_{\mathbf{z}}(\mathbf{X}_i,\mathbf{Z}_i,\mathbf{X},\mathbf{Z})|\mathbf{X}_i,\mathbf{Z}_i]\Big|>A_1\big(x+A_{\overline{g}}n^{1/2}\overline{\sigma}\ln(1/\overline{\sigma})\big)\Big)$$
$$\leq\quad 2\Big(\exp\big(-\frac{A_2x^2}{n\overline{\sigma}^2}\big)+\exp(-A_2x)\Big),$$

or

$$\mathbb{P}\Big(\sup_{\mathbf{z}\in\mathcal{Z}}\frac{1}{n}\Big|\sum_{i=1}^n\overline{g}_i\Big|>4A_1C_K^2\big(x-\frac{A_{\overline{g}}\overline{\sigma}}{n^{1/2}h^p}\ln(\overline{\sigma})\big)\Big)\leq 2\exp\big(-\frac{A_2nh^{2p}x^2}{\overline{\sigma}^2}\big)+2\exp(-A_2nh^px).$$

For any positive t s.t.

$$4A_1C_K^2\frac{(n-1)A_{\overline{g}}\overline{\sigma}}{n^{3/2}h^p}\ln(1/\overline{\sigma})<t/8,$$

note that we can find a real $x>th^p/(16C_K^2A_1)$. Then, we have

$$\mathbb{P}\Big(\sup_{\mathbf{z}\in\mathcal{Z}}\frac{(n-1)}{n^2}\Big|\sum_{i=1}^n\overline{g}_i\Big|>\frac{t}{4}\Big)\leq 2\exp\big(-\frac{A_2nh^pt^2C_K^{-4}}{16^2A_1^2\int K^2f_{\mathbf{z},max}^3(\int|K|)^2}\big)+2\exp(-\frac{A_2nh^pt}{16C_K^2A_1}). \tag{5.16}$$

Therefore, for such t, we obtain from (5.16) and (5.15) that

$$\mathbb{P}\Big(\sup_{\mathbf{z}\in\mathcal{Z}}|\sum_{1\le i\neq j\le n}\big(S_{i,j}(\mathbf{z})-\mathbb{E}[S_{i,j}(\mathbf{z})]\big)|\ge t\Big)\le C_2 D\exp\Big(-\frac{\alpha_2 nh^p t}{8(\int K^2)f_{\mathbf{Z},max}}\Big)$$
$$+\quad 2\exp\big(-\frac{A_2 nh^p t^2 C_K^{-4}}{15^2 A_1^2\int K^2 f_{\mathbf{z},max}^3(\int|K|)^2}\big)+2\exp(-\frac{A_2 nh^p t}{15C_K^2 A_1}).$$

for sufficiently large integers n. The latter inequality, (5.13) and (5.14) yield the exponential upper bound. □

5.5.6 Proof of Proposition 5.7

Let us note that $\tau_{1,2|\mathbf{Z}=\mathbf{z}}=\mathbb{E}\big[g_k(\mathbf{X}_1,\mathbf{X}_2)\big|\mathbf{Z}_1=\mathbf{z},\mathbf{Z}_2=\mathbf{z}\big]$ for every $k=1,2,3$, and that our estimators with the weights (5.2) can be rewritten as $\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}}:=U_n(g_k)\,/\,\{U_n(1)+\epsilon_n\}$ where

$$U_n(g):=\frac{1}{n(n-1)\mathbb{E}[K_h(\mathbf{z}-\mathbf{Z})]^2}\sum_{1\le i\neq j\le n}g(\mathbf{X}_i,\mathbf{X}_j)K_h(\mathbf{z}-\mathbf{Z}_i)K_h(\mathbf{z}-\mathbf{Z}_j)=:\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}g_{i,j},$$

for any measurable bounded function g, with the residual diagonal term $\epsilon_n:=\sum_{i=1}^n K_h^2(\mathbf{z}-\mathbf{Z}_i)/\{n(n-1)\mathbb{E}[K_h(\mathbf{z}-\mathbf{Z})]^2\}$. By Bochner's lemma (see Bosq and Lecoutre [21]), ϵ_n is $O_P((nh^p)^{-1})$, and it will be negligible compared to $U_n(1)$. Since the reasoning will be exactly the same for every estimator $\tau^{(k)}_{1,2|\mathbf{z}}$, i.e. for every function g_k, $k=1,2,3$, we omit the sub-index k. Then, the functions g_k will be simply denoted by g.

The expectation of our U-statistics is

$$\begin{aligned}\mathbb{E}\big[U_n(g)\big]&:=\mathbb{E}\big[g(\mathbf{X}_1,\mathbf{X}_2)K_h(\mathbf{z}-\mathbf{Z}_1)K_h(\mathbf{z}-\mathbf{Z}_2)\big]/\mathbb{E}[K_h(\mathbf{z}-\mathbf{Z})]^2\\&=\int g(\mathbf{x}_1,\mathbf{x}_2)K(\mathbf{t}_1)K(\mathbf{t}_2)f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1,\mathbf{z}+h\mathbf{t}_1)f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_2,\mathbf{z}+h\mathbf{t}_2)d\mathbf{x}_1\,d\mathbf{x}_2\,d\mathbf{t}_1\,d\mathbf{t}_2/\mathbb{E}[K_h(\mathbf{z}-\mathbf{Z})]^2\\&\to\frac{1}{f_{\mathbf{Z}}^2(\mathbf{z})}\int g(\mathbf{x}_1,\mathbf{x}_2)f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1,\mathbf{z})f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_2,\mathbf{z})d\mathbf{x}_1d\mathbf{x}_2=\mathbb{E}\big[g(\mathbf{X}_1,\mathbf{X}_2)\big|\mathbf{Z}_1=\mathbf{z},\mathbf{Z}_2=\mathbf{z}\big],\end{aligned}$$

applying Bochner's lemma to $\mathbf{z}\mapsto\int g(\mathbf{x}_1,\mathbf{x}_2)f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}(\mathbf{x}_1)f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}(\mathbf{x}_2)\,d\mathbf{x}_1\,d\mathbf{x}_2=\tau_{1,2|\mathbf{Z}=\mathbf{z}}$, that is a continuous function by assumption.

Set $\theta_n:=\mathbb{E}[U_n(g)]$, $g^*(\mathbf{x}_1,\mathbf{x}_2):=(g(\mathbf{x}_1,\mathbf{x}_2)+g(\mathbf{x}_2,\mathbf{x}_1))/2$ and $g^*_{i,j}=(g_{i,j}+g_{j,i})/2$ for every (i,j), $i\neq j$. Note that $U_n(g)=U_n(g^*)$. Since g^* is symmetrical, the Hájek projection $\hat{U}_n(g^*)$ of $U_n(g^*)$ satisfies

$$\hat{U}_n(g^*):=\frac{2}{n}\sum_{j=1}^n\mathbb{E}[g^*_{0,j}|\mathbf{X}_j,\mathbf{Z}_j]-\theta_n.$$

Note that $\mathbb{E}[\hat{U}_n(g^*)]=\theta_n=\tau_{1,2|\mathbf{Z}=\mathbf{z}}+o_P(1)$. Since $Var(\hat{U}_n(g^*)=4Var(\mathbb{E}[g^*_{0,j}|\mathbf{X}_j,\mathbf{Z}_j])/n=O((nh^p)^{-1})$, then $\hat{U}_n(g^*)=\theta_n+o_P(1)=\tau_{1,2|\mathbf{Z}=\mathbf{z}}+o_P(1)$.

Moreover, using the notation $\overline{g}_{i,j}:=g^*_{i,j}-\mathbb{E}[g^*_{i,j}|\mathbf{X}_j,\mathbf{Z}_j]-\mathbb{E}[g^*_{i,j}|\mathbf{X}_i,\mathbf{Z}_i]+\theta_n$ for $1\le i\neq j\le n$, we have $U_n(g^*)-\hat{U}_n(g^*)=\sum_{1\le i\neq j\le n}\overline{g}_{i,j}/n(n-1)$. By usual U-statistics calculations, it can be easily checked that

$$Var\big(U_n(g^*)-\hat{U}_n(g^*)\big)\quad=\quad\frac{1}{n^2(n-1)^2}\sum_{1\le i_1\neq j_1\le n}\sum_{1\le i_2\neq j_2\le n}\mathbb{E}[\overline{g}_{i_1,j_1}\overline{g}_{i_2,j_2}]=O(\frac{1}{n^2h^{2p}}).$$

Indeed, when all indices (i_1,i_2,j_1,j_2) are different, or when there is a single identity among them, $\mathbb{E}[\overline{g}_{i_1,j_1}\overline{g}_{i_2,j_2}]$ is zero. The first nonzero terms arise when there are two identities among the indices,

i.e. $i_1 = i_2$ and $j_1 = j_2$ (or $i_1 = j_2$ and $j_1 = i_2$). In the latter case, we get an upper bound as $O((nh^p)^{-2})$ when $f_{\mathbf{Z}}$ is continuous at $\mathbf{z}$, by usual changes of variable techniques and Bochner's Lemma. Then, $U_n(g^*) = \hat{U}_n(g^*) + o_P(1) = \tau_{1,2|\mathbf{Z}=\mathbf{z}} + o_P(1)$. Note that $U_n(1) + \epsilon_n$ tends to one in probability (Bochner's lemma). As a consequence, $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} = U_n(g^*) \,/\, (U_n(1) + \epsilon_n)$ tends to $\tau_{1,2|\mathbf{Z}=\mathbf{z}}/1$ by the continuous mapping theorem. □

5.5.7 Proof of Proposition 5.8

Let us note that

$$\tau_{1,2|\mathbf{Z}=\mathbf{z}} = \mathbb{E}\big[g_k(\mathbf{X}_1, \mathbf{X}_2)\big|\mathbf{Z}_1 = \mathbf{z}, \mathbf{Z}_2 = \mathbf{z}\big] = \int g_k(\mathbf{x}_1, \mathbf{x}_2) f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}(\mathbf{x}_1) f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}(\mathbf{x}_2) d\mathbf{x}_1 d\mathbf{x}_2 = \phi_k(\mathbf{z})/f^2_{\mathbf{Z}}(\mathbf{z})$$

where $\phi_k(\mathbf{z}) := \int g_k(\mathbf{x}_1, \mathbf{x}_2) f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1, \mathbf{z}) f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_2, \mathbf{z}) d\mathbf{x}_1 d\mathbf{x}_2$. We can also write $\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}} = \hat{\phi}_k(\mathbf{z})/\hat{f}^2_{\mathbf{Z}}(\mathbf{z})$, where $\hat{\phi}_k(\mathbf{z}) := n^{-2}\sum_{i,j=1}^n K_h(\mathbf{Z}_i - \mathbf{z}) K_h(\mathbf{Z}_j - \mathbf{z}) g_k(\mathbf{X}_i, \mathbf{X}_j)$ and $\hat{f}_{\mathbf{Z}}(\mathbf{z}) := n^{-1}\sum_{i=1}^n K_h(\mathbf{Z}_i - \mathbf{z})$. Therefore, we have

$$\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}} = \frac{\hat{\phi}_k(\mathbf{z}) - \phi_k(\mathbf{z})}{\hat{f}^2_{\mathbf{Z}}(\mathbf{z})} - \tau_{1,2|\mathbf{Z}=\mathbf{z}} \frac{\hat{f}_{\mathbf{Z}}(\mathbf{z}) - f_{\mathbf{Z}}(\mathbf{z})}{\hat{f}^2_{\mathbf{Z}}(\mathbf{z})} \times \big(\hat{f}_{\mathbf{Z}}(\mathbf{z}) + f_{\mathbf{Z}}(\mathbf{z})\big).$$

By usual uniform consistency results (see for example Bosq and Lecoutre [21]), $\sup_{\mathbf{z}\in\mathcal{Z}} \big|\hat{f}_{\mathbf{Z}}(\mathbf{z}) - f_{\mathbf{Z}}(\mathbf{z})\big| \to 0$ almost surely, as $n \to \infty$. We deduce that

$$\min_{\mathbf{z}\in\mathcal{Z}} \hat{f}^2_{\mathbf{Z}}(\mathbf{z}) \geq f^2_{\mathbf{Z},\min}/2, \text{ and } \max_{\mathbf{z}\in\mathcal{Z}} |\hat{f}_{\mathbf{Z}}(\mathbf{z}) + f_{\mathbf{Z}}(\mathbf{z})| \leq 2 \max_{\mathbf{z}\in\mathcal{Z}} f_{\mathbf{Z}}(\mathbf{z}) \text{ a.s.}$$

This means it is sufficient to prove the uniform strong consistency of $\hat{\phi}_k$ on $\mathcal{Z}$, to obtain that $\sup_{\mathbf{z}\in\mathcal{Z}} \big|\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}} - \tau^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}}\big|$ tends to zero a.s.

Note that, by Bochner's Lemma, $\sup_{\mathbf{z}\in\mathcal{Z}} \big|\mathbb{E}[\hat{\phi}_k(\mathbf{z})] - \phi_k(\mathbf{z})\big| \to 0$. Then, it remains to show that $\sup_{\mathbf{z}\in\mathcal{Z}} \big|\hat{\phi}_k(\mathbf{z}) - \mathbb{E}[\hat{\phi}_k(\mathbf{z})]\big| \to 0$ almost surely. Let $\rho_n > 0$ be such that we cover $\mathcal{Z}$ by the union of l_n open balls $B(\mathbf{t}_l, \rho_n)$, where $\mathbf{t}_1, \ldots, \mathbf{t}_{l_n} \in \mathbb{R}^p$ and $l_n \in \mathbb{N}^*$. Then

$$\sup_{\mathbf{z}\in\mathcal{Z}} \big|\hat{\phi}_k(\mathbf{z}) - \mathbb{E}[\hat{\phi}_k(\mathbf{z})]\big| \leq \sup_{l=1,\ldots l_n} \big|\hat{\phi}_k(\mathbf{t}_l) - \mathbb{E}[\hat{\phi}_k(\mathbf{t}_l)]\big| + A_n,$$

where $A_n := \sup_{l=1,\ldots l_n} \sup_{\mathbf{z}\in B(\mathbf{t}_l,\rho_n)} \big|\hat{\phi}_k(\mathbf{z}) - \hat{\phi}_k(\mathbf{t}_l) - (\mathbb{E}[\hat{\phi}_k(\mathbf{z})] - \mathbb{E}[\hat{\phi}_k(\mathbf{t}_l)])\big|$. For any index $l \in \{1, \ldots, l_n\}$ and any $\mathbf{z} \in B(\mathbf{t}_l, \rho_n)$, a first-order expansion yields

$$\begin{aligned}
&\big|\hat{\phi}_k(\mathbf{z}) - \hat{\phi}_k(\mathbf{t}_l) - (\mathbb{E}[\hat{\phi}_k(\mathbf{z})] - \mathbb{E}[\hat{\phi}_k(\mathbf{t}_l)])\big| \\
&= \Bigg| \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} g_k(\mathbf{X}_i, \mathbf{X}_j) K_h(\mathbf{z} - \mathbf{Z}_i) K_h(\mathbf{z} - \mathbf{Z}_j) \\
&\quad - \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} g_k(\mathbf{X}_i, \mathbf{X}_j) K_h(\mathbf{t}_l - \mathbf{Z}_i) K_h(\mathbf{t}_l - \mathbf{Z}_j) \\
&\quad - \Big(\mathbb{E}\big[g_k(\mathbf{X}_1, \mathbf{X}_2) K_h(\mathbf{z} - \mathbf{Z}_1) K_h(\mathbf{z} - \mathbf{Z}_2)\big] - \mathbb{E}\big[g_k(\mathbf{X}_i, \mathbf{X}_j) K_h(\mathbf{t}_l - \mathbf{Z}_i) K_h(\mathbf{t}_l - \mathbf{Z}_j)\big]\Big)\Bigg| \\
&\leq C_{Lip,K} h^{-2p-1} |\mathbf{z} - \mathbf{t}_l| \Big(\mathbb{E}\big[|g_k(\mathbf{X}_1, \mathbf{X}_2)|\big] + \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} |g_k(\mathbf{X}_i, \mathbf{X}_j)|\Big) \\
&= O(h^{-2p-1}\rho_n) = o(1),
\end{aligned}$$

for some constant $C_{Lip,K}$ and by choosing $\rho_n = o(h_n^{2p+1})$. Actually, we can cover $\mathcal{Z}$ in such a way that $l_n = O(h_n^{-p(2p+1)})$. This is always possible because $\mathcal{Z}$ is a bounded set in $\mathbb{R}^p$. Note that the previous upper bound is uniform w.r.t. l and $\mathbf{z} \in B(\mathbf{t}_l, \rho_n)$, proving that $A_n = o(1)$ everywhere.

Now, for every $l = 1, \dots, l_n$, apply Equation (5.8) for every $\mathbf{z} = \mathbf{t}_l$. For any $t > 0$, this provides

$$\mathbb{P}\left(\frac{1}{n(n-1)}\left|\sum_{i\neq j} g^{(l)}\big((\mathbf{X}_i,\mathbf{Z}_i),(\mathbf{X}_j,\mathbf{Z}_j)\big) - \mathbb{E}\Big[g^{(l)}\big((\mathbf{X}_1,\mathbf{Z}_1),(\mathbf{X}_2,\mathbf{Z}_2)\big)\Big]\right| > t\right) \leq \exp\Big(-\frac{C_0 n h_n^{2p} t^2}{C_1 + C_2 t}\Big),$$

for some positive constants C_0, C_1, C_2, by setting

$$g^{(l)}\big((\mathbf{X}_i,\mathbf{Z}_i),(\mathbf{X}_j,\mathbf{Z}_j)\big) := g_k(\mathbf{X}_i,\mathbf{X}_j) K_h(\mathbf{t}_l - \mathbf{Z}_i) K_h(\mathbf{t}_l - \mathbf{Z}_j).$$

Therefore, we deduce

$$\mathbb{P}\left(\sup_{l=1,\dots l_n} \big|\hat{\phi}_k(\mathbf{t}_l) - \mathbb{E}[\hat{\phi}_k(\mathbf{t}_l)]\big| \geq t\right) \leq C_4 h_n^{-p(2p+1)} \exp\Big(-\frac{C_0 n h_n^{2p} t^2}{C_1 + C_2 t}\Big),$$

for some constant C_4. Finally, applying Borel-Cantelli lemma, $\sup_{\mathbf{z}\in\mathcal{Z}} \big|\hat{\phi}_k(\mathbf{z}) - \mathbb{E}[\hat{\phi}_k(\mathbf{z})]\big|$ tends to zero a.s., proving the result. □

5.5.8 Proof of Proposition 5.9

By Markov's inequality, $\sum_{i=1}^n w_{i,n}^2(\mathbf{z}) = O_P((nh^p)^{-1})$ for any $\mathbf{z}$, that tends to zero. Then, by Slutsky's theorem, we get an asymptotic equivalence between the limiting laws of any $\hat{\tau}^{(k)}_{1,2|\mathbf{z}}$, $k = 1, 2, 3$, and of their linearly transformed versions $\tilde{\tau}_{1,2|\mathbf{z}}$. Thus, we will prove the asymptotic normality of $\hat{\tau}^{(k)}_{1,2|\mathbf{z}}$ for some index $k = 1, 2, 3$, simply denoted by $\hat{\tau}_{1,2|\mathbf{z}}$.

Let $g^*(\mathbf{x}_1, \mathbf{x}_2) := (g_k(\mathbf{x}_1,\mathbf{x}_2) + g_k(\mathbf{x}_2,\mathbf{x}_1))/2$ for some index $k = 1, 2, 3$ (that will be implicit in the proof). We now study the joint behavior of $(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i} - \tau_{1,2|\mathbf{Z}=\mathbf{z}'_i})_{i=1,\dots,n'}$. We will extend Stute [132]'s approach, in the case of multivariate conditioning variable $\mathbf{z}$ and studying the joint distribution of U-statistics at several conditioning points. As in the proof of Proposition 5.7, the estimator with the weights given by (5.2) can be rewritten as $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i} := U_{n,i}(g^*) \,/\, (U_{n,i}(1) + \epsilon_{n,i})$, where

$$U_{n,i}(g) := \frac{1}{n(n-1)\mathbb{E}[K_h(\mathbf{z}'_i - \mathbf{Z})]^2} \sum_{j_1,j_2=1, j_1\neq j_2}^{n} g(\mathbf{X}_{j_1},\mathbf{X}_{j_2}) K_h(\mathbf{z}'_i - \mathbf{Z}_{j_1}) K_h(\mathbf{z}'_i - \mathbf{Z}_{j_2}),$$

for any bounded measurable function $g : \mathbb{R}^4 \to \mathbb{R}$. Moreover, $\sup_{i=1,\dots,n'} |\epsilon_{n,i}| = O_P(n^{-1}h^{-p})$. By a limited expansion of $f_{\mathbf{X},\mathbf{Z}}$ w.r.t. its second argument, and under Assumption 5.3.4, we easily check that $\mathbb{E}\big[U_{n,i}(g)\big] = \tau_{1,2|\mathbf{Z}=\mathbf{z}'_i} + r_{n,i}$, where $|r_{n,i}| \leq C_0 h_n^\alpha / f^2_{\mathbf{Z}}(\mathbf{z}'_i)$, for some constant C_0 that is independent of i.

Now, we prove the joint asymptotic normality of $\big(U_{n,i}(g)\big)_{i=1,\dots,n'}$. The Hájek projection $\hat{U}_{n,i}(g)$ of $U_{n,i}(g)$ satisfies $\hat{U}_{n,i}(g) := 2\sum_{j=1}^n g_{n,i}\big(\mathbf{X}_j,\mathbf{Z}_j\big)/n - \theta_n$, where $\theta_n := \mathbb{E}\big[U_{n,i}(g)\big]$ and

$$g_{n,i}(\mathbf{x},\mathbf{z}) := K_h(\mathbf{z}'_i - \mathbf{z})\mathbb{E}\big[g(\mathbf{X},\mathbf{x}) K_h(\mathbf{z}'_i - \mathbf{Z})\big] \,/\, \mathbb{E}[K_h(\mathbf{z}'_i - \mathbf{Z})]^2.$$

Lemma 5.17. *Under the assumptions of Proposition 5.9, for any measurable bounded function g,*

$$(nh^p)^{1/2}\Big(\hat{U}_{n,i}(g) - \mathbb{E}\big[U_{n,i}(g)\big]\Big)_{i=1,\dots,n'} \xrightarrow{D} \mathcal{N}(0, M_\infty(g)), \text{ as } n \to \infty,$$

where, for $1 \leq i, j \leq n'$,

$$[M_\infty(g)]_{i,j} := \frac{4\int K^2 \mathbb{1}_{\{\mathbf{z}'_i = \mathbf{z}'_j\}}}{f_{\mathbf{Z}}(\mathbf{z}'_i)} \int g\big(\mathbf{x}_1,\mathbf{x}\big) g\big(\mathbf{x}_2,\mathbf{x}\big) f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_i}(\mathbf{x}) f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_i}(\mathbf{x}_1) f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_i}(\mathbf{x}_2) d\mathbf{x}\, d\mathbf{x}_1\, d\mathbf{x}_2.$$

This lemma is proved in Section 5.5.9. Similarly as in the proof of Lemma 2.2 in Stute [132], for every $i = 1, \ldots, n'$ and every bounded symmetrical measurable function g, we have $(nh^p)^{1/2} Var\big[\hat{U}_{n,i}(g) - U_{n,i}(g)\big] = o(1)$, which implies

$$(nh^p)^{1/2}\Big(U_{n,i}(g) - \mathbb{E}\big[U_{n,i}(g)\big]\Big)_{i=1,\ldots,n'} \xrightarrow{D} \mathcal{N}(0, M_\infty(g)), \text{ as } n \to \infty.$$

Considering two measurable and bounded functions g_1 and g_2, we have $U_{n,i}(c_1 g_1 + c_2 g_2) = c_1 U_{n,i}(g_1) + c_2 U_{n,i}(g_2)$ for every real numbers c_1, c_2. By the Cramér-Wold device, we easily state that

$$(nh^p)^{1/2}\left(\Big(U_{n,i}(g_1) - \mathbb{E}\big[U_{n,i}(g_1)\big]\Big)_{i=1,\ldots,n'}, \Big(U_{n,i}(g_2) - \mathbb{E}\big[U_{n,i}(g_2)\big]\Big)_{i=1,\ldots,n'}\right) \xrightarrow{D} \mathcal{N}\left(0, \begin{bmatrix} M_\infty(g_1) & M_\infty(g_1, g_2) \\ M_\infty(g_1, g_2) & M_\infty(g_2) \end{bmatrix}\right),$$

as $n \to \infty$, where

$$[M_\infty(g_1, g_2)]_{i,j} := \frac{4\int K^2 \mathbb{1}_{\{\mathbf{z}_i' = \mathbf{z}_j'\}}}{f_{\mathbf{Z}}(\mathbf{z}_i')} \int g_1(\mathbf{x}_1, \mathbf{x}) g_2(\mathbf{x}_2, \mathbf{x}) f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_i'}(\mathbf{x}) f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_i'}(\mathbf{x}_1) f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_i'}(\mathbf{x}_2) d\mathbf{x}\, d\mathbf{x}_1\, d\mathbf{x}_2.$$

Set $\tilde{\tau}_{1,2|\mathbf{Z}=\mathbf{z}_i'} := U_{n,i}(g^*) \,/\, U_{n,i}(1)$. Since $(nh_n^p)^{1/2}\big(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}_i'} - \tilde{\tau}_{1,2|\mathbf{Z}=\mathbf{z}_i'}\big) = O_P\big((nh_n^p)^{1/2}\epsilon_{n,i}\big)$ is $o_P(1)$, it is sufficient to establish the asymptotic law of $(nh_n^p)^{1/2}\big(\tilde{\tau}_{1,2|\mathbf{Z}=\mathbf{z}_i'} - \tau_{1,2|\mathbf{Z}=\mathbf{z}_i'}\big)$. Since $\mathbb{E}[U_{n,i}(1)] = 1 + o((nh^p)^{-1/2})$ and $\mathbb{E}[U_{n,i}(g^*)] = \tau_{1,2|\mathbf{Z}=\mathbf{z}_i'} + o((nh_n^p)^{-1/2})$, we get

$$(nh^p)^{1/2}\left(\Big(U_{n,i}(g^*) - \tau_{1,2|\mathbf{Z}=\mathbf{z}_i'}\Big)_{i=1,\ldots,n'}, \Big(U_{n,i}(1) - 1\Big)_{i=1,\ldots,n'}\right) \xrightarrow{D} \mathcal{N}\left(0, \begin{bmatrix} M_\infty(g^*) & M_\infty(g^*, 1) \\ M_\infty(g^*, 1) & M_\infty(1) \end{bmatrix}\right), \text{ as } n \to \infty.$$

Now apply the Delta-method with the function $\rho(\mathbf{x}, \mathbf{y}) := \mathbf{x}/\mathbf{y}$ where $\mathbf{x}$ and $\mathbf{y}$ are real-valued vectors of size n' and the division has to be understood component-wise. The Jacobian of ρ is given by the $n' \times 2n'$ matrix

$$J_\rho(\mathbf{x}, \mathbf{y}) = \Big[Diag\big(y_1^{-1}, \ldots y_{n'}^{-1}\big)\,,\, Diag\big(-x_1 y_1^{-2}, \cdots - x_{n'} y_{n'}^{-2}\big)\Big],$$

where, for any vector v of size n', $Diag(v)$ is the diagonal matrix whose diagonal elements are the v_i, with $i = 1, \ldots, n'$. We deduce $(nh^p)^{1/2}\left(\tilde{\tau}_{1,2|\mathbf{Z}=\mathbf{z}_i'} - \tau_{1,2|\mathbf{Z}=\mathbf{z}_i'}\right)_{i=1,\ldots,n'} \xrightarrow{D} \mathcal{N}(0, \mathbb{H})$, as $n \to \infty$, setting

$$\mathbb{H} := J_\rho(\vec{\tau}, \mathbf{e}) \begin{bmatrix} M_\infty(g^*) & M_\infty(g^*, 1) \\ M_\infty(g^*, 1) & M_\infty(1) \end{bmatrix} J_\rho(\vec{\tau}, \mathbf{e})^T,$$

where $\vec{\tau} = \left(\tau_{1,2|\mathbf{Z}=\mathbf{z}_i'}\right)_{i=1,\ldots,n'}$ and $\mathbf{e}$ is the vector of size n' whose all components are equal to 1. Thus, we have $J_\rho(\vec{\tau}, \mathbf{e}) = \Big[Id_{n'}, -Diag(\vec{\tau})\Big]$, denoting by $Id_{n'}$ the identity matrix of size n' and by $Diag(\vec{\tau})$ the diagonal matrix of size n' whose diagonal elements are the $\tau_{1,2|\mathbf{z}_i'}$, for $i = 1, \ldots, n'$. To be specific, we get

$$\mathbb{H} = M_\infty(g^*) - Diag(\vec{\tau}) M_\infty(g^*, 1) - M_\infty(g^*, 1) Diag(\vec{\tau}) + Diag(\vec{\tau}) M_\infty(1) Diag(\vec{\tau}).$$

For i, j in $\{1, \ldots, n'\}$ and using the symmetry of the function g^*, we obtain

$$[M_\infty(g^*)]_{i,j} = \frac{4\int K^2 \mathbb{1}_{\{\mathbf{z}_i' = \mathbf{z}_j'\}}}{f_{\mathbf{Z}}(\mathbf{z}_i')} \mathbb{E}[g^*(\mathbf{X}_1, \mathbf{X}) g^*(\mathbf{X}_2, \mathbf{X}) | \mathbf{Z} = \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}_i'],$$

$$\begin{aligned}[Diag(\vec{\tau}) M_\infty(g^*, 1)]_{i,j} &= \tau_{1,2|\mathbf{Z}=\mathbf{z}_i'} \frac{4\int K^2 \mathbb{1}_{\{\mathbf{z}_i' = \mathbf{z}_j'\}}}{f_{\mathbf{Z}}(\mathbf{z}_i')} \mathbb{E}[g^*(\mathbf{X}_1, \mathbf{X}) | \mathbf{Z} = \mathbf{Z}_1 = \mathbf{z}_i'] \\ &= \frac{4\int K^2 \mathbb{1}_{\{\mathbf{z}_i' = \mathbf{z}_j'\}}}{f_{\mathbf{Z}}(\mathbf{z}_i')} \tau_{1,2|\mathbf{Z}=\mathbf{z}_i'}^2 = [M_\infty(g^*, 1) Diag(\vec{\tau})]_{i,j} = [Diag(\vec{\tau}) M_\infty(1) Diag(\vec{\tau})]_{i,j}.\end{aligned}$$

As a consequence, we obtain

$$[\mathbb{H}]_{i,j} = \frac{4\int K^2 \mathbb{1}_{\{\mathbf{z}_i'=\mathbf{z}_j'\}}}{f_{\mathbf{Z}}(\mathbf{z}_i')}\Big(\mathbb{E}[g^*(\mathbf{X}_1,\mathbf{X})g^*(\mathbf{X}_2,\mathbf{X})|\mathbf{Z}=\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}_i'] - \tau^2_{1,2|\mathbf{Z}=\mathbf{z}_i'}\Big). \ \square$$

5.5.9 Proof of Lemma 5.17

Let us first evaluate the variance-covariance matrix $M_{n,n'} := [Cov(\hat{U}_{n,i},\hat{U}_{n,j})]_{1\le i,j\le n'}$. Note that $\mathbb{E}\big[g_{n,i}(\mathbf{X}_j,\mathbf{Z}_j)\big] = \mathbb{E}\big[\hat{U}_{n,i}\big] = \mathbb{E}\big[U_{n,i}(g)\big]$, and that

$$\Big((nh^p)^{1/2}\big(\hat{U}_{n,i} - \mathbb{E}[U_{n,i}(g)]\big)\Big)_{i=1,\dots,n'} = \frac{2h^{p/2}}{n^{1/2}}\sum_{j=1}^{n}\big(g_{n,i}(\mathbf{X}_j,\mathbf{Z}_j) - \mathbb{E}[U_{n,i}(g)]\big)_{i=1,\dots,n'},$$

that is a sum of independent vectors. Thus, $Cov(\hat{U}_{n,i},\hat{U}_{n,j}) = 4n^{-1}Cov\big(g_{n,i}(\mathbf{X},\mathbf{Z}), g_{n,j}(\mathbf{X},\mathbf{Z})\big)$, for every i,j in $\{1,\dots,n'\}$, and

$$\begin{aligned}
&\mathbb{E}\big[g_{n,i}(\mathbf{X},\mathbf{Z})g_{n,j}(\mathbf{X},\mathbf{Z})\big]\\
&= \int K_h(\mathbf{z}_i'-\mathbf{z})K_h(\mathbf{z}_j'-\mathbf{z})\frac{\mathbb{E}\big[g(\mathbf{X},\mathbf{x})K_h(\mathbf{z}_i'-\mathbf{Z})\big]\mathbb{E}\big[g(\mathbf{X},\mathbf{x})K_h(\mathbf{z}_j'-\mathbf{Z})\big]}{\mathbb{E}[K_h(\mathbf{z}_i'-\mathbf{Z})]^2\mathbb{E}[K_h(\mathbf{z}_j'-\mathbf{Z})]^2} f_{\mathbf{X},\mathbf{Z}}(\mathbf{x},\mathbf{z})d\mathbf{x}\,d\mathbf{z}\\
&\sim \frac{1}{h^p f^2_{\mathbf{Z}}(\mathbf{z}_i')f^2_{\mathbf{Z}}(\mathbf{z}_j')}\int g\big(\mathbf{x}_1,\mathbf{x}\big)g\big(\mathbf{x}_2,\mathbf{x}\big)K_h(\mathbf{z}_i'-\mathbf{z})K_h(\mathbf{z}_j'-\mathbf{z})K_h(\mathbf{z}_i'-\mathbf{w}_1)K_h(\mathbf{z}_j'-\mathbf{w}_2)\\
&\times\ f_{\mathbf{X},\mathbf{Z}}(\mathbf{x},\mathbf{z})f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1,\mathbf{w}_1)f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_2,\mathbf{w}_2)d\mathbf{x}\,d\mathbf{z}\,d\mathbf{x}_1\,d\mathbf{w}_1\,d\mathbf{x}_2\,d\mathbf{w}_2\\
&\sim \frac{1}{h^p f^2_{\mathbf{Z}}(\mathbf{z}_i')f^2_{\mathbf{Z}}(\mathbf{z}_j')}\int g\big(\mathbf{x}_1,\mathbf{x}\big)g\big(\mathbf{x}_2,\mathbf{x}\big)K(\mathbf{u}_1)K(\mathbf{u}_2)K(\mathbf{u})K(\frac{\mathbf{z}_j'-\mathbf{z}_i'}{h}+\mathbf{u})f_{\mathbf{X},\mathbf{Z}}(\mathbf{x},\mathbf{z}_i'-h\mathbf{u})\\
&\times\ f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1,\mathbf{z}_i'-h\mathbf{u}_1)f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_2,\mathbf{z}_j'-h\mathbf{u}_2)d\mathbf{x}\,d\mathbf{u}\,d\mathbf{x}_1\,d\mathbf{u}_1\,d\mathbf{x}_2\,d\mathbf{u}_2.
\end{aligned}$$

If $i\neq j$ and K is compactly supported, the latter term is zero when n is sufficiently large, and

$$Cov(\hat{U}_{n,i},\hat{U}_{n,j}) = -4n^{-1}\mathbb{E}[U_{n,i}]\mathbb{E}[U_{n,j}] \sim -4n^{-1}\tau_{1,2|\mathbf{Z}=\mathbf{z}_i'}\tau_{1,2|\mathbf{Z}=\mathbf{z}_j'} = o\big((nh^p)^{-1}\big).$$

Otherwise, $i=j$ and, as $\mathbb{E}\Big[g_{n,i}\big(\mathbf{X}_1,\mathbf{Z}_1\big)\Big] = O(1)$, we have

$$\begin{aligned}
Var\Big(\big(g_{n,i}(\mathbf{X},\mathbf{Z})\big)^2\Big) &\sim \frac{1}{h^p f^4_{\mathbf{Z}}(\mathbf{z}_i')}\int g\big(\mathbf{x}_1,\mathbf{x}\big)g\big(\mathbf{x}_2,\mathbf{x}\big)K(\mathbf{u}_1)K(\mathbf{u}_2)K^2(\mathbf{u})f_{\mathbf{X},\mathbf{Z}}(\mathbf{x},\mathbf{z}_i'-h\mathbf{u})\\
&\qquad\times f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1,\mathbf{z}_i'-h\mathbf{u}_1)f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_2,\mathbf{z}_i'-h\mathbf{u}_2)\,d\mathbf{x}\,d\mathbf{u}\,d\mathbf{x}_1\,d\mathbf{u}_1\,d\mathbf{x}_2\,d\mathbf{u}_2\\
&\sim \frac{\int K^2}{h^p f_{\mathbf{Z}}(\mathbf{z}_i')}\int g\big(\mathbf{x}_1,\mathbf{x}\big)g\big(\mathbf{x}_2,\mathbf{x}\big)f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_i'}(\mathbf{x})f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_i'}(\mathbf{x}_1)f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_i'}(\mathbf{x}_2)\,d\mathbf{x}\,d\mathbf{x}_1\,d\mathbf{x}_2,
\end{aligned}$$

by Bochner's lemma. We have proved that, for every $i,j\in\{1,\dots,n'\}$,

$$nh^p[M_{n,n'}]_{i,j} \to \frac{4\int K^2\mathbb{1}_{\{\mathbf{z}_i'=\mathbf{z}_j'\}}}{f_{\mathbf{Z}}(\mathbf{z}_i')}\int g\big(\mathbf{x}_1,\mathbf{x}\big)g\big(\mathbf{x}_2,\mathbf{x}\big)f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_i'}(\mathbf{x})f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_i'}(\mathbf{x}_1)f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_i'}(\mathbf{x}_2)\,d\mathbf{x}\,d\mathbf{x}_1\,d\mathbf{x}_2,$$

as $n\to\infty$. Therefore, $nh^pM_{n,n'}$ tends to M_∞.

We now verify Lyapunov's condition with third-order moments, so that the usual multivariate central limit theorem would apply. It is then sufficient to show that

$$\Big(\frac{h^{p/2}}{n^{1/2}}\Big)^3\sum_{j=1}^{n}\mathbb{E}\Big[\big|g_{n,i}(\mathbf{X}_j,\mathbf{Z}_j) - \mathbb{E}[U_{n,i}(g)]\big|^3\Big] = o(1). \tag{5.17}$$

For any $j=1,\dots,n$, we have

$$\begin{aligned}
&\mathbb{E}\Big[\big|g_{n,i}(\mathbf{X}_j,\mathbf{Z}_j) - \mathbb{E}[U_{n,i}(g)]\big|^3\Big]\\
&\sim \int\Big|\frac{1}{f^2_{\mathbf{Z}}(\mathbf{z}_i')}\int g(\mathbf{x}_1,\mathbf{x})K_h(\mathbf{z}_i'-\mathbf{z}_1)K_h(\mathbf{z}_i'-\mathbf{z})f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1,\mathbf{z}_1)d\mathbf{x}_1\,d\mathbf{z}_1 - \mathbb{E}\big[U_{n,i}(g)\big]\Big|^3 f_{\mathbf{X},\mathbf{Z}}(\mathbf{x},\mathbf{z})d\mathbf{x}\,d\mathbf{z}.
\end{aligned}$$

By the change of variable $\mathbf{z}_1 = \mathbf{z}_i' - h\mathbf{t}_1$ and $\mathbf{z} = \mathbf{z}_i' - h\mathbf{t}$, we get

$$\mathbb{E}\Big[\big|g_{n,i}(\mathbf{X}_j, \mathbf{Z}_j) - \mathbb{E}[U_{n,i}(g)]\big|^3\Big] \sim h^{-2p} \int \Big| \frac{1}{f_{\mathbf{Z}}^2(\mathbf{z}_i')} \int g(\mathbf{x}_1, \mathbf{x}) K(\mathbf{t}_1) K(\mathbf{t}) f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1, \mathbf{z}_i' - h\mathbf{t}_1) d\mathbf{x}_1 \, d\mathbf{t}_1$$
$$- \quad h^p \mathbb{E}\big[U_{n,i}(g)\big] \Big|^3 f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}, \mathbf{z}_i' - h\mathbf{t}) d\mathbf{x} \, d\mathbf{t} = O(h^{-2p}),$$

because of Bochner's lemma, under our assumptions. Therefore, we have obtained

$$\Big(\frac{h^{p/2}}{n^{1/2}}\Big)^3 \sum_{j=1}^{n} \mathbb{E}\Big[\big|g_{n,i}(\mathbf{X}_j, \mathbf{Z}_j) - \mathbb{E}[U_{n,i}(g)]\big|^3\Big] = O(h^{3p/2} n^{-3/2} n h^{-2p}) = O((nh^p)^{-1/2}) = o(1).$$

Therefore, we have checked Lyapunov's condition and the result follows. □

Chapter 6

About Kendall's regression

Abstract

Conditional Kendall's tau is a measure of dependence between two random variables, conditionally on some covariates. We assume a regression-type relationship between conditional Kendall's tau and some covariates, in a parametric setting with a large number of transformations of a small number of regressors. This model may be sparse, and the underlying parameter is estimated through a penalized criterion. We prove non-asymptotic bounds with explicit constants that hold with high probabilities. We derive the consistency of a two-step estimator, its asymptotic law and some oracle properties. Some simulations and applications to real data conclude the paper.

Keywords: Conditional dependence measures, kernel smoothing, regression-type models, conditional Kendall's tau.

Based on [39]: Derumigny, A., & Fermanian, J. D., About Kendall's regression. *ArXiv preprint*, arXiv:1802.07613, 2018.

6.1 Introduction

In dependence modeling, it is common to work with scalar dependence measures which are margin-free. They can be used to quantify the positive or negative relationship between two random variables X_1 and X_2. One of the most popular of them is Kendall's tau, a dependence measure defined by

$$\tau_{1,2} := \mathbb{P}\big((X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2}) > 0\big) - \mathbb{P}\big((X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2}) < 0\big),$$

where $(X_{i,1}, X_{i,2})$, $i = 1, 2$ are i.i.d. copies of (X_1, X_2), see [106]. When a covariate $\mathbf{Z}$ is available, it is natural to work with the conditional version of this, i.e. the conditional Kendall's tau. It is defined as

$$\begin{aligned}\tau_{1,2|\mathbf{Z}=\mathbf{z}} := \mathbb{P}\big((X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2}) > 0 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big)\\ - \quad \mathbb{P}\big((X_{1,1} - X_{2,1})(X_{1,2} - X_{2,2}) < 0 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big),\end{aligned}$$

where $(X_{i,1}, X_{i,2}, \mathbf{Z}_i)$, $i = 1, 2$ are i.i.d. copies of $(X_1, X_2, \mathbf{Z})$. In such a model, the goal is to study to what extent a p-dimensional covariate $\mathbf{z}$ can affect the dependence between the two variables of interest X_1 and X_2.

Most often, it is difficult to have a clear intuition about the functional link between some measure of dependence and the underlying explanatory variables. Sometimes, it is even unclear whether the

covariates have an influence on the dependence between the variables of interest. This is the so-called "simplifying assumption", well-known in the world of copula modeling (see [38] and the references therein). This issue is particularly crucial with pair-copula constructions, as pointed out in [68], [5], [87], among others. In our case, we will evaluate an explicit and flexible link between some dependence measure, the Kendall's tau, and the vector of covariates. As a sub-product of our model, we will be able to provide a test of the "simplifying assumption".

Given a dataset $(X_{i,1}, X_{i,2}, \mathbf{Z}_i)$, $i = 1, \ldots, n$, we will focus on the function $\mathbf{z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$ for $\mathbf{z} \in \mathcal{Z}$, where $\mathcal{Z}$ denotes a compact subset of $\mathbb{R}^p$. This $\mathcal{Z}$ represents a set of "reasonable" values for $\mathbf{z}$, so that the density $f_{\mathbf{Z}}$ is bounded from below on $\mathcal{Z}$. In order to simplify notation, the reference to the conditioning event $\mathbf{Z} \in \mathcal{Z}$ will be omitted. A first natural choice would be to invoke a nonparametric estimator of $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ as in [63], [141] and [40]. Here, we prefer to obtain parameters that can be interpreted and that would sum up the information about the conditional Kendall's tau. Moreover, kernel-based estimation can be very costly under a computational point of view: for m values of $\mathbf{z}$, the prediction of all these conditional Kendall's taus has a total cost of $O(mn^2)$, that can be large if a large number m is required. Other estimators of the conditional Kendall's tau, based on classification methods, are proposed in [41].

In this paper, our idea is to decompose the function $\mathbf{z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$ on some functional basis $(\psi_i)_{i\geq 1}$, as any element of a space of functions from $\mathcal{Z}$ to $\mathbb{R}$. First note that a Kendall's tau takes its values in the interval $[-1, 1]$, and not on the whole real line. Nevertheless, for some known increasing and continuously differentiable function $\Lambda : [-1, 1] \to \mathbb{R}$, the function $\mathbf{z} \mapsto \Lambda\left(\tau_{1,2|\mathbf{Z}=\mathbf{z}}\right)$ takes values on up to the whole real line potentially, and it can be decomposed on any basis $(\psi_i)_{i\geq 1}$. Typical transforms are $\Lambda(\tau) = \log\left(\frac{1+\tau}{1-\tau}\right)$ (the Fisher transform) or $\Lambda(\tau) = \log(-\log((1-\tau)/2))$. We will assume that only a finite number of elements are necessary to represent this function. This means that we have

$$\Lambda\left(\tau_{1,2|\mathbf{Z}=\mathbf{z}}\right) = \sum_{i=1}^{p'} \psi_i(\mathbf{z})\beta_i^* = \boldsymbol{\psi}(\mathbf{z})^T\beta^*, \tag{6.1}$$

for all $\mathbf{z} \in \mathcal{Z}$, with $p' > 0$ and a "true" unknown parameter $\beta^* \in \mathbb{R}^{p'}$. The function $\boldsymbol{\psi}(\cdot) := \big(\psi_1(\cdot), \ldots, \psi_{p'}(\cdot)\big)^T$ from $\mathbb{R}^p$ to $\mathbb{R}^{p'}$ is known and corresponds to deterministic transformations of the covariates $\mathbf{z}$. In practice, it is not easy to have intuition about which kind of basis to use, especially in our framework of conditional dependence measurement. Therefore, the most simple solution is the use of a lot of different functions : polynomials, exponentials, sinuses and cosinuses, indicator functions, etc... They allow to take into account potential non-linearities and even discontinuities of conditional Kendall's taus with respect to $\mathbf{z}$. For the sake of identifiability, we only require their linear independence, as seen in the following proposition (whose straightforward proof is omitted).

Proposition 6.1. *The parameter β^* in Model (6.1) is identifiable if and only if the functions $(\psi_1, \ldots, \psi_{p'})$ are linearly independent $\mathbb{P}_{\mathbf{Z}}$-a.e. in the sense that, for any given vector $\mathbf{t} = (t_1, \ldots, t_{p'}) \in \mathbb{R}^{p'}$, $\mathbb{P}_{\mathbf{Z}}\big(\boldsymbol{\psi}(\mathbf{Z})^T\mathbf{t} = 0\big) = 1$ implies $\mathbf{t} = 0$.*

With such a large choice among flexible classes of functions, it is unlikely we will be able to guess the right ones ex ante. Therefore, it will be necessary to consider a large number of functions ψ_i under a sparsity constraint: the cardinality of $\mathcal{S}$, the set of non-zero components of β^*, is less than some $s \in \{1, \ldots, p'\}$. It is denoted by $|\mathcal{S}| = |\beta^*|_0$, where $|\cdot|_0$ yields the number of non-zero components of any vector in $\mathbb{R}^{p'}$. Note that, in this framework, p' can be moderately large, for example 10 or 30 while the original dimension p is small, for example $p = 1$ or 2. This corresponds to the decomposition of a function, defined on a small-dimension domain, in a mildly large basis.

Once an estimator $\hat{\beta}$ of β^* has been computed, the prediction of all the conditional Kendall's tau's for m values of $\mathbf{z}$, which is just the computation of $\Lambda^{(-1)}\big(\boldsymbol{\psi}(\mathbf{z})^T\hat{\beta}\big)$ can be done in $O(ms)$, that is much faster than what was previously required with a kernel-based estimator for large m, as soon as $s \leq n^2$ (see Section 6.4.1 for a discussion).

Estimating Model (6.1) not only provides an estimator of the conditional Kendall's tau $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$, but also easily provides estimators of the marginal effects of $\mathbf{z}$ as by-product. For example, given $\mathbf{z} \in \mathcal{Z}$, the marginal effect of z_1, i.e. $\partial\tau_{1,2|\mathbf{Z}=\mathbf{z}}(\mathbf{z})/\partial z_1$, can be directly estimated by $\big(\partial_{z_1}\boldsymbol{\psi}(\mathbf{z})\big)^T\hat{\beta}\cdot\Lambda^{(-1)\prime}\big(\boldsymbol{\psi}(\mathbf{z})^T\hat{\beta}\big)$, assuming that $\boldsymbol{\psi}$ and $\Lambda^{(-1)}$ are differentiable respectively at $\mathbf{z}$ and $\boldsymbol{\psi}(\mathbf{z})^T\hat{\beta}$. Such sensitivities can be useful in many applications.

A desirable empirical feature of Model (6.1) would be the possibility of obtaining very high/low levels of dependence between X_1 and X_2, for some $\mathbf{Z}$ values, i.e. $\Lambda^{(-1)}(\boldsymbol{\psi}(\mathbf{z})^T\beta^*)$ should be close (or even equal) to 1 or -1 for some $\mathbf{z}$. This can be the case even if $\mathcal{Z}$ is compact, that is here required for theoretical reasons. Indeed, the image of $\{\tau_{1,2|\mathbf{z}}|\mathbf{z} \in \mathcal{Z}\} = [\tau_{\min}, \tau_{\max}]$ through Λ is an interval $[\Lambda_{\min}, \Lambda_{\max}]$. If $\boldsymbol{\psi}(\mathbf{z})^T\beta^* \geq \Lambda_{\max}$ (resp. $\boldsymbol{\psi}(\mathbf{z})^T\beta^* \leq \Lambda_{\min}$), then simply set $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = \tau_{\max}$ or even one (resp. $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = \tau_{\min}$ or even (-1)).

Contrary to more usual models, the "explained variable" - the conditional Kendall's tau $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ - is not observed in (6.1). Therefore, a direct estimation of the parameter β^* (for example, by the ordinary least squares, or by the Lasso) is unfeasible. In other words, even if the function $\mathbf{z} \mapsto \Lambda\big(\tau_{1,2|\mathbf{Z}=\mathbf{z}}\big)$ is deterministic, finding the best β in Model (6.1) is far from being just a numerical analysis problem since the function to be decomposed is unknown. Nevertheless, we will replace $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ by a nonparametric estimate $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}$, and use it as an approximation of the explained variable. More precisely, we fix a finite collection of points $\mathbf{z}'_1, \ldots, \mathbf{z}'_{n'} \in \mathcal{Z}^{n'}$ and we estimate $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}$ for each of these points. Then, $\hat{\beta}$ is estimated as the minimizer of the l_1-penalized criteria

$$\hat{\beta} := \arg\min_{\beta\in\mathbb{R}^{p'}} \Big[\frac{1}{n'}\sum_{i=1}^{n'}\big(\Lambda(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i}) - \boldsymbol{\psi}(\mathbf{z}'_i)^T\beta\big)^2 + \lambda|\beta|_1\Big], \tag{6.2}$$

where λ is a positive tuning parameter (that may depend on n and n'), and $|\cdot|_q$ denotes the l_q norm, for $1 \leq q \leq \infty$. This procedure is summed up in the following Algorithm 5. Note that even if we study the general case with any $\lambda \geq 0$, the properties of the unpenalized estimator can be derived by choosing the particular case $\lambda = 0$.

Algorithm 5: Two-step estimation of β

Input: A dataset $(X_{i,1}, X_{i,2}, \mathbf{Z}_i)$, $i = 1, \ldots, n$
Input: A finite collection of points $\mathbf{z}'_1, \ldots, \mathbf{z}'_{n'} \in \mathcal{Z}^{n'}$
for $j \leftarrow 1$ **to** n' **do**
  Compute the estimator $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_j}$ using the sample $(X_{i,1}, X_{i,2}, \mathbf{Z}_i)$, $i = 1, \ldots, n$;
end
Compute the minimizer $\hat{\beta}$ of (6.2) using the $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_j}$, $j = 1, \ldots, n'$, estimated in the above step ;
Output: An estimator $\hat{\beta}$.

Several nonparametric estimators of $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_j}$ can potentially be used. We refer to [40] for a detailed analysis of their statistical properties. They are of the form

$$\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} := \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,n}(\mathbf{z})w_{j,n}(\mathbf{z})g^*(\mathbf{X}_i, \mathbf{X}_j), \tag{6.3}$$

where g^* is a bounded function, $\mathbf{X}_i := (X_{i,1}, X_{i,2})$ for $i = 1, \dots, n$ and $w_{i,n}(\mathbf{z}) := K_h(\mathbf{Z}_i - \mathbf{z}) / \sum_{j=1}^n K_h(\mathbf{Z}_j - \mathbf{z})$, $h = h(n) > 0$ denoting the bandwidth sequence. In the same way, the conditional Kendall's tau can be rewritten as $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = \mathbb{E}[g^*(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}]$ for the same choices of g^*. Possible choices of g^* are given in Section 6.9.

In Section 6.2, we state non-asymptotic results for the our estimator $\hat{\beta}$ that hold with high probability. In Section 6.3, its asymptotic properties are stated. In particular, we will study the cases when n' is fixed and $n \to \infty$, and when both indices tend to the infinity. We also give some oracle properties and suggest a related adaptive estimator. Sections 6.4 and 6.5 illustrate respectively the numerical performances of $\hat{\beta}$ on simulated and real data. All proofs and two supplementary figures have been postponed at the end of the chapter.

Remark 6.2. *At first sight, in Model (6.1), there seems to be no noise perturbing the variable of interest. In fact, this is a simple consequence of our formulation of the model. In the same way, a classical linear model $Y = \mathbf{X}^T\beta^* + \varepsilon$ can be rewritten as $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = \mathbf{x}^T\beta^*$ without any explicit noise. By definition, $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ is a deterministic function of a given $\mathbf{x}$. In our case, $\Lambda\big(\tau_{1,2|\mathbf{Z}=\mathbf{z}}\big)$ is a deterministic function of the variable $\mathbf{z}$. This means that we cannot formally write a model with noise, such as $\Lambda\big(\tau_{1,2|\mathbf{Z}=\mathbf{z}}\big) = \boldsymbol{\psi}(\mathbf{z})^T\beta^* + \varepsilon$ where ε is independent of the choice of $\mathbf{z}$. Indeed, the left-hand side of the latter equality is a $\mathbf{z}$-mesurable quantity, unless ε is constant almost surely.*

Remark 6.3. *Note that the conditioning event of Model (6.1) is unusual: usual regression models consider $\mathbb{E}[g(\mathbf{X})|\mathbf{Z} = \mathbf{z}]$ as a function of the conditioning variable $\mathbf{z}$. Here, the probabilities of concordant/discordant pairs are made conditionally on $\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}$. This unusual conditioning event will necessitate some peculiar theoretical treatments.*

Remark 6.4. *Instead of a fixed design setting $(\mathbf{z}'_i)_{i=1,\dots,n'}$ in the optimization program, it would be possible to consider a random design: simply draw n' realizations of $\mathbf{Z}$, independently of the n-sample that has been used for the estimation of the conditional Kendall's taus. The differences between fixed and random designs are mainly a matter of presentation and the reader could easily rewrite our results in a random design setting. We have preferred the former one to study the finite distance properties and asymptotics when n' is fixed (Section 6.3.1). When n and n' will tend to the infinity (Section 6.3.3), both designs are encompassed de facto because we will assume the weak convergence of the empirical distribution associated to the sample $(\mathbf{z}'_i)_{i=1,\dots,n'}$, when $n' \to \infty$.*

6.2 Finite-distance bounds on $\hat{\beta}$

Our first goal is to prove finite-distance bounds in probability for the estimator $\hat{\beta}$. Let $\mathbb{Z}'$ be the matrix of size $n' \times p'$ whose lines are $\boldsymbol{\psi}(\mathbf{z}'_i)^T$, $i = 1, \dots, n'$, and let $\mathbf{Y} \in \mathbb{R}^{n'}$ be the column vector whose components are $Y_i = \Lambda(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i})$, $i = 1, \dots, n'$. For a vector $\mathbf{v} \in \mathbb{R}^{p'}$, denote by $||\mathbf{v}||_{n'} := |\mathbf{v}|_2/\sqrt{n'}$ its empirical norm. We can then rewrite the criterion (6.2) as $\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^{p'}} \Big[||\mathbf{Y} - \mathbb{Z}'\beta||^2_{n'} + \lambda|\beta|_1\Big]$, where $\mathbf{Y}$ and $\mathbb{Z}'$ may be considered as "observed", so that the practical problem is reduced to a standard Lasso estimation procedure. Define some "residuals" by $\xi_{i,n} := \Lambda(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i}) - \boldsymbol{\psi}(\mathbf{z}'_i)^T\beta^* = \Lambda(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i}) - \Lambda(\tau_{1,2|\mathbf{Z}=\mathbf{z}'_i})$, for $i = 1, \dots, n'$. Note that these $\xi_{i,n}$ are not "true residuals" in the sense that they do not depend on the estimator $\hat{\beta}$, but on the true parameter β^*. We also emphasized the dependence on n in the notation $\xi_{i,n}$, which is a consequence of the estimated conditional Kendall's tau.

To get non-asymptotic bounds on $\hat{\beta}$, assume the *Restricted Eigenvalue* (RE) condition, introduced by [19]. For $c_0 > 0$ and $s \in \{1, \dots, p\}$, assume

$RE(s, c_0)$ **condition :** *The design matrix $\mathbb{Z}'$ satisfies*

$$\kappa(s, c_0) := \min_{\substack{J_0 \subset \{1,\dots,p'\} \\ Card(J_0) \leq s}} \min_{\substack{\delta \neq 0 \\ |\delta_{J_0^C}|_1 \leq c_0 |\delta_{J_0}|_1}} \frac{|\mathbb{Z}'\delta|_2}{\sqrt{n'}|\delta|_2} > 0.$$

Note that this condition is very mild, and is satisfied with a high probability for a large class of random matrices: see [12, Section 8.1] for references and a discussion.

Assumption 6.2.1. *The function $\mathbf{z} \mapsto \boldsymbol{\psi}(\mathbf{z})$ are bounded on $\mathcal{Z}$ by a constant $C_{\boldsymbol{\psi}}$. Moreover, $\Lambda(\cdot)$ is continuously differentiable. Let $\mathcal{T}$ be the range of $\mathbf{z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$, from $\mathcal{Z}$ towards $[-1, 1]$. On an open neighborhood of $\mathcal{T}$, the derivative of $\Lambda(\cdot)$ is bounded by a constant $C_{\Lambda'}$.*

Theorem 6.5 (Fixed design case)**.** *Suppose that Assumptions 6.9.1-6.9.4 and 6.2.1 hold and that the design matrix $\mathbb{Z}'$ satisfies the $RE(s, 3)$ condition. Choose the tuning parameter as $\lambda = \gamma t$, with $\gamma \geq 4$ and $t > 0$, and assume that we choose h small enough such that*

$$h^\alpha \leq \min\left(\frac{f_{\mathbf{Z},min}\alpha!}{4\, C_{K,\alpha}} \,,\, \frac{f^4_{\mathbf{Z},min}\alpha!\, t}{8\, C_\psi C_{\Lambda'}(f^2_{\mathbf{Z},min} + 8 f^2_{\mathbf{Z},max}) C_{\mathbf{XZ},\alpha}} \right). \tag{6.4}$$

Then, we have

$$\begin{aligned} \mathbb{P}\Big(||\mathbb{Z}'(\hat{\beta} - \beta^*)||_{n'} \leq \frac{4(\gamma+1)\tilde{t}\sqrt{s}}{\kappa(s,3)} \text{ and } |\hat{\beta} - \beta^*|_q \leq \frac{4^{2/q}(\gamma+1)\tilde{t}s^{1/q}}{\kappa^2(s,3)}, \text{ for every } 1 \leq q \leq 2 \Big) \\ \geq 1 - 2n' \exp\Big(- nh^p C_1 \Big) - 2n' \exp\Big(- \frac{(n-1)h^{2p}t^2}{C_2 + C_3 t} \Big) \\ - 2\exp\Big(- \frac{nh^p (f_{\mathbf{Z},max} - C_{\tilde{K},2}h^2)^2}{C_4 + 4C_{\tilde{K}}(f_{\mathbf{Z},min} - C_{\tilde{K},2}h^2)/3} \Big), \end{aligned} \tag{6.5}$$

where $C_1 := f^2_{\mathbf{Z},min}/\big(32 f_{\mathbf{Z},max}\int K^2 + (8/3)C_K f_{\mathbf{Z},min}\big)$, $C_2 := \{16 C_\psi C_{\Lambda'}(f^2_{\mathbf{Z},min} + 8 f^2_{\mathbf{Z},max}) f_{\mathbf{Z},max}\int K^2\}^2 / f^8_{\mathbf{Z},min}$, $C_3 := (64/3) C_\psi C_{\Lambda'} C_K^2 (f^2_{\mathbf{Z},min} + 8 f^2_{\mathbf{Z},max})/f^4_{\mathbf{Z},min}$, $C_4 := 8 f_{\mathbf{Z},max}\int \tilde{K}^2$, *and* $\tilde{t} := t + 3\int K^2/(nh^p f_{\mathbf{Z},min})$.

This theorem, proved in Section 6.6.2, yields some bounds that hold in probability for the prediction error $||\mathbb{Z}'(\hat{\beta} - \beta^*)||_{n'}$ and for the estimation error $|\hat{\beta} - \beta^*|_q$, $1 \leq q \leq 2$, under the specification (6.1). Note that the influence of n' and p' is hidden through the Restricted Eigenvalue number $\kappa(s, 3)$. The result depends on three parameters γ, t and h. Apparently, the choice of γ seems to be easy, as a larger γ deteriorates the upper bounds. Nonetheless, it is a bit misleading because $\hat{\beta}$ implicitly depends on λ and then on γ (for a fixed t). Nonetheless, choosing $\gamma = 4$ is a reasonable "by default" choice. Moreover, a lower t provides a smaller upper bound, but at the same time the probability of this event is lowered. This induces a trade-off between the probability of the desired event and the size of the bound, as we want the smallest possible bound with the highest probability. Moreover, we cannot choose a too small t, because of the lower bound (6.4): t is limited by a value proportional to h^α. The latter h cannot be chosen as too small, otherwise the probability in Equation (6.5) will decrease. To be short: *low values of h and t yield a sharper upper bound with a lower probability, and the opposite*. Therefore, a trade-off has to be found, depending of the kind of result we are interested in.

Clearly, we would like to exhibit the sharpest upper bounds in (6.5), with the "highest probabilities". Let us look for parameters of the form $t \propto n^{-a}$ and $h \propto n^{-b}$, with $a, b > 0$. The assumptions of Theorem 6.5 imply $b\alpha \geq a$ (to satisfy (6.4)) and $1 - 2a - 2pb > 0$ (so that the right-hand side of (6.5) tends to 1 as $n \to \infty$, i.e. $nh^p \to \infty$ and $nt^2h^{2p} \to \infty$). For fixed α and p, what are the "optimal" choices a and b under the constraints $b\alpha \geq a$ and $1 - 2a - 2pb > 0$? The latter domain is the interior of a triangle in the plane

$(a,b) \in \mathbb{R}_+^2$, whose vertices are $O := (0,0)$, $A := (0, 1/(2p))$ and $B := (\alpha/(2p+2\alpha), 1/(2p+2\alpha))$, plus the segment $]0, B[$. All points in such a domain would provide admissible couples (a,b) and then admissible tuning parameters (t,h). In particular, choosing the neighborhood of B, i.e. $a = \alpha(1-\epsilon)/(2p+2\alpha)$ and $b = 1/(2p+2\alpha)$ for some (small) $\epsilon > 0$, will be nice because the upper bounds will be minimized.

Corollary 6.6. *For $0 < \epsilon < 1$, choosing the parameters $\lambda = 4t$, $t = (n-1)^{-\alpha(1-\epsilon)/(2\alpha+2p)}$ and*

$$h = c_h (n-1)^{-1/(2\alpha+2p)},\ c_h := \Big(\frac{f_{\mathbf{Z},min}^4 \alpha!}{2\, C_\psi C_{\Lambda'} (f_{\mathbf{Z},min}^2 + 16 f_{\mathbf{Z},max}^2) C_{\mathbf{XZ},\alpha}} \Big)^{1/\alpha},$$

we have, if n is sufficiently large so that (6.4) is satisfied,

$$\begin{aligned}
\mathbb{P}\Big(||\mathbb{Z}'(\hat\beta - \beta^*)||_{n'} \le \frac{20\sqrt{s}}{\kappa(s,3)(n-1)^{\alpha(1-\epsilon)/(2\alpha+2p)}} \ \textit{and} \\
|\hat\beta - \beta^*|_q \le \frac{5.4^{2/q} s^{1/q}}{\kappa^2(s,3)(n-1)^{\alpha(1-\epsilon)/(2\alpha+2p)}},\ \textit{for every } 1 \le q \le 2 \Big) \\
\ge 1 - 2n' \exp\Big(- C_1 c_h^p (n-1)^{(2\alpha+p)/(2\alpha+2p)} \Big) \\
- 2n' \exp\Big(- \frac{c_h^{2p} (n-1)^{2\alpha\epsilon/(2p+2\alpha)}}{C_2 + C_3 (n-1)^{-\alpha(1-\epsilon)/(2\alpha+2p)}} \Big) \\
- 2 \exp\Big(- \frac{c_h^p (n-1)^{(2\alpha+p)/(2\alpha+2p)} (f_{\mathbf{Z},max} - C_{\tilde K,2} h^2)^2}{C_4 + 4 C_{\tilde K} (f_{\mathbf{Z},min} - C_{\tilde K,2} h^2)/3} \Big).
\end{aligned}$$

6.3 Asymptotic behavior of $\hat\beta$

6.3.1 Asymptotic properties of $\hat\beta$ when $n \to \infty$ and for fixed n'

In this part, n' is still supposed to be fixed and we state the consistency and the asymptotic normality of $\hat\beta$ as $n \to \infty$. As above, we adopt a fixed design: the $\mathbf{z}_i'$ are arbitrarily fixed or, equivalently, our reasonings are made conditionally on the second sample.

For $n, n' > 0$, denote by $\hat\beta_{n,n'}$ the estimator (6.2) with $h = h_n$ and $\lambda = \lambda_{n,n'}$. The following lemma, proved in Section 6.7.1, provides another representation of this estimator $\hat\beta_{n,n'}$ that will be useful hereafter.

Lemma 6.7. *We have $\hat\beta_{n,n'} = \arg\min_{\beta \in \mathbb{R}^{p'}} \mathbb{G}_{n,n'}(\beta)$, where*

$$\mathbb{G}_{n,n'}(\beta) := \frac{2}{n'} \sum_{i=1}^{n'} \xi_{i,n} \boldsymbol{\psi}(\mathbf{z}_i')^T (\beta^* - \beta) + \frac{1}{n'} \sum_{i=1}^{n'} \big\{ \boldsymbol{\psi}(\mathbf{z}_i')^T (\beta^* - \beta) \big\}^2 + \lambda_{n,n'} |\beta|_1. \tag{6.6}$$

We will invoke a *convexity argument*: "Let g_n and g_∞ be random convex functions taking minimum values at x_n and x_∞, respectively. If all finite dimensional distributions of g_n converge weakly to those of g_∞ and x_∞ is the unique minimum point of g_∞ with probability one, then x_n converges weakly to x_∞" (see [79], e.g).

Theorem 6.8 (Consistency of $\hat\beta$)**.** *Under the assumptions of Lemma 6.23, if n' is fixed and $\lambda = \lambda_{n,n'} \to \lambda_0$, then, given $\mathbf{z}_1', \ldots, \mathbf{z}_{n'}'$ and as n tends to the infinity, $\hat\beta_{n,n'} \xrightarrow{\mathbb{P}} \beta^{**} := \inf_\beta \mathbb{G}_{\infty,n'}(\beta)$, where $\mathbb{G}_{\infty,n'}(\beta) := \sum_{i=1}^{n'} \big(\boldsymbol{\psi}(\mathbf{z}_i')^T (\beta^* - \beta) \big)^2 / n' + \lambda_0 |\beta|_1$. In particular, if $\lambda_0 = 0$ and $< \boldsymbol{\psi}(\mathbf{z}_1'), \ldots, \boldsymbol{\psi}(\mathbf{z}_{n'}') > = \mathbb{R}^{p'}$, then $\hat\beta_{n,n'} \xrightarrow{\mathbb{P}} \beta^*$.*

Proof : By Lemma 6.23, the first term in the r.h.s. of (6.6) converges to 0 as $n \to \infty$. The third term in the r.h.s. of (6.6) converges to $\lambda_0|\beta|_1$ by assumption. We have just proven that $\mathbb{G}_{n,n'} \to \mathbb{G}_{\infty,n'}$ pointwise as $n \to \infty$. We can now apply the convexity argument, because $\mathbb{G}_{n,n'}$ and $\mathbb{G}_{\infty,n'}$ are convex functions. As a consequence, $\arg\min_\beta \mathbb{G}_{n,n'}(\beta) \to \arg\min_\beta \mathbb{G}_{\infty,n'}(\beta)$ in law. Since we have adopted a fixed design setting, β^{**} is non random, given $(\mathbf{Z}'_1, \dots, \mathbf{Z}'_{n'})$. The convergence in law towards a deterministic quantity implies convergence in probability, which concludes the proof. Moreover, when $\lambda_0 = 0$, β^* is the minimum of $\mathbb{G}_{\infty,n'}$ because the vectors $\boldsymbol{\psi}(\mathbf{z}'_i)$, $i = 1, \dots, p'$ generate the space $\mathbb{R}^{p'}$. Therefore, this implies the consistency of $\hat{\beta}_{n,n'}$. $\square$

To evaluate the limiting behavior of $\hat{\beta}_{n,n'}$, we need the joint asymptotic normality of $(\xi_{1,n}, \dots, \xi_{n',n})$, when $n \to \infty$ and given $\mathbf{z}'_1, \dots, \mathbf{z}'_{n'}$. By applying the Delta-method to the function $\Lambda(\,\cdot\,)$ component-wise, this is given by the following corollary of Lemma 6.24.

Corollary 6.9. *Under the assumptions of Lemma 6.24,* $(nh_n^p)^{1/2}\left[\xi_{1,n}, \dots, \xi_{n',n}\right]^T$ *tends in law towards a random vector* $\mathcal{N}\big(0, \tilde{\mathbb{H}}\big)$ *given* $(\mathbf{z}'_1, \dots, \mathbf{z}'_{n'})$*, where* $\tilde{\mathbb{H}}$ *is a* $n' \times n'$ *real matrix defined, for every integers* $1 \le i, j \le n'$*, by*

$$[\tilde{\mathbb{H}}]_{i,j} := \frac{4\int K^2 \mathbb{1}_{\{\mathbf{z}'_i = \mathbf{z}'_j\}}}{f_{\mathbf{Z}}(\mathbf{z}'_i)} \Big(\Lambda'\big(\tau_{1,2|\mathbf{Z}=\mathbf{z}'_i}\big)\Big)^2 \times \Big\{\mathbb{E}[\tilde{g}(\mathbf{X}_1, \mathbf{X})\tilde{g}(\mathbf{X}_2, \mathbf{X})|\mathbf{Z} = \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}'_i] - \tau^2_{1,2|\mathbf{Z}=\mathbf{z}'_i}\Big\},$$

where $\tilde{g}$ *is the symmetrized version* $\tilde{g}(\mathbf{x}_1, \mathbf{x}_2) := (g^*(\mathbf{x}_1, \mathbf{x}_2) + g^*(\mathbf{x}_2, \mathbf{x}_1))/2$*.*

Theorem 6.10 (Asymptotic law of the estimator)**.** *Under the assumptions of Lemma 6.24, and if* $\lambda_{n,n'}(nh^p_{n,n'})^{1/2}$ *tends to* ℓ *when* $n \to \infty$*, we have* $(nh^p_{n,n'})^{1/2}(\hat{\beta}_{n,n'} - \beta^*) \xrightarrow{D} \mathbf{u}^* := \arg\min_{\mathbf{u}\in\mathbb{R}^{p'}} \mathbb{F}_{\infty,n'}(\mathbf{u})$*, given* $\mathbf{z}'_1, \dots, \mathbf{z}'_{n'}$*, where*

$$\mathbb{F}_{\infty,n'}(\mathbf{u}) := \frac{2}{n'}\sum_{i=1}^{n'}\sum_{j=1}^{p'} W_i\psi_j(\mathbf{z}'_i)u_j + \frac{1}{n'}\sum_{i=1}^{n'}\big(\boldsymbol{\psi}(\mathbf{z}'_i)^T\mathbf{u}\big)^2 + \ell\sum_{i=1}^{p'}\big(|u_i|\mathbb{1}_{\{\beta^*_i=0\}} + u_i\,\mathrm{sign}(\beta^*_i)\mathbb{1}_{\{\beta^*_i\neq 0\}}\big),$$

with $\mathbf{W} = (W_1, \dots, W_{n'}) \sim \mathcal{N}\left(0, \tilde{\mathbb{H}}\right)$.

This theorem is proved in Section 6.7.2. When $\ell = 0$, we can say more about the limiting law in general. Indeed, in such a case, $\mathbf{u}^* = \arg\min_{\mathbf{u}\in\mathbb{R}^{p'}} \mathbb{F}_{\infty,n'}(\mathbf{u})$ is the solution of the first order conditions $\nabla\mathbb{F}_{\infty,n'}(\mathbf{u}) = 0$, that are written as $\sum_{i=1}^{n'} W_i\boldsymbol{\psi}(\mathbf{z}'_i) + \sum_{i=1}^{n'}\boldsymbol{\psi}(\mathbf{z}'_i)\boldsymbol{\psi}(\mathbf{z}'_i)^T\mathbf{u} = 0$. Therefore,

$$\mathbf{u}^* = -\Big(\sum_{i=1}^{n'}\boldsymbol{\psi}(\mathbf{z}'_i)\boldsymbol{\psi}(\mathbf{z}'_i)^T\Big)^{-1}\sum_{i=1}^{n'} W_i\boldsymbol{\psi}(\mathbf{z}'_i),$$

when $\Sigma_{n'} := \sum_{i=1}^{n'}\boldsymbol{\psi}(\mathbf{z}'_i)\boldsymbol{\psi}(\mathbf{z}'_i)^T$ is invertible. Then, the limiting law of $(nh^p_{n,n'})^{1/2}(\hat{\beta}_{n,n'} - \beta^*)$ is Gaussian, and its asymptotic covariance is $V_{as} := \Sigma^{-1}_{n'}\sum_{i,j=1}^{n'}[\tilde{\mathbb{H}}]_{i,j}\boldsymbol{\psi}(\mathbf{z}'_i)\boldsymbol{\psi}(\mathbf{z}'_j)^T\Sigma^{-1}_{n'}$.

The previous results on the asymptotic normality of $\hat{\beta}_{n,n'} - \beta^*$ can be used to test $\mathcal{H}_0 : \beta^* = 0$ against the opposite. As said in the introduction, this would constitute a test of the "simplifying assumption", i.e. the fact that the conditional copula of (X_1, X_2) given $\mathbf{Z}$ does not depend on this covariate. Some tests of significance of β^* would be significantly simpler than most of the tests of the simplifying assumption that have been proposed in the literature until now. Indeed, the latter ones have been built on nonparametric estimates of conditional copulas and, as sub-products of the weak convergence of the associated processes, the test statistics behaviors are obtained. Therefore, such statistics depend on a preliminary non-parametric estimation of conditional marginal distributions (see [141], [38], e.g.), a source of complexities and statistical noise. At the opposite, some tests of $\mathcal{H}_0$ based on $\hat{\beta}_{n,n'}$ do not require this

stage, at the cost of a (probably small) loss of power. For instance, in the case of $\ell = 0$, we propose the Wald-type test statistics

$$\mathcal{W}_n := nh_{n,n'}^p(\hat{\beta}_{n,n'} - \beta^*)^T V_n(\hat{\beta}_{n,n'} - \beta^*),\ V_n := \Sigma_{n'}^{-1} \sum_{i,j=1}^{n'} \hat{\mathbb{H}}_{i,j} \boldsymbol{\psi}(\mathbf{z}_i')\boldsymbol{\psi}(\mathbf{z}_j')^T \Sigma_{n'}^{-1}.$$

$$\hat{\mathbb{H}}_{i,j} := \frac{4\int K^2 \mathbb{1}_{\{\mathbf{z}_i' = \mathbf{z}_j'\}}}{\hat{f}_{\mathbf{Z}}(\mathbf{z}_i')} \Big(\Lambda'\big(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}_i'}\big)\Big)^2 \times \Big\{\mathcal{G}_n(\mathbf{z}_i') - \hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}_i'}^2\Big\},$$

where $\hat{f}_{\mathbf{Z}}(\mathbf{Z})$ and $\mathcal{G}_n(\mathbf{z})$ denote consistent estimators of $f_{\mathbf{Z}}(\mathbf{z})$ and $\mathbb{E}[\tilde{g}(\mathbf{X}_1, \mathbf{X})\tilde{g}(\mathbf{X}_2, \mathbf{X})|\mathbf{Z} = \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}]$ respectively. Under $\mathcal{H}_0$, $\mathcal{W}_n$ tends to a chi-square distribution with n' degrees of freedom. For instance, with the notation of Section 6.1, we propose

$$\mathcal{G}_n(\mathbf{z}) = \sum_{i,j,k=1, i\neq j\neq k}^{n} w_{i,n}(\mathbf{z})w_{j,n}(\mathbf{z})w_{k,n}(\mathbf{z})\tilde{g}(\mathbf{X}_i, \mathbf{X}_k)\tilde{g}(\mathbf{X}_j, \mathbf{X}_k).$$

Note that if there is an intercept, i.e. if one of the functions in $\boldsymbol{\psi}$ (say, ψ_1) is constant to 1, it should be removed in the statistics above. The corresponding coefficients of $\hat{\beta}$ should be removed as well. Indeed, in this case the simplifying assumption does not correspond to $\beta^* = 0$, but rather to $\beta_{-1}^* = 0$ where β_{-i}^* denotes the vector β^* where the i-th coefficient has been removed.

6.3.2 Oracle property and a related adaptive procedure

Let remember that $\mathcal{S} := \{j : \beta_j^* \neq 0\}$ and assume that $|\mathcal{S}| = s < p$ so that the true model depends on a subset of predictors. In the same spirit as [48], we say that an estimator $\hat{\beta}$ satisfies the oracle property if

- $v_n(\hat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}^*)$ converges in law towards a continuous random vector, for some conveniently chosen rate of convergence (v_n), and
- we identify the nonzero components of the true parameter β^* with probability one when the sample size n is large, i.e. the probability of the event $\big(\{j : \hat{\beta}_j \neq 0\} = \mathcal{S}\big)$ tends to one.

As above, let us fix n' and n will tend to the infinity. Then, denote $\{j : \hat{\beta}_j \neq 0\}$ by $\mathcal{S}_n$, that will implicitly depend on n'. It is well-known that the usual Lasso estimator does not fulfill the oracle property, see [146]. Here, this is still the case. The following proposition is proved in Section 6.7.3.

Proposition 6.11. *Under the assumptions of Theorem 6.10,* $\limsup_n \mathbb{P}\left(\mathcal{S}_n = \mathcal{S}\right) = c < 1$.

A usual way of obtaining the oracle property is to modify our estimator in an "adaptive" way. Following [146], consider a preliminary "rough" estimator of β^*, denoted by $\tilde{\beta}_n$, or more simply $\tilde{\beta}$. Moreover $\nu_n(\tilde{\beta}_n - \beta^*)$ is assumed to be asymptotically normal, for some deterministic sequence (ν_n) that tends to the infinity. Now, let us consider the same optimization program as in (6.2) but with a random tuning parameter given by $\lambda_{n,n'} := \mu_{n,n'}/|\tilde{\beta}_n|^\delta$, for some constant $\delta > 0$ and some positive deterministic sequence $(\mu_{n,n'})$. The corresponding adaptive estimator (solution of the modified Equation (6.2)) will be denoted by $\check{\beta}_{n,n'}$, or simply $\check{\beta}$. Hereafter, we still set $\mathcal{S}_n = \{j : \check{\beta}_j \neq 0\}$. The following theorem is proved in Section 6.7.4.

Theorem 6.12 (Asymptotic law of the adaptive estimator of β)**.** *Under the assumptions of Lemma 6.24, if* $\mu_{n,n'}(nh_{n,n'}^p)^{1/2} \to \ell \geq 0$ *and* $\mu_{n,n'}(nh_{n,n'}^p)^{1/2}\nu_n^\delta \to \infty$ *when* $n \to \infty$*, we have*

$$(nh_{n,n'}^p)^{1/2}(\check{\beta}_{n,n'} - \beta^*)_{\mathcal{S}} \xrightarrow{D} \mathbf{u}_{\mathcal{S}}^{**} := \underset{\mathbf{u}_{\mathcal{S}} \in \mathbb{R}^s}{\arg\min}\, \mathbb{F}_{\infty,n'}(\mathbf{u}_{\mathcal{S}}),\ \textit{where}$$

$$\check{\mathbb{F}}_{\infty,n'}(\mathbf{u}_{\mathcal{S}}) := \frac{2}{n'}\sum_{i=1}^{n'}\sum_{j\in\mathcal{S}} W_i\psi_j(\mathbf{z}'_i)u_j + \frac{1}{n'}\sum_{i=1}^{n'}\Big(\sum_{j\in\mathcal{S}}\psi_j(\mathbf{z}'_i)u_j\Big)^2 + \ell\sum_{i\in\mathcal{S}}\frac{u_i}{|\beta^*_i|^\delta}\,\mathrm{sign}(\beta^*_i),$$

with $\mathbf{W} = (W_1,\ldots,W_{n'}) \sim \mathcal{N}\big(0,\tilde{\mathbb{H}}\big)$. *Moreover, when* $\ell = 0$*, the oracle property is fulfilled:* $\mathbb{P}\left(\mathcal{S}_n = \mathcal{S}\right) \underset{n}{\to} 1$.

6.3.3 Asymptotic properties of $\hat{\beta}$ when n and n' jointly tend to $+\infty$

Now, we consider a framework in which both n and n' are going to the infinity, while the dimensions p and p' stay fixed. To be specific, n and n' will not be allowed to independently go to the infinity. In particular, for a given n, the other size $n'(n)$ (simply denoted as n') will be constrained, as detailed in the assumptions below. In this section, we still work conditionally on $\mathbf{z}'_1,\ldots,\mathbf{z}'_{n'},\ldots$. The latter vectors are considered as "fixed", inducing a deterministic sequence. Alternatively, we could consider randomly drawn $\mathbf{z}'_i$ from a given law. The latter case can easily been stated from the results below but its specific statement is left to the reader.

Theorem 6.13 (Consistency of $\hat{\beta}_{n,n'}$, jointly in (n,n'))**.** *Assume that Assumptions 6.9.1-6.9.4 and 6.2.1 are satisfied. Assume that* $\sum_{i=1}^{n'}\boldsymbol{\psi}(\mathbf{z}'_i)\boldsymbol{\psi}(\mathbf{z}'_i)^T/n'$ *converges to a matrix* $M_{\psi,\mathbf{z}'}$*, as* $n'\to\infty$*. Assume that* $\lambda_{n,n'}\to\lambda_0$ *and* $n'\exp(-Anh^{2p})\to 0$ *for every* $A>0$*, when* $(n,n')\to\infty$*. Then* $\hat{\beta}_{n,n'}\xrightarrow{\mathbb{P}}\arg\min_{\beta\in\mathbb{R}^{p'}}\mathbb{G}_{\infty,\infty}(\beta)$*, as* $(n,n')\to\infty$*, where* $\mathbb{G}_{\infty,\infty}(\beta) := (\beta^*-\beta)M_{\psi,\mathbf{z}'}(\beta^*-\beta)^T + \lambda_0|\beta|_1$*. Moreover, if* $\lambda_0 = 0$ *and* $M_{\psi,\mathbf{z}'}$ *is invertible, then* $\hat{\beta}_{n,n'}$ *is consistent and tends to the true value* β^**.*

Proof of this theorem is provided in Section 6.7.5. Note that, since the sequence $(\mathbf{z}'_i)$ is deterministic, we just assume the usual convergence of $\sum_{i=1}^{n'}\boldsymbol{\psi}(\mathbf{z}'_i)\boldsymbol{\psi}(\mathbf{z}'_i)^T/n'$ in $\mathbb{R}^{p'^2}$. Moreover, if the "second subset" $(\mathbf{z}'_i)_{i=1,\ldots,n'}$ were a random sample (drawn along the law $\mathbb{P}_{\mathbf{Z}}$), the latter convergence would be understood "in probability". And if $\mathbb{P}_{\mathbf{Z}}$ satisfies the identifiability condition (Proposition 6.1), then $M_{\psi,\mathbf{z}'}$ would be invertible and $\hat{\beta}_{n,n'}\to\beta^*$ in probability. Now, we want to go one step further and derive the asymptotic law of the estimator $\hat{\beta}_{n,n'}$.

Assumption 6.3.1.

(i) The support of the kernel $K(\cdot)$ *is included into* $[-1,1]^p$*. Moreover, for all* n,n' *and every* $(i,j)\in\{1,\ldots,n'\}^2$*,* $i\neq j$*, we have* $|\mathbf{z}'_i-\mathbf{z}'_j|_\infty > 2h_{n,n'}$*.*

(ii) (a) $n'(nh^{p+4\alpha}_{n,n'} + h^{2\alpha}_{n,n'} + (nh^p_{n,n'})^{-1})\to 0$*, (b)* $\lambda_{n,n'}(n'\,n\,h^p_{n,n'})^{1/2}\to 0$*,*
(c) $n\,h^{p+\alpha}_{n,n'}/\ln n'\to\infty$*.*

(iii) The distribution $\mathbb{P}_{\mathbf{z}',n'} := \sum_{i=1}^{n'}\delta_{\mathbf{z}'_i}/n'$ *weakly converges as* $n'\to\infty$*, to a distribution* $\mathbb{P}_{\mathbf{z}',\infty}$ *on* $\mathbb{R}^p$*, with a density* $f_{\mathbf{z}',\infty}$ *with respect to the* p*-dimensional Lebesgue measure.*

(iv) The matrix $V_1 := \int\boldsymbol{\psi}(\mathbf{z}')\boldsymbol{\psi}(\mathbf{z}')^T f_{\mathbf{z}',\infty}(\mathbf{z}')d\mathbf{z}'$ *is non-singular.*

(v) $\Lambda(\cdot)$ *is two times continuously differentiable. Let* $\mathcal{T}$ *be the range of* $\mathbf{z}\mapsto\tau_{1,2|\mathbf{Z}=\mathbf{z}}$*, from* $\mathcal{Z}$ *towards* $[-1,1]$*. On an open neighborhood of* $\mathcal{T}$*, the second derivative of* $\Lambda(\cdot)$ *is bounded by a constant* $C_{\Lambda''}$*.*

Part (i) of the latter assumption forbids the design points $(\mathbf{z}'_i)_{i\geq 1}$ from being too close to each other and too fast, with respect to the rate of convergence $(h_{n,n'})$ to 0. This can be guaranteed by choosing an appropriate design. For example, if $p=1$ and $\mathcal{Z}=[0,1]$, choose the dyadic sequence $1/2,1/4,3/4,1/8,3/8,5/8,7/8,\ldots$

Part (ii) can be ensured by first choosing a slowly growing sequence $n'(n)$, and then by choosing h that would tend to 0 fast enough. Note that a compromise has to be found concerning these two rates.

The sequence $\lambda_{n,n'}$ should be chosen at last, so that (b) is satisfied. Interestingly, it is always possible to choose the asymptotically optimal bandwidth, i.e. $h \propto n^{-1/(2\alpha+p)}$. In this case, we can set $n' = n^a$, with any $a \in]0, 2\alpha/(2\alpha+p)[$ and the constraints are satisfied.

The design points $\mathbf{z}'_i$ are deterministic, similarly to all results in the present paper. For a given n', we can invoke the non-random measure $\mathbb{P}_{\mathbf{z}',n'} := n'^{-1} \sum_{i=1}^{n'} \delta_{z'_i}$. Equivalently, all results can be seen as given conditionally on the sample $(\mathbf{z}'_i)_{i\geq 1}$. In (iii), we impose the weak convergence of $\mathbb{P}_{\mathbf{z}',n'}$ to a measure with density w.r.t. the Lebesgue measure. Intuitively, this means we do not want to observe some design points that would be repeated infinitely often (this would result in a Dirac component in $\mathbb{P}_{\mathbf{z}',\infty}$). An optimal choice of the density $f_{\mathbf{z}',\infty}$ is not an easy task. Indeed, even if we knew exactly the true density $f_{\mathbf{Z}}$, there is no obvious reasons why we should select the $\mathbf{z}'_i$ along $f_{\mathbf{Z}}$ (at least in the limit). If we want a small asymptotic variance $\tilde{V}_{as}$ (see below), the distribution of the design should concentrate the $\mathbf{z}'_i$ in the regions where $\Lambda'\big(\tau_{1,2|\mathbf{Z}=\mathbf{z}'}\big)^2$ is small and where $\boldsymbol{\psi}(\mathbf{z}')\boldsymbol{\psi}(\mathbf{z}')^T$ is big.

Part (iv) of the assumption is usual, and ensure that the design is somehow "asymptotically full rank". This matrix V_1 will also appear in the asymptotic variance of $\hat{\beta}_{n,n'}$.

Part (v) allow us to control a remainder term in a Taylor expansion of Λ. Notice that this technical assumption was not necessary in the previous section, where we used the Delta-method on the vector $(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i} - \tau_{1,2|\mathbf{Z}=\mathbf{z}'_i})_{i=1,\dots,n'}$. But when the number of terms n' tends to infinity, we have to invoke second derivatives to control remainder terms.

The proof of the next theorem is provided in Section 6.8.

Theorem 6.14 (Asymptotic law of $\hat{\beta}_{n,n'}$, jointly in (n, n'))**.** *Under Assumptions 6.3.1 and 6.9.1-6.9.4, we have*

$$(nn'h^p_{n,n'})^{1/2}(\hat{\beta}_{n,n'} - \beta^*) \xrightarrow{D} \mathcal{N}(0, \tilde{V}_{as}),$$

where $\tilde{V}_{as} := V_1^{-1}V_2V_1^{-1}$*,* V_1 *is the matrix defined in Assumption 6.3.1(iv), and*

$$V_2 := \int K^2 \int (\tilde{g}(\mathbf{x}_1, \mathbf{x}_3)\tilde{g}(\mathbf{x}_2, \mathbf{x}_3) - \tau_{1,2|\mathbf{z}'_1=\mathbf{z}'_2=\mathbf{z}})\Lambda'\big(\tau_{1,2|\mathbf{Z}=\mathbf{z}}\big)^2 \boldsymbol{\psi}(\mathbf{z})\boldsymbol{\psi}(\mathbf{z})^T,$$
$$\times f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_1|\mathbf{Z}=\mathbf{z}) f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_2|\mathbf{Z}=\mathbf{z}) f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_3|\mathbf{Z}=\mathbf{z}) \frac{f_{\mathbf{z}',\infty}(\mathbf{z})}{f_{\mathbf{Z}}(\mathbf{z})}\, d\mathbf{x}_1\, d\mathbf{x}_2\, d\mathbf{x}_3\, d\mathbf{z}'.$$

6.4 Simulations

6.4.1 Numerical complexity

Let us take a short numerical application to compare the complexity of our new estimator with the kernel-based ones. Assume that the size of our dataset is $n = 1.000$, with a fixed small p, and $p' = 100$. We want to estimate the conditional Kendall's tau on $m = 10.000$ given points $\mathbf{z}_1, \dots, \mathbf{z}_m$. Using simple kernel-based estimation, the total number of operations is of the order of $n^2 \times m = 1.000^2 \times 10.000 = 10^{10}$. On the contrary, using our new parametric estimators, the cost can be decomposed in the following way:

1. We choose the design points $\mathbf{z}'_1, \dots, \mathbf{z}'_{n'}$ (say, equi-spaced) with $n' = 100$.
2. We estimate the kernel-based estimator on these n' points (cost: $n^2 \times n' = 1.000^2 \times 100 = 10^8$).
3. We run the Lasso optimization, which is a convex program, so its computation time is linear in n' and p' (cost: $n' \times p' = 100 \times 100 = 10^4$).

4. Finally, for each $\mathbf{z}_i$, we compute the prediction $\Lambda^{(-1)}\big(\hat{\beta}^T\mathbf{z}_i\big)$, and let us assume that $s = 50$ (cost: $m \times s = 10.000 \times 50 = 5 \times 10^5$).

Summing up, the computational cost of this realistic experiment is around 10^8, which is 100 times faster than the kernel-based estimator. Moreover, each new point $\mathbf{z}_{m+1}$ will result in a marginal supplementary cost of 50 operations, compared with a marginal cost of $n^2 = 1.000^2 = 10^6$ for the kernel-based estimator. Such a huge difference is due to the fact that we have transformed what was previously available as U-statistic of order 2 with a $O(n^2)$ computational cost for each prediction, into a linear parametric model with s non-zero parameters, giving a cost of $O(s)$ operations for each prediction.

6.4.2 Choice of tuning parameters and estimation of the components of β

Now, we evaluate the numerical performance of our estimates through a simulation study. In this subsection, we have chosen $n = 3000$, $n' = 100$ and $p = 1$. The univariate covariate Z follows a uniform distribution between 0 and 1. The marginals $X_1|Z = z$ and $X_2|Z = z$ follow some Gaussian distributions $\mathcal{N}(z, 1)$. The conditional copula of $(X_1, X_2)|Z = z$ belongs to the Gaussian copula family. Therefore, it will be parameterized by its (conditional) Kendall's tau $\tau_{1,2|Z=z}$, and is denoted by $C_{\tau_{1,2|Z=z}}$. Obviously, $\tau_{1,2|Z=z}$ is given by Model (6.1). The dependence between X_1 and X_2, given $Z = z$, is specified by $\tau_{1,2|Z=z} := 3z(1-z) = 3/4 - (3/4)(z-1/2)^2$.

We will choose Λ as the identity function and the $\mathbf{z}'_i$ as a uniform grid on $[0.01, 0.99]$. The values 0 and 1 for the $\mathbf{z}'_i$ are excluded to avoid boundaries numerical problems. As for regressors, we will consider $p' = 12$ functions of Z, namely $\psi_1(z) = 1$, $\psi_{i+1}(z) = 2^{-i}(z-0.5)^i$ for $i = 1, \ldots, 5$, $\psi_{5+2i}(z) = \cos(2i\pi z)$ and $\psi_{6+2i}(z) = \sin(2i\pi z)$ for $i = 1, 2$, $\psi_{11}(z) = \mathbb{1}\{z \leq 0.4\}$, $\psi_{12}(z) = \mathbb{1}\{z \leq 0.6\}$. They cover a mix of polynomial, trigonometric and step-functions. Then, the true parameter is $\beta^* = (3/4, 0, -3/4, \mathbf{0}_9)$, where $\mathbf{0}_9$ is the null vector of size 9.

Our reference value of the tuning parameter h is given by the usual rule-of-thumb, i.e. $h = \hat{\sigma}(Z)n^{-1/5}$, where $\hat{\sigma}$ is the estimated standard deviation of Z. Data-driven choices of the bandwidth h of the first estimator are presented in [40]. Moreover, we designed a cross validation procedure (see Algorithm 6) whose output is a data-driven choice for the tuning parameter $\hat{\lambda}^{cv}$. Finally, we perform the convex optimization of the Lasso criterion using the R package `glmnet` by [55].

Algorithm 6: Cross-validation algorithm for choosing λ.

Divide the dataset $\mathcal{D} = (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\ldots,n}$ into N disjoint blocks $\mathcal{D}_1, \ldots, \mathcal{D}_N$;
foreach λ **do**
  for $k \leftarrow 1$ **to** N **do**
    Estimate the conditional Kendall's taus $\big(\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}'_i}\big)_{i=1,\ldots,n'}$ on the dataset $\mathcal{D}_k$;
    Estimate $\hat{\beta}^{(-k)}$ by Equation (6.2) on the dataset $\mathcal{D}\backslash\mathcal{D}_k$ using the tuning parameter λ ;
    Compute $Err_k(\lambda) := \sum_{i=1,\ldots,n'} \Big(\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}'_i} - \boldsymbol{\psi}(\mathbf{z}'_i)^T\hat{\beta}^{(-k)}\Big)^2$;
  end
end
Return $\hat{\lambda}^{cv} := \arg\min_\lambda \sum_k Err_k(\lambda)$.

In our simulations, we observed that the estimation of $\hat{\beta}$ is not very satisfying if the family of function ψ_i is far too large. Indeed, our model will "learn the noise" produced by the kernel estimation, and there

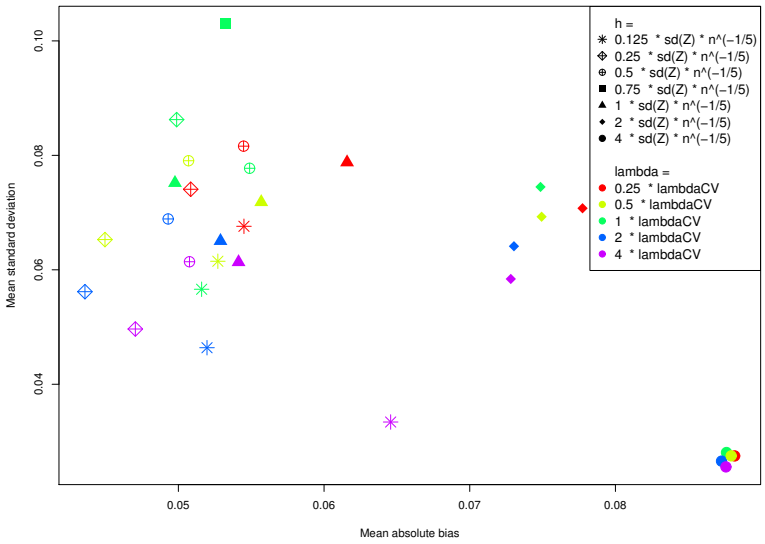


Figure 6.1: Mean absolute bias $\sum_{i=1}^{12} |\mathbb{E}[\hat{\beta}_i] - \beta_i^*|/12$ and mean standard deviation $\sum_{i=1}^{12} \sigma(\hat{\beta}_i)/12$, for different data-driven choices of the tuning parameters h and λ.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$
True value	0.75	0	-0.75	0	0	0	0	0	0	0	0	0
Bias	-0.13	3.6e-05	0.26	0.0033	-0.045	-0.0051	-0.011	-2e-04	-3.2e-05	0.073	-0.0013	0.00021
Std. dev.	0.15	0.00041	0.18	0.035	0.078	0.041	0.022	0.0051	0.00037	0.15	0.007	0.0041
Prob.	1	0.015	0.96	0.015	0.4	0.069	0.36	0.076	0.0076	0.33	0.038	0.023

Table 6.1: Estimated bias, standard deviation and probability of being non-null for each estimated component of β ($h = 0.25\,\hat{\sigma}(Z)n^{-1/5}$ and $\lambda = 2\hat{\lambda}^{cv}$).

will be "overfitting" in the sense that the function $\Lambda^{(-1)}\big(\boldsymbol{\psi}(\cdot)^T\hat{\beta}\big)$ will be very close to $\hat{\tau}_{1,2|\mathbf{Z}=\cdot}$, but not to the target $\tau_{1,2|\mathbf{Z}=\cdot}$. Therefore, we have to find a compromise between misspecification (to choose a family of ψ_i that is not rich enough), and over-fitting (to choose a family of ψ_i that is too rich).

We have led 100 simulations for couples of tuning parameters (λ, h), where $\lambda \propto \hat{\lambda}^{cv}$, and $h \propto \hat{\sigma}(Z)n^{-1/5}$. The results in term of empirical bias and standard deviation of $\hat{\beta}$ are displayed in Figure 6.1. Empirically, we find the smallest h tend to perform better than the largest ones. The influence of the tuning parameter λ (around reasonable values) is less clear. Finally, we selected $h = 0.25\hat{\sigma}(Z)n^{-1/5}$ and $\lambda = 2\hat{\lambda}^{cv}$. With the latter choice, the coefficient by coefficient results are provided in Table 6.1. The empirical results are relatively satisfying, despite a small amount of over-fitting. In particular, the estimation procedure is able to identify the non-zero coefficients almost systematically. To give a complete picture, for one particular simulated sample, we show the results of the estimation procedure, as displayed in Figures 6.4 and 6.5 in Section 6.10.

6.4.3 Comparison between parametric and nonparametric estimators of the conditional Kendall's tau

We will now compare our estimator of the conditional Kendall's tau, i.e. $\mathbf{z} \mapsto \Lambda^{(-1)}\big(\boldsymbol{\psi}(\mathbf{z})^T\hat{\beta}\big)$ with the kernel-based estimator, i.e. the first-step estimator. For this, we will consider six different settings:

1. as previously, a Gaussian copula parameterized by its conditional Kendall's tau, given by $\tau_{1,2|Z=z} :=$

$3z(1-z) = 3/4 - (3/4)(z-1/2)^2$ (well-specified model) ;

2. a badly-specified model, with a Frank copula whose parameter is given by $\theta(z) = \tan(\pi z/2)$. Note that the parameter θ of the Frank family belongs to $\mathbb{R}\backslash\{0\}$ and that its Kendall's tau is not written in terms of standard functions of its parameter θ, see [106, p.171] ;

3. an intermediate model with a Frank copula calibrated to have the same conditional Kendall's tau as in the first setting ;

4. another intermediate model with a Gaussian copula calibrated to have the same conditional Kendall's tau as in the second setting ;

5. a Gaussian copula with a conditional Kendall's tau constant equal to 0.5 ;

6. a Frank copula with a conditional Kendall's tau constant equal to 0.5.

This setting will allows to see the effect of good/bad specifications and of changes in terms of copula families. In Table 6.2, for each setting, we provide five numerical measures of performance of a given estimator:

- the integrated bias: $IBias := \int_z \big(\mathbb{E}[\hat{\tau}_{1,2|Z=z}] - \tau_{1,2|Z=z}\big)dz$;
- the integrated variance: $IVar := \int_z \mathbb{E}\Big[\big(\hat{\tau}_{1,2|Z=z} - \mathbb{E}[\hat{\tau}_{1,2|Z=z}]\big)^2\Big]dz$;
- the integrated standard deviation: $ISd := \int_z \mathbb{E}\Big[\big(\hat{\tau}_{1,2|Z=z} - \mathbb{E}[\hat{\tau}_{1,2|Z=z}]\big)^2\Big]^{1/2}dz$;
- the integrated mean square-error: $IMSE := \int_z \mathbb{E}\Big[\big(\hat{\tau}_{1,2|Z=z} - \tau_{1,2|Z=z}\big)^2\Big]dz$;
- the CPU time used for the computation.

Note that integrals have been approximately computed using a discrete grid $\{0.0005 \times i, i = 0, \ldots, 2000\}$. Globally, in terms of IMSE, the parametric estimator of $\tau_{1,2|\mathbf{z}}$ is doing a better work than a kernel estimator almost systematically (with the single exception of setting 3) and not only in terms of computation time. Surprisingly, even under mis-specification, this conclusion applies whatever the sample size. The differences are particularly striking when the conditional Kendall's tau is a constant function (i.e. under the simplifying assumption).

6.4.4 Comparison with the tests of the simplifying assumption

Now, under the six previous settings, we compare the test of the simplifying assumption $\mathcal{H}_0$ developed in Section 6.3.1 with some of the bootstrapped-based tests of the latter assumption that has been introduced in [38]. In particular, they propose a nonparametric test, using the statistic $\mathcal{T}^0_{CvM}$ defined by

$$\mathcal{T}^0_{CvM} := \int_{[0,1]^3} \Big(\hat{C}_{1,2|Z=\hat{F}_Z^{-1}(u_3)}(u_1,u_2) - \hat{C}_{s,1,2|Z}(u_1,u_2)\Big)^2 du_1 du_2 du_3,$$

where $\hat{C}_{1,2|Z=z}$ is a kernel-based nonparametric estimator of the conditional copula of $(X_1, X_2)|Z = z$ and $\hat{C}_{s,1,2|Z}(u_1,u_2) := n^{-1}\sum_{i=1}^n \hat{C}_{1,2|Z=Z_i}(u_1,u_2)$. We will also invoke their parametric test statistic

$$\mathcal{T}^c_2 := \int_0^1 \Big(\hat{\theta}\big(\hat{F}_Z^{-1}(u)\big) - \hat{\theta}\Big)^2 du,$$

	Kernel-based estimator						Two-step estimator with $n' = 100$ points					
Setting	1	2	3	4	5	6	1	2	3	4	5	6
						$n = 500$						
$IBias$	-29.3	-14.9	-31.5	-6.35	-32.2	-29.9	-23.9	-19.5	-26	-10.5	-31.6	-29.9
$IVar$	17.4	26.4	16.9	26.2	18.5	16.8	27	17.1	28	16.8	1.9	1.65
ISd	123	158	120	157	132	126	43.3	62.5	43.8	56.4	29.7	26.6
$IMSE$	17.4	26.5	16.9	26.4	18.5	16.8	27	17.1	28	16.9	1.91	1.65
CPU time (s)	4.63	5.83	4.62	4.85	4.74	4.9	1.47	1.72	1.42	1.45	1.52	1.54
						$n = 1000$						
$IBias$	-16.6	-11.6	-15.8	-2.97	-16.6	-17.7	-12.6	-12.3	-12.3	-5.42	-16.6	-17.6
$IVar$	8.92	17.3	8.23	13.8	8.82	8.52	8.06	7.59	9.03	6.31	0.622	0.659
ISd	89.2	116	84.5	115	92.2	90.5	30.2	47.8	35.5	43.1	18.2	18.6
$IMSE$	9.01	17.4	8.31	14	8.88	8.57	8.07	7.61	9.04	6.34	0.624	0.661
CPU time (s)	13	12.5	12.8	12.3	12.3	12.7	3.44	3.58	3.73	3.59	3.63	3.68
						$n = 2000$						
$IBias$	-9.94	-4.96	-10	-4.47	-10.7	-10.5	-6.99	-6.55	-7.27	-5.81	-10.6	-10.5
$IVar$	4.76	7.62	4.49	7.81	4.94	4.65	3.09	2.49	3.3	2.44	0.345	0.351
ISd	65.2	85	62.6	86.4	69.4	67.3	22.7	31.4	22.3	32.3	14.7	15.2
$IMSE$	4.77	7.63	4.5	7.83	4.95	4.66	3.09	2.49	3.3	2.44	0.345	0.352
CPU time (s)	67.7	68.6	67.2	73.4	72.3	59.2	15.1	15.1	15.1	16.4	17.9	14.8

Table 6.2: Comparison of the performance between the two estimators. Integrated measures have been multiplied by 10^3, for readability.

where $\hat{\theta}(z)$ estimates the parameter of the Gaussian (resp. Frank) copula given $Z = z$, assuming we know the right family of conditional copula, and $\hat{\theta}$ consistently estimates the parameter of the corresponding simplified copula (under the null). Moreover, $\hat{F}_Z^{-1}$ denotes the empirical quantile function that is associated to the Z-sample. The latter test statistics depends on an a priori chosen parametric copula family. To evaluate the risk of mis-specification, we also include in our table the parametric test $\mathcal{T}_2^c$ assuming that the data come from a Clayton copula, whereas the true copula is Gaussian or Frank. For these three tests, p-values are computed by the usual nonparametric bootstrap, with 100 resampling: see Table 6.3. Globally, the test based on $\mathcal{W}_n$ performs very well under all settings, compared to the alternative nonparametric test. It is only beaten by $\mathcal{T}_2^c$ that is obtained by choosing the right copula family, a not very realistic situation. When it is not the case, $\mathcal{W}_n$ does a better work.

	Not under $\mathcal{H}_0$				Under $\mathcal{H}_0$	
	1	2	3	4	5	6
$\mathcal{W}_n$	88.7	99.8	87.3	100	12	12.1
$\mathcal{T}_{CvM}^0$	59.5	52	64.7	37.5	0	0
$\mathcal{T}_2^c$	100	100	100	100	0.2	2.6
$\mathcal{T}_2^c$ (Clayton)	68	13	100	100	1.8	1.8

Table 6.3: Comparison of the performance between different tests of the simplifying assumption under the six settings of Section 6.4.3, with $n = 500$.

6.4.5 Dimension 2 and choice of $\boldsymbol{\psi}$

In this section, we will fix the sample size $n = 3000$ and the dimension $p = 2$. The random vector $\mathbf{Z}$ will follow a uniform distribution on $[0,1]^2$, $X_1|\mathbf{Z} = \mathbf{z} \sim \mathcal{N}(0, z_1)$, $X_2|\mathbf{Z} = \mathbf{z} \sim \mathcal{N}(0, z_1)$. Given $\mathbf{Z} = \mathbf{z}$, the conditional copula of X_1 and X_2 is Gaussian. We consider three different choices for the functional form of its conditional Kendall's tau :

Setting 1. $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = (3/4) \times (z_1 - z_2)$;

Setting 2. $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = (4/8) \times \cos(2\pi z_1) + (2/8) \times \sin(2\pi z_2)$;

Setting 3. $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = (3/4) \times \tanh(z_1/z_2)$,

where $\mathbf{z} = (z_1, z_2)$. We try different choices of dictionaries $\boldsymbol{\psi}$. For convenience, define $p_0(x) := 1$, $p_i(x) := 2^{-i}(x-0.5)^i$, $trig_0(x) := 1$, and $trig_i(x) := \big(\cos(2i\pi x), \sin(2i\pi x)\big)$, for $x \in \mathbb{R}$ and $i \in \mathbb{N}^*$. We will use the notation $(g_1, g_2) \otimes (g_3, g_4) := (g_1 g_3, g_1 g_4, g_2 g_3, g_2 g_4)$. We are interested in the following functions $\boldsymbol{\psi}$, that are defined for every $\mathbf{z} \in \mathbb{R}^p$ by

$$
\begin{aligned}
\boldsymbol{\psi}^{(1)}(\mathbf{z}) &:= \Big(1, \big(p_i(z_1)\big)_{i=1,\dots,5}, \big(p_i(z_2)\big)_{i=1,\dots,5}\Big) = \Big(p_i(z_1) \times p_j(z_2)\Big)_{\min(i,j)=0,\, \max(i,j)\le 5} \in \mathbb{R}^{11},\\
\boldsymbol{\psi}^{(2)}(\mathbf{z}) &:= \Big(p_i(z_1) \times p_j(z_2)\Big)_{\min(i,j)\le 1,\, \max(i,j)\le 5} \in \mathbb{R}^{20},\\
\boldsymbol{\psi}^{(3)}(\mathbf{z}) &:= \Big(p_i(z_1) \times p_j(z_2)\Big)_{\min(i,j)\le 2,\, \max(i,j)\le 5} \in \mathbb{R}^{27},\\
\boldsymbol{\psi}^{(4)}(\mathbf{z}) &:= \Big(p_i(z_1) \times p_j(z_2)\Big)_{\max(i,j)\le 5} \in \mathbb{R}^{36},\\
\boldsymbol{\psi}^{(5)}(\mathbf{z}) &:= \Big(1, \big(trig_i(z_1)\big)_{i=1,\dots,5}, \big(trig_i(z_2)\big)_{i=1,\dots,5}\Big) \in \mathbb{R}^{21},\\
\boldsymbol{\psi}^{(6)}(\mathbf{z}) &:= \Big(trig_i(z_1) \otimes trig_j(z_2)\Big)_{\min(i,j)\le 1,\, \max(i,j)\le 5} \in \mathbb{R}^{57},\\
\boldsymbol{\psi}^{(7)}(\mathbf{z}) &:= \Big(trig_i(z_1) \otimes trig_j(z_2)\Big)_{\min(i,j)\le 2,\, \max(i,j)\le 5} \in \mathbb{R}^{85},\\
\boldsymbol{\psi}^{(8)}(\mathbf{z}) &:= \Big(trig_i(z_1) \otimes trig_j(z_2)\Big)_{\max(i,j)\le 5} \in \mathbb{R}^{121},\\
\boldsymbol{\psi}^{(9)}(\mathbf{z}) &:= \Big(\boldsymbol{\psi}^{(1)}(\mathbf{z}), \boldsymbol{\psi}^{(5)}(\mathbf{z})\Big) \in \mathbb{R}^{31}, \qquad \boldsymbol{\psi}^{(10)}(\mathbf{z}) := \Big(\boldsymbol{\psi}^{(2)}(\mathbf{z}), \boldsymbol{\psi}^{(6)}(\mathbf{z})\Big) \in \mathbb{R}^{76},\\
\boldsymbol{\psi}^{(11)}(\mathbf{z}) &:= \Big(\boldsymbol{\psi}^{(3)}(\mathbf{z}), \boldsymbol{\psi}^{(7)}(\mathbf{z})\Big) \in \mathbb{R}^{137}, \qquad \boldsymbol{\psi}^{(12)}(\mathbf{z}) := \Big(\boldsymbol{\psi}^{(4)}(\mathbf{z}), \boldsymbol{\psi}^{(8)}(\mathbf{z})\Big) \in \mathbb{R}^{156},
\end{aligned}
$$

where in the last 4 dictionaries, we count the function constant to 1 only once. We choose $n' = 400$ and the design points $\mathbf{z}'_i$ are chosen as an equispaced grid on $[0.1, 0.9]^2$. We consider similar measures of performance for our estimators as in Section 6.4.3. The only difference is that the integration in $\mathbf{z}$ is now done on the unit square $[0,1]^2$. In practice, integrals are discretized, and estimated by a sum over the points $\{(0.01 \times i, 0.01 \times j), 0 \le i, j \le 100\}$. Results are displayed in the following Table 6.4.

We note that the size of the family $\boldsymbol{\psi}$ seems to have a tiny influence on the computation time, which lies always between 6 and 8 seconds. In all settings, polynomial families ($\boldsymbol{\psi}^{(1)}$ to $\boldsymbol{\psi}^{(4)}$) give the best $IMSE$, even when the true function is trigonometric (Setting 2) or under misspecification (Setting 3). Nevertheless, using trigonometric functions can help to reduce the integrated biais and standard deviation. Indeed, in Setting 2, trigonometric families ($\boldsymbol{\psi}^{(5)}$ to $\boldsymbol{\psi}^{(8)}$) do a fair job according to these two measures of performance. Similarly, in Setting 3, mixed families ($\boldsymbol{\psi}^{(9)}$ to $\boldsymbol{\psi}^{(12)}$) achieve an acceptable performance. In Settings 1 and 2, they often yield improvement other a msispecified family, especially in terms of integrated standard deviation.

Comparisons between three indicators $IMSE$, $IBias$ and ISd may be surprising at first sight, but there is no direct link between their values. Indeed, for every point $\mathbf{z}$, $MSE(\mathbf{z}) = Bias(\mathbf{z})^2 + Sd(\mathbf{z})^2$, while $IMSE = \int MSE(\mathbf{z})d\mathbf{z}$, $IBias = \int Bias(\mathbf{z})d\mathbf{z}$ and $ISd = \int Sd(\mathbf{z})d\mathbf{z}$. Therefore, a procedure

	Setting 1				Setting 2				Setting 3			
	IBias	ISd	IMSE	Time	IBias	ISd	IMSE	Time	IBias	ISd	IMSE	Time
$\psi^{(1)}$	0.577	19.4	1.44	6.82	-0.632	24	1.4	6.75	-7.71	17	6.79	6.67
$\psi^{(2)}$	0.309	18.9	1.43	6.77	-0.166	23.7	1.35	6.66	-7.57	16.9	6.8	6.66
$\psi^{(3)}$	0.728	19.9	1.63	6.77	-0.36	27.1	1.9	6.67	-7.63	23.7	3.45	7.06
$\psi^{(4)}$	0.513	18.9	1.81	6.77	-0.245	26.5	2.22	6.68	-7.29	25	2.06	7.52
$\psi^{(5)}$	1.5	25.7	15.7	6.77	0.0616	15	2.67	6.66	-8.38	21.6	14.9	7.51
$\psi^{(6)}$	1.64	26	15.7	6.79	0.269	15	2.61	6.66	-8.23	21.9	14.9	7.52
$\psi^{(7)}$	0.311	26.1	17	6.79	0.0167	15	3.14	6.69	-7.33	23.1	15.1	7.26
$\psi^{(8)}$	1.2	26	17.3	6.88	-0.113	14.6	3.15	6.7	-7.6	22.9	15.3	7.2
$\psi^{(9)}$	0.596	17.7	2.05	6.79	0.492	15.8	2.72	6.67	-7.93	16.3	7.04	7.19
$\psi^{(10)}$	-0.0921	18	2.08	6.77	-0.493	16.6	2.75	6.66	-7.65	16.7	6.94	7.19
$\psi^{(11)}$	0.529	17.3	2.57	6.83	-0.165	15.8	3.08	6.7	-6.87	23	4.76	7.21
$\psi^{(12)}$	0.5	16.9	2.64	6.92	-0.078	16.4	3.24	6.76	-7.07	25.5	4.43	7.54

Table 6.4: Comparison of the estimation using different ψ families. All integrated measures have been multiplied by 1000. Computation time is given in seconds.

that minimize both $Ibias$ and ISd still may not minimize $IMSE$, and conversely. This is due to the non-linearity of the square function, combined with the integration.

6.5 Real data application

Now, we apply the model given by (6.1) to a real dataset. From the website of the World Factbook of the Central Intelligence Agency, we have collected data of male and female life expectancy and GDP per capita for $n = 206$ countries in the world. We seek to analyze the dependence between male and female life expectancies conditionally on the GDP per capita, i.e. given the explanatory variable $Z = \log_{10}(GDP/capita)$. This dataset and these variables are similar as those in the first example studied in [63].

We use $n' = 100$, $h = 2\sigma(Z)n^{-1/5}$ and the same family of functions ψ_i as in Section 6.4.2 above (once composed with a linear transform to be defined on $[\min(Z), \max(Z)]$). The results are displayed in Figure 6.2. As expected, the levels of conditional dependence between male and female expectancies are strong overall. Many poor countries suffer from epidemics, malnutrition or even wars. In such cases, life expectancies of both genders are exposed to the same "exogenous" factors, inducing high Kendall's taus. Logically, we observe a monotonic decrease of such Kendall's taus when Z is larger, up to $Z \simeq 4.5$, as already noticed by [63]. Indeed, when countries become richer, more developed and safe, men and women less and less depend on their environment (and on its risks of death, potentially). Nonetheless, when Z become even larger (the richest countries in the world), conditional dependencies between male and female life expectancies interestingly increase again, because men and women behave similarly in terms of way of life. In particular, they can benefit from the same levels of security and health and are exposed to the same lethal risks.

6.6 Proofs of finite-distance results for $\hat{\beta}$

In this section, we will use the notation $\mathbf{u} := \hat{\beta} - \beta^*$ and $\xi = [\xi_{i,n}]_{i=1,\dots,n'}$, $\xi_{i,n} = Y_i - (\mathbb{Z}'\beta)_i$.

6.6.1 Technical lemmas

Lemma 6.15. *We have* $||\mathbb{Z}'\mathbf{u}||_{n'}^2 \leq \lambda|\mathbf{u}|_1 + \frac{1}{n'}\langle \xi\ ,\ \mathbb{Z}'\mathbf{u}\rangle$.

Proof : As $\hat{\beta}$ is optimal, through the Karush-Kuhn-Tucker conditions, we have $(1/n')\mathbb{Z}'^T(\mathbf{Y} - \mathbb{Z}'\hat{\beta}) \in \partial(\lambda|\hat{\beta}|_1)$, where $\partial(\lambda|\hat{\beta}|_1)$ is the subdifferential of the norm $\lambda|\cdot|_1$ evaluated at $\hat{\beta}$. The dual norm of $|\cdot|_1$ is $|\cdot|_\infty$, so there exists $\mathbf{v}$ such that $|\mathbf{v}|_\infty \leq 1$ and $(1/n')\mathbb{Z}'^T(\mathbf{Y} - \mathbb{Z}'\hat{\beta}) + \lambda\mathbf{v} = 0$. We deduce successively $\mathbb{Z}'^T\mathbb{Z}'(\beta^* - \hat{\beta})/n' + \mathbb{Z}'^T\xi/n' + \lambda\mathbf{v} = 0$,

$$\frac{1}{n'}|\mathbb{Z}'(\beta^* - \hat{\beta})|_2^2 + \frac{1}{n'}(\beta^* - \hat{\beta})^T\mathbb{Z}'^T\xi + \lambda(\beta^* - \hat{\beta})^T\mathbf{v} = 0, \text{ and finally}$$

$$||\mathbb{Z}'(\beta^* - \hat{\beta})||_{n'}^2 \leq \frac{1}{n'}\left\langle \mathbb{Z}'(\hat{\beta} - \beta^*)\ ,\ \xi\right\rangle + \lambda|\beta^* - \hat{\beta}|_1. \ \square$$

Lemma 6.16. *We have* $|\mathbf{u}_{\mathcal{S}^C}|_1 \leq |\mathbf{u}_{\mathcal{S}}|_1 + \frac{2}{\lambda n'}\langle \xi\ ,\ \mathbb{Z}'\mathbf{u}\rangle$.

Proof : By definition, $\hat{\beta}$ is a minimizer of $||\mathbf{Y} - \mathbb{Z}'\beta||_{n'}^2 + \lambda|\beta|_1$. Therefore, we have

$$||\mathbf{Y} - \mathbb{Z}'\hat{\beta}||_{n'}^2 + \lambda|\hat{\beta}|_1 \leq ||\mathbf{Y} - \mathbb{Z}'\beta^*||_{n'}^2 + \lambda|\beta^*|_1.$$

After some algebra, we derive $||\mathbf{Y} - \mathbb{Z}'\hat{\beta}||_{n'}^2 - ||\mathbf{Y} - \mathbb{Z}'\beta^*||_{n'}^2 \leq \lambda\big(|(\beta^* - \hat{\beta})_{\mathcal{S}}|_1 - |(\hat{\beta} - \beta^*)_{\mathcal{S}^C}|_1\big)$. Moreover, the mapping $\beta \mapsto ||\mathbf{Y} - \mathbb{Z}'\beta||_{n'}^2$ is convex and its gradient at β^* is $-2\,\mathbb{Z}'^T(\mathbf{Y} - \mathbb{Z}'\beta^*)/n' = -2\,\mathbb{Z}'^T\xi/n'$. So, we obtain

$$||\mathbf{Y} - \mathbb{Z}'\hat{\beta}||_{n'}^2 - ||\mathbf{Y} - \mathbb{Z}'\beta^*||_{n'}^2 \geq \frac{-2}{n'}\left\langle \mathbb{Z}'^T\xi\ ,\ \hat{\beta} - \beta^*\right\rangle.$$

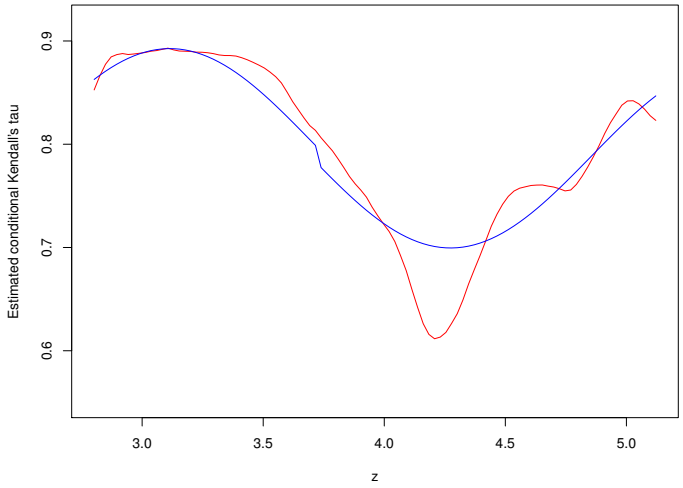


Figure 6.2: Estimated conditional Kendall's tau $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}$ (red curve), and prediction $\Lambda^{(-1)}\big(\boldsymbol{\psi}(\mathbf{z})^T\hat{\beta}\big)$ (blue curve) as a function of $\mathbf{z}$ for the application on real data, where the estimated non-zero coefficients are $\hat{\beta}_1 = 0.78$, $\hat{\beta}_7 = -0.043$, $\hat{\beta}_8 = 0.069$ and $\hat{\beta}_{11} = 0.020$.

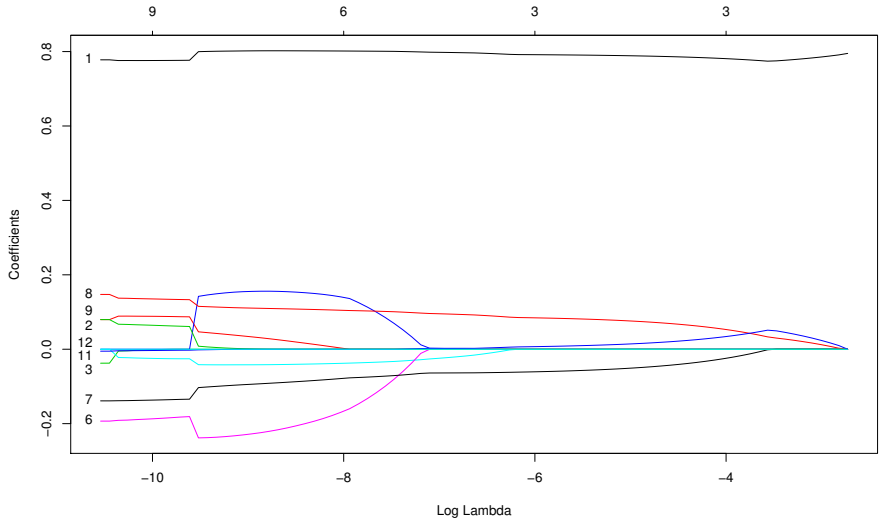


Figure 6.3: Evolution of the estimated non-zero coefficients as a function of the regularization parameter λ for the application on real data. All the other non-displayed ψ_i coefficients are zero.

Combining the two previous equations, we get

$$(-2)\left\langle \mathbb{Z}'^T\xi, \hat{\beta}-\beta^*\right\rangle/n' \leq \lambda\big(|(\beta^*-\hat{\beta})_{\mathcal{S}}|_1 - |(\hat{\beta}-\beta^*)_{\mathcal{S}^c}|_1\big). \ \square$$

Lemma 6.17. *Assume that* $\max_{j=1,\dots,p'}\left|\frac{1}{n'}\sum_{i=1}^{n'} Z'_{i,j}\xi_{i,n}\right| \leq t$, *for some* $t>0$, *that the assumption* $RE(s,3)$ *is satisfied, and that the tuning parameter is given by* $\lambda=\gamma t$, *with* $\gamma\geq 4$. *Then,* $||\mathbb{Z}'(\hat{\beta}-\beta^*)||_{n'} \leq \frac{4(\gamma+1)t\sqrt{s}}{\kappa(s,3)}$ *and* $|\hat{\beta}-\beta^*|_q \leq \frac{4^{2/q}(\gamma+1)ts^{1/q}}{\kappa^2(s,3)}$, *for every* $1\leq q\leq 2$.

Proof : Under the first assumption, we have the upper bound

$$\frac{1}{n'}\left\langle \mathbb{Z}'^T\xi\ ,\ \mathbf{u}\right\rangle \leq |\mathbf{u}|_1 \max_{j=1,\dots,p'}\left|\frac{1}{n'}\sum_{i=1}^{n'} Z'_{i,j}\xi_{i,n}\right| \leq |\mathbf{u}|_1 t.$$

We first show that $\mathbf{u}$ belongs to the cone $\left\{\delta\in\mathbb{R}^{p'} : |\delta_{\mathcal{S}^c}|_1 \leq 3|\delta_{\mathcal{S}}|_1, Card(\mathcal{S})\leq s\right\}$, so that we will be able to use the $RE(s,3)$ assumption with $J_0=\mathcal{S}$. From Lemma 6.16, $|\mathbf{u}_{\mathcal{S}^c}|_1 \leq |\mathbf{u}_{\mathcal{S}}|_1 + 2t|\mathbf{u}|_1/\lambda$. With our choice of λ, we deduce $|\mathbf{u}_{\mathcal{S}^c}|_1 \leq |\mathbf{u}_{\mathcal{S}}|_1 + 2|\mathbf{u}|_1/\gamma$. Using the decomposition $|\mathbf{u}|_1 = |\mathbf{u}_{\mathcal{S}^c}|_1 + |\mathbf{u}_{\mathcal{S}}|_1$, we get $|\mathbf{u}_{\mathcal{S}^c}|_1 \leq |\mathbf{u}_{\mathcal{S}}|_1(\gamma+2)/(\gamma-2) \leq 3|\mathbf{u}_{\mathcal{S}}|_1$. As a consequence, we have

$$|\mathbf{u}|_1 = |\mathbf{u}_{\mathcal{S}^c}|_1 + |\mathbf{u}_{\mathcal{S}}|_1 \leq 4|\mathbf{u}_{\mathcal{S}}|_1 \leq 4\sqrt{s}|\mathbf{u}|_2 \leq 4\sqrt{s}\,||\mathbb{Z}'\mathbf{u}||_{n'}/\kappa(s,3).$$

By Lemma 6.15,

$$||\mathbb{Z}'\mathbf{u}||_{n'}^2 \leq \lambda|\mathbf{u}|_1 + \frac{1}{n'}\left\langle \xi\ ,\ \mathbb{Z}'\mathbf{u}\right\rangle \leq \lambda|\mathbf{u}|_1 + |\mathbf{u}|_1 t \leq |\mathbf{u}|_1(\gamma+1)t \leq \frac{4\sqrt{s}}{\kappa(s,3)}\,||\mathbb{Z}'\mathbf{u}||_{n'}(\gamma+1)t$$

We can now simplify and we get

$$||\mathbb{Z}'\mathbf{u}||_{n'} \leq \frac{4(\gamma+1)t}{\kappa(s,3)}\sqrt{s}, \quad |\mathbf{u}|_2 \leq \frac{4(\gamma+1)t}{\kappa^2(s,3)}\sqrt{s}, \text{ and } |\mathbf{u}|_1 \leq \frac{16(\gamma+1)t}{\kappa^2(s,3)}\,s.$$

Now, we compute a general bound for $|\mathbf{u}|_q$, with $1\leq q\leq 2$, using the Hölder norm interpolation inequality:

$$|\mathbf{u}|_q \leq |\mathbf{u}|_1^{2/q-1}|\mathbf{u}|_2^{2-2/q} \leq \frac{4^{2/q}(\gamma+1)ts^{1/q}}{\kappa^2(s,3)}\cdot \ \square$$

6.6.2 Proof of Theorem 6.5

Using Lemma 6.21, for every $t_1,t_2>0$ such that $C_{K,\alpha}h^\alpha/\alpha! + t_1 \leq f_{\mathbf{Z},min}/2$, with probability greater than $1-2n'\exp\left(-\left(nh^pt_1^2\right)/\left(2f_{\mathbf{Z},max}\int K^2 + (2/3)C_K t_1\right)\right) - 2n'\exp\left(-(n-1)h^{2p}t_2^2 f_{\mathbf{Z},min}^4 \,/\, \left(4f_{\mathbf{Z},max}^2(\int K^2)^2 + (8/3)C_K^2 f_{\mathbf{Z},min}^2 t_2\right)\right) - 2\exp\left(-\left(nh^p(f_{\mathbf{z}}(\mathbf{z}) - C_{\tilde{K},2}h^2)^2\right)/\left(8f_{\mathbf{Z},max}\int \tilde{K}^2 + 4C_{\tilde{K}}(f_{\mathbf{z}}(\mathbf{z}) - C_{\tilde{K},2}h^2)/3\right)\right)$, we have

$$\begin{aligned}\max_{j=1,\dots,p'}\left|\frac{1}{n'}\sum_{i=1}^{n'} Z'_{i,j}\xi_{i,n}\right| &\leq C_\psi \max_{i=1,\dots,n'}\left|\xi_{i,n}\right| \leq C_\psi C_{\Lambda'} \max_{i=1,\dots,n'}\left|\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i} - \tau_{1,2|\mathbf{Z}=\mathbf{z}'_i}\right| \\ &\leq 4C_\psi C_{\Lambda'}\left(1+\frac{16f_{\mathbf{Z},max}^2}{f_{\mathbf{Z},min}^3}\left(\frac{C_{K,\alpha}h^\alpha}{\alpha!}+t_1\right)\right)\left(\frac{C_{\mathbf{XZ},\alpha}h^\alpha}{f_{\mathbf{Z},min}^2\alpha!}+t_2\right).\end{aligned}$$

We choose $t_1 := f_{\mathbf{Z},min}/4$ so that, because of Condition (6.4), we get $C_{K,\alpha}h^\alpha/\alpha! + t_1 \leq f_{\mathbf{Z},min}/2$. Now we choose $t_2 := tf_{\mathbf{Z},min}^2/\{8C_\psi C_{\Lambda'}(f_{\mathbf{Z},min}^2 + 8f_{\mathbf{Z},max}^2)\}$. By Condition (6.4), $C_{\mathbf{XZ},\alpha}h^\alpha/(f_{\mathbf{Z},min}^2\alpha!) \leq t_2$, so that we have

$$4C_\psi C_{\Lambda'}\left(1+\frac{8f_{\mathbf{Z},max}^2}{f_{\mathbf{Z},min}^2}\right)\times\left(\frac{C_{\mathbf{XZ},\alpha}h^\alpha}{f_{\mathbf{Z},min}^2\alpha!}+t_2\right) \leq 8t_2C_\psi C_{\Lambda'}\left(1+\frac{8f_{\mathbf{Z},max}^2}{f_{\mathbf{Z},min}^2}\right) \leq t.$$

As a consequence, we obtain that

$$\begin{aligned}
&\mathbb{P}\left(\max_{j=1,\dots,p'} \Big| \frac{1}{n'} \sum_{i=1}^{n'} Z'_{i,j}\xi_{i,n} \Big| > t + \frac{3\int K^2}{nh^p f_{\mathbf{Z},min}} \right) \\
&\leq 2n' \exp\left(- \frac{nh^p f_{\mathbf{Z},min}^2}{32 f_{\mathbf{Z},max}\int K^2 + (8/3)C_K f_{\mathbf{Z},min}} \right) + 2n' \exp\left(- \frac{(n-1)h^{2p}t^2}{C_2 + C_3 t} \right) \\
&+ 2\exp\left(- \frac{nh^p (f_{\mathbf{Z},max} - C_{\tilde{K},2}h^2)^2}{8 f_{\mathbf{Z},max}\int \tilde{K}^2 + 4C_{\tilde{K}}(f_{\mathbf{Z},min} - C_{\tilde{K},2}h^2)/3} \right),
\end{aligned}$$

and we can apply Lemma 6.17 to get the claimed result. □

6.7 Proofs of asymptotic results for $\hat{\beta}_{n,n'}$

6.7.1 Proof of Lemma 6.7

Using the definition (6.2) of $\hat{\beta}_{n,n'}$, we get

$$\begin{aligned}
\hat{\beta}_{n,n'} &:= \arg\min_{\beta\in\mathbb{R}^{p'}} \frac{1}{n'} \sum_{i=1}^{n'} \big(\Lambda(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i}) - \boldsymbol{\psi}(\mathbf{z}'_i)^T\beta\big)^2 + \lambda_{n,n'}|\beta|_1 \\
&= \arg\min_{\beta\in\mathbb{R}^{p'}} \frac{1}{n'} \sum_{i=1}^{n'} \big(\xi_{i,n} + \boldsymbol{\psi}(\mathbf{z}'_i)^T\beta^* - \boldsymbol{\psi}(\mathbf{z}'_i)^T\beta\big)^2 + \lambda_{n,n'}|\beta|_1 \\
&= \arg\min_{\beta\in\mathbb{R}^{p'}} \frac{1}{n'} \sum_{i=1}^{n'} \xi_{i,n}^2 + \frac{2}{n'}\sum_{i=1}^{n'} \xi_{i,n}\boldsymbol{\psi}(\mathbf{z}'_i)^T(\beta^* - \beta) + \frac{1}{n'}\sum_{i=1}^{n'} \big(\boldsymbol{\psi}(\mathbf{z}'_i)^T(\beta^* - \beta)\big)^2 + \lambda_{n,n'}|\beta|_1 \\
&= \arg\min_{\beta\in\mathbb{R}^{p'}} \frac{2}{n'}\sum_{i=1}^{n'} \xi_{i,n}\boldsymbol{\psi}(\mathbf{z}'_i)^T(\beta^* - \beta) + \frac{1}{n'}\sum_{i=1}^{n'} \big(\boldsymbol{\psi}(\mathbf{z}'_i)^T(\beta^* - \beta)\big)^2 + \lambda_{n,n'}|\beta|_1. \ \square
\end{aligned}$$

6.7.2 Proof of Theorem 6.10

Let us define $r_{n,n'} := (nh^p_{n,n'})^{1/2}$, $\mathbf{u} := r_{n,n'}(\beta - \beta^*)$ and $\hat{\mathbf{u}}_{n,n'} := r_{n,n'}(\hat{\beta}_{n,n'} - \beta^*)$, so that $\hat{\beta}_{n,n'} = \beta^* + \hat{\mathbf{u}}_{n,n'}/r_{n,n'}$. By Lemma 6.7, $\hat{\beta}_{n,n'} = \arg\min_{\beta\in\mathbb{R}^{p'}} \mathbb{G}_{n,n'}(\beta)$. We have therefore

$$\hat{\mathbf{u}}_{n,n'} = \arg\min_{\mathbf{u}\in\mathbb{R}^{p'}} \Big[\frac{-2}{n'}\sum_{i=1}^{n'} \xi_{i,n}\boldsymbol{\psi}(\mathbf{z}'_i)^T\frac{\mathbf{u}}{r_{n,n'}} + \frac{1}{n'}\sum_{i=1}^{n'} \big(\boldsymbol{\psi}(\mathbf{z}'_i)^T\frac{\mathbf{u}}{r_{n,n'}}\big)^2 + \lambda_{n,n'}\big|\beta^* + \frac{\mathbf{u}}{r_{n,n'}}\big|_1 \Big],$$

or $\hat{\mathbf{u}}_{n,n'} = \arg\min_{\mathbf{u}\in\mathbb{R}^{p'}} \mathbb{F}_{n,n'}(\mathbf{u})$, where, for every $\mathbf{u}\in\mathbb{R}^{p'}$,

$$\mathbb{F}_{n,n'}(\mathbf{u}) := \frac{-2r_{n,n'}}{n'}\sum_{i=1}^{n'} \xi_{i,n}\boldsymbol{\psi}(\mathbf{z}'_i)^T\mathbf{u} + \frac{1}{n'}\sum_{i=1}^{n'} \big(\boldsymbol{\psi}(\mathbf{z}'_i)^T\mathbf{u}\big)^2 + \lambda_{n,n'}r_{n,n'}^2\Big(\big|\beta^* + \frac{\mathbf{u}}{r_{n,n'}}\big|_1 - \big|\beta^*\big|_1\Big).$$

Note that, by Corollary 6.9, we have

$$\frac{2r_{n,n'}}{n'}\sum_{i=1}^{n'} \xi_{i,n}\boldsymbol{\psi}(\mathbf{z}'_i)^T\mathbf{u} = \frac{2}{n'}\sum_{i=1}^{n'}\sum_{j=1}^{p'} r_{n,n'}\xi_{i,n}\psi_j(\mathbf{z}'_i)u_j \xrightarrow{D} \frac{2}{n'}\sum_{i=1}^{n'}\sum_{j=1}^{p'} W_i\psi_j(\mathbf{z}'_i)u_j.$$

We also have, for any (fixed) $\mathbf{u}$ and when n is large enough,

$$\big|\beta^* + \frac{\mathbf{u}}{r_{n,n'}}\big|_1 - \big|\beta^*\big|_1 = \sum_{i=1}^{p'}\left(\frac{|u_i|}{r_{n,n'}}\mathbb{1}_{\{\beta_i^*=0\}} + \frac{u_i}{r_{n,n'}}\,\mathrm{sign}(\beta_i^*)\mathbb{1}_{\{\beta_i^*\neq 0\}} \right).$$

Therefore $\lambda_{n,n'} r_{n,n'}^2 \Big(\big|\beta^* + \mathbf{u}/r_{n,n'}\big|_1 - \big|\beta^*\big|_1 \Big) \to \ell \sum_{i=1}^{p'} \big(|u_i| \mathbb{1}_{\{\beta_i^*=0\}} + u_i \operatorname{sign}(\beta_i^*) \mathbb{1}_{\{\beta_i^* \neq 0\}} \big)$.

We have shown that $\mathbb{F}_{n,n'}(\mathbf{u}) \xrightarrow{D} \mathbb{F}_{\infty,n'}(\mathbf{u})$. Those functions are convex, hence the conclusion follows from the convexity argument. □

6.7.3 Proof of Proposition 6.11

The proof closely follows Proposition 1 in [146]. It starts by noting that $\mathbb{P}\left(\mathcal{S}_n = \mathcal{S}\right) \leq \mathbb{P}\left(\hat{\beta}_j = 0,\ \forall j \notin \mathcal{S}\right)$. Because of the weak limit of $\hat{\beta}$ (Theorem 6.10 and the notation therein), this implies

$$\limsup_n \mathbb{P}\left(\hat{\beta}_j = 0,\ \forall j \notin \mathcal{S}\right) \leq \mathbb{P}\left(u_j^* = 0,\ \forall j \notin \mathcal{S}\right).$$

If $\ell = 0$, then $\mathbf{u}^*$ is asymptotically normal, and the latter probability is zero. Otherwise, $\ell \neq 0$ and define the Gaussian random vector $\vec{W}_{\boldsymbol{\psi}} := 2\sum_{i=1}^{n'} W_i \boldsymbol{\psi}(\mathbf{z}_i')/n'$. The KKT conditions applied to $\mathbb{F}_{\infty,n'}$ provide

$$\vec{W}_{\boldsymbol{\psi}} + \frac{2}{n'} \sum_{i=1}^{n'} \boldsymbol{\psi}(\mathbf{z}_i') \boldsymbol{\psi}(\mathbf{z}_i')^T \mathbf{u}^* + \ell \mathbf{v}^* = 0,$$

for some vector $\mathbf{v}^* \in \mathbb{R}^p$ whose components v_j^* are less than one in absolute value when $j \notin \mathcal{S}$, and $v_j^* = \operatorname{sign}(\beta_j^*)$ when $j \in \mathcal{S}$. If $u_j^* = 0$ for all $j \notin \mathcal{S}$, we deduce

$$(\vec{W}_{\boldsymbol{\psi}})_{\mathcal{S}} + \Big[\frac{2}{n'} \sum_{i=1}^{n'} \boldsymbol{\psi}(\mathbf{z}_i') \boldsymbol{\psi}(\mathbf{z}_i')^T\Big]_{\mathcal{S},\mathcal{S}} \mathbf{u}_{\mathcal{S}}^* + \ell \operatorname{sign}(\beta_{\mathcal{S}}^*) = 0, \text{ and} \tag{6.7}$$

$$\left| (\vec{W}_{\boldsymbol{\psi}})_{\mathcal{S}^c} + \Big[\frac{2}{n'} \sum_{i=1}^{n'} \boldsymbol{\psi}(\mathbf{z}_i') \boldsymbol{\psi}(\mathbf{z}_i')^T\Big]_{\mathcal{S}^c,\mathcal{S}} \mathbf{u}_{\mathcal{S}}^* \right| \leq \ell, \tag{6.8}$$

componentwise and with obvious notation. Combining the two latter equations provides

$$\left| (\vec{W}_{\boldsymbol{\psi}})_{\mathcal{S}^c} - \Big[\sum_{i=1}^{n'} \boldsymbol{\psi}(\mathbf{z}_i') \boldsymbol{\psi}(\mathbf{z}_i')^T\Big]_{\mathcal{S}^c,\mathcal{S}} \Big[\sum_{i=1}^{n'} \boldsymbol{\psi}(\mathbf{z}_i') \boldsymbol{\psi}(\mathbf{z}_i')^T\Big]_{\mathcal{S},\mathcal{S}}^{-1} \Big(\vec{W}_{\boldsymbol{\psi}})_{\mathcal{S}} + \ell \operatorname{sign}(\beta_{\mathcal{S}}^*) \Big) \right| \leq \ell, \tag{6.9}$$

componentwise. Since the latter event is of probability strictly lower than one, this is still the case for the event $\{u_j^* = 0,\ \forall j \notin \mathcal{S}\}$. □

6.7.4 Proof of Theorem 6.12

The beginning of the proof is similar to the proof of Theorem 6.10. With obvious notation, $\check{\mathbf{u}}_{n,n'} = \arg\min_{\mathbf{u} \in \mathbb{R}^{p'}} \check{\mathbb{F}}_{n,n'}(\mathbf{u})$, where for every $\mathbf{u} \in \mathbb{R}^{p'}$,

$$\begin{aligned} \check{\mathbb{F}}_{n,n'}(\mathbf{u}) := & \frac{-2r_{n,n'}}{n'} \sum_{i=1}^{n'} \xi_{i,n} \boldsymbol{\psi}(\mathbf{z}_i')^T \mathbf{u} + \frac{1}{n'} \sum_{i=1}^{n'} \big(\boldsymbol{\psi}(\mathbf{z}_i')^T \mathbf{u} \big)^2 \\ + & \quad \mu_{n,n'} r_{n,n'}^2 \sum_{i=1}^{p'} \frac{1}{|\tilde{\beta}_i|^\delta} \left(|\beta_i^* + \frac{u_i}{r_{n,n'}}| - |\beta_i^*| \right). \end{aligned}$$

If $\beta_i^* \neq 0$, then

$$\frac{\mu_{n,n'} r_{n,n'}^2}{|\tilde{\beta}_i|^\delta} \left(|\beta_i^* + \frac{u_i}{r_{n,n'}}| - |\beta_i^*| \right) = \frac{\mu_{n,n'} r_{n,n'}}{|\tilde{\beta}_i|^\delta} u_i \operatorname{sign}(\beta_i^*) = \frac{\ell}{|\beta_i^*|^\delta} u_i \operatorname{sign}(\beta_i^*) + o_P(1).$$

If $\beta_i^* = 0$, then

$$\frac{\mu_{n,n'} r_{n,n'}^2}{|\tilde{\beta}_i|^\delta} \left(|\beta_i^* + \frac{u_i}{r_{n,n'}}| - |\beta_i^*| \right) = \frac{\mu_{n,n'} r_{n,n'} \nu_n^\delta}{|\nu_n \tilde{\beta}_i|^\delta} |u_i|.$$

By assumption $\nu_n\tilde{\beta}_i = O_p(1)$, and the latter term tends to the infinity in probability iff $u_i \neq 0$. As a consequence, if there exists some $i \notin \mathcal{S}$ s.t. $u_i \neq 0$, then $\check{\mathbb{F}}_{n,n'}(\mathbf{u})$ tends to the infinity. Otherwise, $u_i = 0$ when $i \notin \mathcal{S}$ and $\check{\mathbb{F}}_{n,n'}(\mathbf{u}) \to \check{\mathbb{F}}_{\infty,n'}(\mathbf{u}_{\mathcal{S}})$. Since $\check{\mathbb{F}}_{\infty,n'}$ is convex, we deduce [79] that $\check{\mathbf{u}}_{\mathcal{S}} \to \mathbf{u}^*_{\mathcal{S}}$, and $\check{\mathbf{u}}_{\mathcal{S}^c} \to 0_{\mathcal{S}^c}$, proving the asymptotic normality of $\check{\beta}_{n,n',\mathcal{S}}$.

Now, let us prove the oracle property. If $j \in \mathcal{S}$, then $\check{\beta}_j$ tends to β_j in probability and $\mathbb{P}(j \in \mathcal{S}_n) \to 1$. It suffices to show that $\mathbb{P}(j \in \mathcal{S}_n) \to 0$ when $j \notin \mathcal{S}$. If $j \notin \mathcal{S}$ and $j \in \mathcal{S}_n$, the KKT conditions on $\check{\mathbb{F}}_{n,n'}$ provide

$$\frac{-2r_{n,n'}}{n'}\sum_{i=1}^{n'}\xi_{i,n}\psi_j(\mathbf{z}'_i) + \frac{2}{n'}\sum_{i=1}^{n'}\psi_j(\mathbf{z}'_i)\boldsymbol{\psi}(\mathbf{z}'_i)^T\check{\mathbf{u}}_{n,n'} = -\frac{\mu_{n,n'}r_{n,n'}\nu_n^{\delta}}{|\nu_n\tilde{\beta}_j|^{\delta}}sign(\check{u}_j).$$

Due to the asymptotic normality of $\check{\beta}$ (that implies the one of $\check{\mathbf{u}}_{n,n'}$), the left hand side of the previous equation is asymptotically normal, when $\ell = 0$. On the other side, the r.h.s. tends to the infinity in probability because $\nu_n\tilde{\beta}_j = O_P(1)$. Therefore, the probability of the latter event tends to zero when $n \to \infty$. □

6.7.5 Proof of Theorem 6.13

By Lemma 6.7, we have $\hat{\beta}_{n,n'} = \arg\min_{\beta\in\mathbb{R}^{p'}}\mathbb{G}_{n,n'}(\beta)$, where

$$\mathbb{G}_{n,n'}(\beta) := \frac{2}{n'}\sum_{i=1}^{n'}\xi_{i,n}\boldsymbol{\psi}(\mathbf{z}'_i)^T(\beta^* - \beta) + \frac{1}{n'}\sum_{i=1}^{n'}\big(\boldsymbol{\psi}(\mathbf{z}'_i)^T(\beta^* - \beta)\big)^2 + \lambda_{n,n'}|\beta|_1.$$

Define also $\mathbb{G}_{\infty,n'}(\beta) := \sum_{i=1}^{n'}\big(\boldsymbol{\psi}(\mathbf{z}'_i)^T(\beta^* - \beta)\big)^2/n' + \lambda_0|\beta|_1$. We have

$$\big|\mathbb{G}_{n,n'}(\beta) - \mathbb{G}_{\infty,n'}(\beta)\big| \leq \left|\frac{2}{n'}\sum_{i=1}^{n'}\xi_{i,n}\boldsymbol{\psi}(\mathbf{z}'_i)^T(\beta^* - \beta)\right| + |\lambda_{n,n'} - \lambda_0| \times |\beta|_1.$$

By assumption, the second term on the r.h.s. converges to 0. We now show that the first term on the r.h.s. is negligible. Indeed, for every $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\Big(\big\|\frac{1}{n'}\sum_{i=1}^{n'}\xi_{i,n}\boldsymbol{\psi}(\mathbf{z}'_i)\big\| > \epsilon\Big) &\leq \mathbb{P}\Big(\frac{\|C_{\Lambda'}\|}{n'}\sum_{i=1}^{n'}|\hat{\tau}_{\mathbf{z}'_i} - \tau_{\mathbf{z}'_i}| \times \|\boldsymbol{\psi}(\mathbf{z}'_i)\| > \epsilon\Big)\\ &\leq \sum_{i=1}^{n'}\mathbb{P}\left(|\hat{\tau}_{\mathbf{z}'_i} - \tau_{\mathbf{z}'_i}| > Cst\epsilon\right),\end{aligned}$$

where Cst is the constant $(\|C_{\Lambda'}\| \times \|C_{\psi}\|)^{-1}$. Apply Lemma 6.21 with the $t = f_{\mathbf{Z},min}/4$ and t'/ϵ is a sufficiently small constant. When n is sufficiently large, we get

$$\mathbb{P}\Big(|\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}}| > Cst\epsilon\Big) \leq 4\exp\Big(-nh^{2p}Cst'\Big),$$

for some constant $Cst' > 0$. Thus, $\sum_{i=1}^{n'}\xi_{i,n}\boldsymbol{\psi}(\mathbf{z}'_i)/n' = o_{\mathbb{P}}(1)$, and $\mathbb{G}_{n,n'}(\beta) = \mathbb{G}_{\infty,n'}(\beta) + o_{\mathbb{P}}(1)$ for every β.

Since $\sum_{i=1}^{n'}\psi(\mathbf{z}'_i)\psi(\mathbf{z}'_i)^T/n'$ tends towards a matrix $M_{\psi,\mathbf{z}'}$, deduce that $\mathbb{G}_{\infty,n'}(\beta)$ tends to $\mathbb{G}_{\infty,\infty}(\beta)$ when $n' \to \infty$. Therefore, for all $\beta \in \mathbb{R}^{p'}$, $\mathbb{G}_{n,n'}(\beta)$ weakly tends to $\mathbb{G}_{\infty,\infty}(\beta)$. By the convexity argument, we deduce that $\arg\min_\beta \mathbb{G}_{n,n'}(\beta)$ weakly converges to $\arg\min_\beta \mathbb{G}_{\infty,\infty}(\beta)$. Since the latter minimizer is non random, the same convergence is true in probability. □

6.8 Proof of Theorem 6.14

We start as in the proof of Theorem 6.10. Define $\tilde{r}_{n,n'} := (nn' h^p_{n,n'})^{1/2}$, $\mathbf{u} := \tilde{r}_{n,n'}(\beta - \beta^*)$ and $\hat{\mathbf{u}}_{n,n'} := \tilde{r}_{n,n'}(\hat{\beta}_{n,n'} - \beta^*)$, so that $\hat{\beta}_{n,n'} = \beta^* + \hat{\mathbf{u}}_{n,n'}/\tilde{r}_{n,n'}$. We define for every $\mathbf{u} \in \mathbb{R}^{p'}$,

$$\mathbb{F}_{n,n'}(\mathbf{u}) := \frac{-2\tilde{r}_{n,n'}}{n'} \sum_{i=1}^{n'} \xi_{i,n} \boldsymbol{\psi}(\mathbf{z}'_i)^T \mathbf{u} + \frac{1}{n'} \sum_{i=1}^{n'} \big(\boldsymbol{\psi}(\mathbf{z}'_i)^T \mathbf{u}\big)^2 + \lambda_{n,n'} \tilde{r}^2_{n,n'} \Big(\big|\beta^* + \frac{\mathbf{u}}{\tilde{r}_{n,n'}}\big|_1 - \big|\beta^*\big|_1 \Big), \quad (6.10)$$

and we obtain $\hat{\mathbf{u}}_{n,n'} = \arg\min_{\mathbf{u} \in \mathbb{R}^{p'}} \mathbb{F}_{n,n'}(\mathbf{u})$.

Lemma 6.18. *Under the same assumptions as in Theorem 6.14,* $T_1 := (\tilde{r}_{n,n'}/n') \sum_{i=1}^{n'} \xi_{i,n} \boldsymbol{\psi}(\mathbf{z}'_i)$ *tends in law towards a Gaussian random vector* $\mathcal{N}(0, V_2)$.

This lemma is proved in Section 6.8.1. It will help to control the first term of Equation (6.10), which is simply $-2T_1^T \mathbf{u}$.

Concerning the second term of Equation (6.10), using Assumption 6.3.1(iii), we have for every $\mathbf{u} \in \mathbb{R}^{p'}$

$$\frac{1}{n'} \sum_{i=1}^{n'} \big(\boldsymbol{\psi}(\mathbf{z}'_i)^T \mathbf{u}\big)^2 \to \int \big(\boldsymbol{\psi}(\mathbf{z}')^T \mathbf{u}\big)^2 f_{\mathbf{z}',\infty}\, d\mathbf{z}'. \quad (6.11)$$

This has to be read as a convergence of a sequence of real numbers indexed by $\mathbf{u}$, because the design points $\mathbf{z}'_i$ are deterministic. We also have, for any $\mathbf{u} \in \mathbb{R}^{p'}$ and when n is large enough,

$$\big|\beta^* + \frac{\mathbf{u}}{\tilde{r}_{n,n'}}\big|_1 - \big|\beta^*\big|_1 = \sum_{i=1}^{p'} \Big(\frac{|u_i|}{\tilde{r}_{n,n'}} \mathbb{1}_{\{\beta^*_i = 0\}} + \frac{u_i}{\tilde{r}_{n,n'}} \operatorname{sign}(\beta^*_i) \mathbb{1}_{\{\beta^*_i \neq 0\}} \Big).$$

Therefore, by Assumption 6.3.1(ii)(b), for every $\mathbf{u} \in \mathbb{R}^{p'}$,

$$\lambda_{n,n'} \tilde{r}^2_{n,n'} \Big(\big|\beta^* + \frac{\mathbf{u}}{\tilde{r}_{n,n'}}\big|_1 - \big|\beta^*\big|_1 \Big) \to 0, \quad (6.12)$$

when (n, n') tends to the infinity. Combining Lemma 6.18 and Equations (6.10-6.12), and defining the function $\mathbb{F}_{\infty,\infty}$ by

$$\mathbb{F}_{\infty,\infty}(\mathbf{u}) := 2\tilde{\mathbf{W}}^T \mathbf{u} + \int \big(\boldsymbol{\psi}(\mathbf{z}')^T \mathbf{u}\big)^2 f_{\mathbf{z}',\infty}(\mathbf{z}') d\mathbf{z}',\ \mathbf{u} \in \mathbb{R}^{p'},$$

where $\tilde{W} \sim \mathcal{N}(0, V_2)$, we obtain that every finite-dimensional margin of $\mathbb{F}_{n,n'}$ converges weakly to the corresponding margin of $\mathbb{F}_{\infty,\infty}$. Now, applying the convexity lemma, we get

$$\hat{\mathbf{u}}_{n,n'} \xrightarrow{D} \mathbf{u}_{\infty,\infty}, \text{ where } \mathbf{u}_{\infty,\infty} := \arg \min_{\mathbf{u} \in \mathbb{R}^{p'}} \mathbb{F}_{\infty,\infty}(\mathbf{u}).$$

Since $\mathbb{F}_{\infty,\infty}(\mathbf{u})$ is a continuously differentiable convex function, we apply the first-order condition $\nabla \mathbb{F}_{\infty,\infty}(\mathbf{u}) = 0$, which yields $2\tilde{\mathbf{W}} + 2\int \boldsymbol{\psi}(\mathbf{z}')\boldsymbol{\psi}(\mathbf{z}')^T \mathbf{u}_{\infty,\infty} f_{\mathbf{z}',\infty}(\mathbf{z}') d\mathbf{z}' = 0$. As a consequence $\mathbf{u}_{\infty,\infty} = -V_1^{-1}\tilde{\mathbf{W}} \sim \mathcal{N}(0, \tilde{V}_{as})$, using Assumption 6.3.1(iv). We finally obtain $\tilde{r}_{n,n'}\big(\hat{\beta}_{n,n'} - \beta^*\big) \xrightarrow{D} \mathcal{N}\big(0, \tilde{V}_{as}\big)$, as claimed. □

6.8.1 Proof of Lemma 6.18 : convergence of T_1

Using a Taylor expansion, we have

$$T_1 := \frac{\tilde{r}_{n,n'}}{n'} \sum_{i=1}^{n'} \xi_{i,n} \boldsymbol{\psi}(\mathbf{z}'_i) = \frac{\tilde{r}_{n,n'}}{n'} \sum_{i=1}^{n'} \Big(\Lambda\big(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i}\big) - \Lambda\big(\tau_{1,2|\mathbf{Z}=\mathbf{z}'_i}\big) \Big) \boldsymbol{\psi}(\mathbf{z}'_i) = T_2 + T_3,$$

where the main term is

$$T_2 := \frac{\tilde{r}_{n,n'}}{n'} \sum_{i=1}^{n'} \Lambda'\big(\tau_{1,2|\mathbf{Z}=\mathbf{z}'_i}\big)\big(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i} - \tau_{1,2|\mathbf{Z}=\mathbf{z}'_i}\big)\boldsymbol{\psi}(\mathbf{z}'_i),$$

and the remainder is

$$T_3 := \frac{\tilde{r}_{n,n'}}{n'} \sum_{i=1}^{n'} \alpha_{3,i}\big(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i} - \tau_{1,2|\mathbf{Z}=\mathbf{z}'_i}\big)^2\boldsymbol{\psi}(\mathbf{z}'_i),$$

with $\forall i = 1, \ldots, n'$, $|\alpha_{3,i}| \leq C_{\Lambda''}/2$, by Assumption 6.3.1(v).

Using the definition (6.3) of $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}$, the definition of the weights $w_{i,n}(\mathbf{z})$ and the notation $\overline{\boldsymbol{\psi}}(\mathbf{z}) := \Lambda'\big(\tau_{1,2|\mathbf{Z}=\mathbf{z}}\big)\boldsymbol{\psi}(\mathbf{z})$, we rewrite $T_2 =: T_4 + T_5$, where

$$\begin{aligned} T_4 := \frac{\tilde{r}_{n,n'}}{n'n^2} \sum_{i=1}^{n'} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} & \frac{K_h(\mathbf{z}'_i - \mathbf{Z}_{j_1})K_h(\mathbf{z}'_i - \mathbf{Z}_{j_2})}{f^2_{\mathbf{Z}}(\mathbf{z}'_i)} \\ & \times \Big(g^*(\mathbf{X}_{j_1}, \mathbf{X}_{j_2}) - \mathbb{E}\big[g^*(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}'_i\big]\Big)\overline{\boldsymbol{\psi}}(\mathbf{z}'_i), \end{aligned} \tag{6.13}$$

$$\begin{aligned} T_5 := \frac{\tilde{r}_{n,n'}}{n'n^2} \sum_{i=1}^{n'} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} & K_h(\mathbf{z}'_i - \mathbf{Z}_{j_1})K_h(\mathbf{z}'_i - \mathbf{Z}_{j_2})\left(\frac{1}{\hat{f}_{\mathbf{Z}}(\mathbf{z}'_i)^2} - \frac{1}{f_{\mathbf{Z}}(\mathbf{z}'_i)^2}\right) \\ & \times \Big(g^*(\mathbf{X}_{j_1}, \mathbf{X}_{j_2}) - \mathbb{E}\big[g^*(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}'_i\big]\Big)\overline{\boldsymbol{\psi}}(\mathbf{z}'_i). \end{aligned} \tag{6.14}$$

Note that we can put together the terms (j_1, j_2) and (j_2, j_1). This corresponds to the substitution of g^* by its symmetrized version $\tilde{g}$. In the following, we will therefore assume that g^* has been symmetrized without loss of generality. The random variable T_4 can be seen (see Equation (6.13)) as a sum of (indexed by i) U-statistics of order 2. Its Hájek projection will yield the asymptotically normal dominant term of T_2.

To lighten notation, we denote $\tau_i := \tau_{1,2|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}'_i}$, $f(\cdot,\cdot) = f_{\mathbf{X},\mathbf{Z}}(\cdot,\cdot)$ and

$$g_{i,j_1,j_2} := g^*(\mathbf{X}_{j_1}, \mathbf{X}_{j_2}) - \mathbb{E}\big[g^*(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{z}'_i\big] = g^*(\mathbf{X}_{j_1}, \mathbf{X}_{j_2}) - \tau_i.$$

Implicitly, all the expectations we will consider are expectations conditionally on the sequence of $\mathbf{z}'_i$, $i \geq 1$.

First note that, by usual α-order limited expansions, we have

$$\begin{aligned} \mathbb{E}[T_4] = & \frac{\tilde{r}_{n,n'}}{n'n^2} \sum_{i=1}^{n'} n(n-1) \int \frac{K_h(\mathbf{z}'_i - \mathbf{z}_1)K_h(\mathbf{z}'_i - \mathbf{z}_2)}{f^2_{\mathbf{Z}}(\mathbf{z}'_i)}\big(g^*(\mathbf{x}_1, \mathbf{x}_2) - \tau_i\big) \\ & \times \overline{\boldsymbol{\psi}}(\mathbf{z}'_i)f(\mathbf{x}_1, \mathbf{z}_1)f(\mathbf{x}_2, \mathbf{z}_2)\, d\mathbf{x}_1\, d\mathbf{x}_2\, d\mathbf{z}_1\, d\mathbf{z}_2 \\ & - \frac{\tilde{r}_{n,n'}}{n'n} \sum_{i=1}^{n'} \tau_i\overline{\boldsymbol{\psi}}(\mathbf{z}'_i) \int \frac{K^2_h(\mathbf{z}'_i - \mathbf{z})}{f^2_{\mathbf{Z}}(\mathbf{z}'_i)} f(\mathbf{x}, \mathbf{z}) d\mathbf{x}\, d\mathbf{z} \\ = & \frac{(n-1)\tilde{r}_{n,n'}}{n'n} \sum_{i=1}^{n'} \int \frac{K(\mathbf{t}_1)K(\mathbf{t}_2)}{f^2_{\mathbf{Z}}(\mathbf{z}'_i)}\big(g^*(\mathbf{x}_1, \mathbf{x}_2) - \tau_i\big) \\ & \times \overline{\boldsymbol{\psi}}(\mathbf{z}'_i)f(\mathbf{x}_1, \mathbf{z}'_i - h\mathbf{t}_1)f(\mathbf{x}_2, \mathbf{z}'_i - h\mathbf{t}_1)\, d\mathbf{x}_1\, d\mathbf{x}_2\, d\mathbf{t}_1\, d\mathbf{t}_2 \\ & - \frac{\tilde{r}_{n,n'}}{n'nh^p} \sum_{i=1}^{n'} \tau_i\overline{\boldsymbol{\psi}}(\mathbf{z}'_i) \int \frac{K^2(\mathbf{t})}{f^2_{\mathbf{Z}}(\mathbf{z}'_i)} f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}, \mathbf{z}'_i - h\mathbf{t}) d\mathbf{x}\, d\mathbf{t} \end{aligned}$$

$$
\begin{aligned}
&= \frac{(n-1)\tilde{r}_{n,n'}h^{2\alpha}}{n'n}\sum_{i=1}^{n'}\int \frac{K(\mathbf{t}_1)K(\mathbf{t}_2)}{f^2_{\mathbf{Z}}(\mathbf{z}'_i)}\big(g^*(\mathbf{x}_1,\mathbf{x}_2)-\tau_i\big)\\
&\times \overline{\boldsymbol{\psi}}(\mathbf{z}'_i)d^{(\alpha)}_{\mathbf{Z}}f(\mathbf{x}_1,\mathbf{z}^*_i)\cdot\mathbf{t}_1^{(\alpha)}d^{(\alpha)}_{\mathbf{Z}}f(\mathbf{x}_2,\mathbf{z}^*_i)\cdot\mathbf{t}_2^{(\alpha)}\,d\mathbf{x}_1\,d\mathbf{x}_2\,d\mathbf{t}_1\,d\mathbf{t}_2\\
&- \frac{\tilde{r}_{n,n'}}{n'nh^p}\sum_{i=1}^{n'}\tau_i\int K^2\int\frac{\overline{\boldsymbol{\psi}}(\mathbf{z}'_i)}{f^2_{\mathbf{Z}}(\mathbf{z}'_i)}f(\mathbf{x},\mathbf{z}^*_i)\,d\mathbf{x}\\
&= O\Big(\tilde{r}_{n,n'}h^{2\alpha}+\tilde{r}_{n,n'}/(nh^p)\Big)=O\Big(\sqrt{nn'h^{p+4\alpha}}+\sqrt{n'/(nh^p)}\Big)=o(1),
\end{aligned}
$$

under Assumption 6.3.1 (ii). Above, we have denoted by $\mathbf{z}^*_i$ some vectors in $\mathbb{R}^p$ s.t. $\|\mathbf{z}'_i-\mathbf{Z}^*_i\|_\infty<1$. They depend on $\mathbf{z}'_i$, $\mathbf{x}_1$, $\mathbf{x}_2$ or $\mathbf{x}$, respectively.

Moreover, set

$$
T_4-\mathbb{E}[T_4]=\frac{\tilde{r}_{n,n'}}{n'n^2}\sum_{i=1}^{n'}\sum_{j_1,j_2=1}^{n}\zeta_{i,j_1,j_2}, \tag{6.15}
$$

$$
\zeta_{i,j_1,j_2}=\Big(K_h(\mathbf{z}'_i-\mathbf{Z}_{j_1})K_h(\mathbf{z}'_i-\mathbf{Z}_{j_2})g_{i,j_1,j_2}-\mathbb{E}\big[K_h(\mathbf{z}'_i-\mathbf{Z}_{j_1})K_h(\mathbf{z}'_i-\mathbf{Z}_{j_2})g_{i,j_1,j_2}\big]\Big)\frac{\overline{\boldsymbol{\psi}}(\mathbf{z}'_i)}{f^2_{\mathbf{Z}}(\mathbf{z}'_i)}.
$$

Note that $Var(T_4)=\mathbb{E}[T_4T_4^T]+o(1)$ and

$$
\mathbb{E}[T_4T_4^T]=\frac{\tilde{r}^2_{n,n'}}{(n')^2n^4}\sum_{i_1,i_2=1}^{n'}\sum_{j_1,j_2=1}^{n}\sum_{j_3,j_4=1}^{n}\mathbb{E}[\zeta_{i_1,j_1,j_2}\zeta^T_{i_2,j_3,j_4}].
$$

By independence, $\mathbb{E}[\zeta_{i,j_1,j_2}\zeta^T_{i,j_3,j_4}]=0$ when $\{j_1,j_2\}\cap\{j_3,j_4\}=\emptyset$.

Otherwise, assume that $j_1=j_3=j$ and there are no other identities among the four indices (j_1,j_2,j_3,j_4). Set

$$
\overline{\zeta}_i:=\mathbb{E}\big[K_h(\mathbf{z}'_i-\mathbf{Z}_1)K_h(\mathbf{z}'_i-\mathbf{Z}_2)g_{i,1,2}\big]\frac{\overline{\boldsymbol{\psi}}(\mathbf{z}'_i)}{f^2_{\mathbf{Z}}(\mathbf{z}'_i)}. \tag{6.16}
$$

Then,

$$
\mathbb{E}[\zeta_{i_1,j,j_2}\zeta^T_{i_2,j,j_4}]=\zeta_{i_1,j,j_2,i_2,j,j_4}-\overline{\zeta}_{i_1}\overline{\zeta}^T_{i_2},
$$

where

$$
\begin{aligned}
\zeta_{i_1,j,j_2,i_2,j,j_4}&:=\mathbb{E}\Big[K_h(\mathbf{z}'_{i_1}-\mathbf{Z}_j)K_h(\mathbf{z}'_{i_1}-\mathbf{Z}_{j_2})K_h(\mathbf{z}'_{i_2}-\mathbf{Z}_j)K_h(\mathbf{z}'_{i_2}-\mathbf{Z}_{j_4})g_{i_1,j,j_2}g^T_{i_2,j,j_4}\Big]\\
&\times\quad\frac{\overline{\boldsymbol{\psi}}(\mathbf{z}'_{i_1})\overline{\boldsymbol{\psi}}(\mathbf{z}'_{i_2})^T}{f^2_{\mathbf{Z}}(\mathbf{z}'_{i_1})f^2_{\mathbf{Z}}(\mathbf{z}'_{i_2})}\\
&=\quad\frac{\overline{\boldsymbol{\psi}}(\mathbf{z}'_{i_1})\overline{\boldsymbol{\psi}}(\mathbf{z}'_{i_2})^T}{h^pf^2_{\mathbf{Z}}(\mathbf{z}'_{i_1})f^2_{\mathbf{Z}}(\mathbf{z}'_{i_2})}\int K(\mathbf{t}_1)K(\mathbf{t}_2)K\big(\frac{\mathbf{z}'_{i_2}-\mathbf{z}'_{i_1}}{h}+\mathbf{t}_1\big)K(\mathbf{t}_4)\big(g^*(\mathbf{x}_1,\mathbf{x}_2)-\tau_{i_1}\big)\\
&\times\quad\big(g^*(\mathbf{x}_1,\mathbf{x}_4)-\tau_{i_2}\big)f(\mathbf{x}_1,\mathbf{z}'_{i_1}-h\mathbf{t}_1)f(\mathbf{x}_2,\mathbf{z}'_{i_1}-h\mathbf{t}_2)f(\mathbf{x}_4,\mathbf{z}'_{i_4}-h\mathbf{t}_4)\,d\mathbf{x}_1\,d\mathbf{x}_2\,d\mathbf{x}_4\,d\mathbf{t}_1\,d\mathbf{t}_2\,d\mathbf{t}_4.
\end{aligned}
$$

By assumption, $\zeta_{i_1,j,j_2,i_2,j,j_4}$ is zero when $i_1\neq i_2$. Otherwise, when $i_1=i_2=i$,

$$
\begin{aligned}
\zeta_{i,j,j_2,i,j,j_4}&\simeq\frac{\overline{\boldsymbol{\psi}}(\mathbf{z}'_i)\overline{\boldsymbol{\psi}}(\mathbf{z}'_i)^T}{h^pf_{\mathbf{Z}}(\mathbf{z}'_i)}\int K^2\int\big(g^*(\mathbf{x}_1,\mathbf{x}_2)-\tau_i\big)\big(g^*(\mathbf{x}_1,\mathbf{x}_4)-\tau_i\big)\\
&\times\quad f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_1|\mathbf{z}'_i)f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_2|\mathbf{z}'_i)f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_4|\mathbf{z}'_i)\,d\mathbf{x}_1\,d\mathbf{x}_2\,d\mathbf{x}_4:=C_{i,1,2,4}/h^p.
\end{aligned}
$$

It is easy to check that the terms with other identities among the four indices j_k, as ζ_{i,j,j_2,i,j,j_2} or $\zeta_{i,j,j_2,i,j,j}$ will induce negligible remainder terms. Therefore, we get

$$
\frac{\tilde{r}^2_{n,n'}}{(n')^2n^4}\sum_{i_1,i_2=1}^{n'}\sum_{j,j_2,j_4=1}^{n}\zeta_{i_1,j,j_2,i_2,j,j_4}\simeq\frac{1}{n'}\sum_{i=1}^{n'}C_{i,1,2,4}.
$$

Concerning the terms induced by the product of two $\overline{\zeta}_i$, note that, by limited expansions,

$$\begin{aligned}\overline{\zeta}_i &= \frac{\overline{\psi}(\mathbf{z}_i')}{f_{\mathbf{Z}}^2(\mathbf{z}_i')}\int K_h(\mathbf{z}_i'-\mathbf{z}_1)K_h(\mathbf{z}_i'-\mathbf{z}_2)\big(g^*(\mathbf{x}_1,\mathbf{x}_2)-\tau_i\big)f(\mathbf{x}_1,\mathbf{z}_1)f(\mathbf{x}_2,\mathbf{z}_2)\,d\mathbf{x}_1\,d\mathbf{z}_1\,d\mathbf{x}_2\,d\mathbf{z}_2\\ &= \frac{\overline{\psi}(\mathbf{z}_i')}{f_{\mathbf{Z}}^2(\mathbf{z}_i')}\int K(\mathbf{t}_1)K(\mathbf{t}_2)\big(g^*(\mathbf{x}_1,\mathbf{x}_2)-\tau_i\big)f(\mathbf{x}_1,\mathbf{z}_i'-h\mathbf{t}_1)f(\mathbf{x}_2,\mathbf{z}_i'-h\mathbf{t}_2)\,d\mathbf{x}_1\,d\mathbf{t}_1\,d\mathbf{x}_2\,d\mathbf{t}_2\\ &= \frac{h^{2\alpha}\overline{\psi}(\mathbf{z}_i')}{f_{\mathbf{Z}}^2(\mathbf{z}_i')}\int K(\mathbf{t}_1)K(\mathbf{t}_2)\big(g^*(\mathbf{x}_1,\mathbf{x}_2)-\tau_i\big)d_{\mathbf{Z}}^{(\alpha)}f(\mathbf{x}_1,\mathbf{z}_i^*)\cdot\mathbf{t}_1^{(\alpha)}d_{\mathbf{Z}}^{(\alpha)}f(\mathbf{x}_2,\mathbf{z}_i^*)\cdot\mathbf{t}_2^{(\alpha)}\,d\mathbf{x}_1\,d\mathbf{t}_1\,d\mathbf{x}_2\,d\mathbf{t}_2,\end{aligned}$$

with the same notation as above. As a consequence, $\sup_i\overline{\zeta}_i = O(h^{2\alpha})$ and

$$\frac{\tilde{r}_{n,n'}^2}{(n')^2n^4}\sum_{i_1,i_2=1}^{n'}\sum_{j,j_2,j_4=1}^{n}\overline{\zeta}_{i_1}\overline{\zeta}_{i_2}\simeq\frac{\tilde{r}_{n,n'}^2}{n}\Big(\frac{1}{n'}\sum_{i=1}^{n'}\overline{\zeta}_{i,1,2}\Big)^2=O\Big(\frac{h^{4\alpha}\tilde{r}_{n,n'}^2}{n}\Big)=O(n'h^{4\alpha+p})=o(1).$$

Therefore, we obtain

$$\frac{\tilde{r}_{n,n'}^2}{(n')^2n^4}\sum_{i_1,i_2=1}^{n'}\sum_{j,j_2,j_4=1}^{n}\mathbb{E}[\zeta_{i_1,j,j_2}\zeta_{i_2,j,j_4}^T]\simeq\frac{1}{n'}\sum_{i=1}^{n'}C_{i,1,2,4}.$$

To calculate $\mathbb{E}[T_4T_4^T]$, there are three other similar terms, that respectively correspond to the cases $j_1=j_4$, $j_2=j_3$ or $j_2=j_4$. Therefore, we deduce

$$\begin{aligned}Var(T_4)\simeq\mathbb{E}[T_4T_4^T]&\simeq\frac{4}{n'}\sum_{i=1}^{n'}C_{i,1,2,4}\\ &\simeq 4\int K^2\int\frac{\overline{\psi}(\mathbf{z})\overline{\psi}(\mathbf{z})^T}{f_{\mathbf{Z}}(\mathbf{z})}\int\big(g^*(\mathbf{x}_1,\mathbf{x}_2)-\tau_{1,2|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}}\big)\big(g^*(\mathbf{x}_1,\mathbf{x}_4)-\tau_{1,2|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}}\big)\\ &\times\ f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_1|\mathbf{z})f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_2|\mathbf{z})f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_4|\mathbf{z})f_{\mathbf{z}',\infty}(\mathbf{z})\,d\mathbf{x}_1\,d\mathbf{x}_2\,d\mathbf{x}_4\,d\mathbf{z},\end{aligned}$$

that is equal to the so-called variance-covariance matrix V_2. Now assume that $T_4-\mathbb{E}[T_4]$ is asymptotically normal, i.e. $T_4-\mathbb{E}[T_4]\xrightarrow{D}\mathcal{N}(0,V_2)$. This result will be proved in Subsection 6.8.2.

Let us decompose the term T_5, as defined in Equation (6.14). For every $i=1,\ldots,n'$, a usual Taylor expansion yields

$$\frac{1}{\hat{f}_{\mathbf{Z}}^2(\mathbf{z}_i')}-\frac{1}{f_{\mathbf{Z}}^2(\mathbf{z}_i')}=\frac{1}{f_{\mathbf{Z}}^2(\mathbf{z}_i')}\Big\{\frac{1}{\Big(1+\dfrac{\hat{f}_{\mathbf{Z}}(\mathbf{z}_i')-f_{\mathbf{Z}}(\mathbf{z}_i')}{f_{\mathbf{Z}}(\mathbf{z}_i')}\Big)^2}-1\Big\}=-2\frac{\hat{f}_{\mathbf{Z}}(\mathbf{z}_i')-f_{\mathbf{Z}}(\mathbf{z}_i')}{f_{\mathbf{Z}}^3(\mathbf{z}_i')}+T_{7,i},$$

where

$$T_{7,i}=\frac{3}{f_{\mathbf{Z}}^2(\mathbf{z}_i')}(1+\alpha_{7,i})^{-4}\Big(\frac{\hat{f}_{\mathbf{Z}}(\mathbf{z}_i')-f_{\mathbf{Z}}(\mathbf{z}_i')}{f_{\mathbf{Z}}(\mathbf{z}_i')}\Big)^2,\ \text{for some}\ |\alpha_{7,i}|\le\Big|\frac{\hat{f}_{\mathbf{Z}}(\mathbf{z}_i')-f_{\mathbf{Z}}(\mathbf{z}_i')}{f_{\mathbf{Z}}(\mathbf{z}_i')}\Big|.$$

Therefore, we obtain the decomposition $T_5=-2\,T_6+T_7$, where

$$\begin{aligned}T_6&:=\frac{\tilde{r}_{n,n'}}{n'n^2}\sum_{i=1}^{n'}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}K_h(\mathbf{z}_i'-\mathbf{Z}_{j_1})K_h(\mathbf{z}_i'-\mathbf{Z}_{j_2})\Big(\frac{\hat{f}_{\mathbf{Z}}(\mathbf{z}_i')-f_{\mathbf{Z}}(\mathbf{z}_i')}{f_{\mathbf{Z}}^3(\mathbf{z}_i')}\Big)\\ &\qquad\times\Big(g^*(\mathbf{X}_{j_1},\mathbf{X}_{j_2})-\mathbb{E}\big[g^*(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}_i'\big]\Big)\overline{\psi}(\mathbf{z}_i'),\\ T_7&:=\frac{\tilde{r}_{n,n'}}{n'n^2}\sum_{i=1}^{n'}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}K_h(\mathbf{z}_i'-\mathbf{Z}_{j_1})K_h(\mathbf{z}_i'-\mathbf{Z}_{j_2})T_{7,i}\\ &\qquad\times\Big(g^*(\mathbf{X}_{j_1},\mathbf{X}_{j_2})-\mathbb{E}\big[g^*(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}_i'\big]\Big)\overline{\psi}(\mathbf{z}_i').\end{aligned}$$

Summing up all the previous equations, we get

$$T_1 = \big(T_4 - \mathbb{E}[T_4]\big) - 2\,T_6 + T_7 + T_3 + o(1). \tag{6.17}$$

Afterwards, we will prove that all the remainders terms T_6, T_7 and T_3 are negligible, i.e. they tend to zero in probability. These results are respectively proved in Subsections 6.8.3, 6.8.4 and 6.8.5. Combining all these elements with the asymptotic normality of T_4 (proved in Subsection 6.8.2), we get $T_1 \xrightarrow{D} \mathcal{N}(0, V_2)$, as claimed. □

6.8.2 Proof of the asymptotic normality of T_4

We will lead the usual Hájek projection of T_4. To weaken notation, denote $\mathbb{E}[\zeta_{i,j_1,j_2}|\mathbf{X}_{j_1}, \mathbf{Z}_{j_1}] := \mathbb{E}[\zeta_{i,j_1,j_2}|j_1]$. Then, recalling (6.15), we can write

$$T_4 - \mathbb{E}[T_4] = T_{4,1} + T_{4,2} + T_{4,3}, \text{ with}$$

$$T_{4,1} := \frac{2\tilde{r}_{n,n'}}{n'n^2} \sum_{i=1}^{n'} \sum_{j_1,j_2=1}^{n} \mathbb{1}(j_1 \neq j_2)\mathbb{E}[\zeta_{i,j_1,j_2}|j_1],$$

$$T_{4,2} := \frac{2\tilde{r}_{n,n'}}{n'n^2} \sum_{i=1}^{n'} \sum_{j=1}^{n} \mathbb{E}[\zeta_{i,j,j}|j], \text{ and}$$

$$T_{4,3} := \frac{\tilde{r}_{n,n'}}{n'n^2} \sum_{i=1}^{n'} \sum_{j_1,j_2=1}^{n} \Big(\zeta_{i,j_1,j_2} - \mathbb{E}[\zeta_{i,j_1,j_2}|j_1] - \mathbb{E}[\zeta_{i,j_1,j_2}|j_2]\Big).$$

We will prove that $T_{4,2}$ and $T_{4,3}$ are $o_P(1)$. Therefore, the asymptotic normality of T_4 reduces to the one of $T_{4,1}$.

Note that $nT_{4,1}/2(n-1) = \sum_{j=1}^{n} \beta_{j,n,n'}$, where

$$\beta_{j,n,n'} := \frac{\tilde{r}_{n,n'}}{n'n} \sum_{i=1}^{n'} \mathbb{E}[\zeta_{i,j,0}|j],\ j = 1, \ldots, n,$$

by formally considering a random vector $\mathbf{Z}_0$ that is independent of the other $\mathbf{Z}_j$, $j \geq 1$. Therefore, we get a triangular array of random vectors $(\beta_{j,n,n'})_{j=1,\ldots,n}$, s.t., for a fixed n, the variables $\beta_{j,n,n'}$ are mutually independent given the vectors $\mathbf{z}'_i$, $i \geq 1$. Let us check Lyapunov's sufficient condition, that will imply the asymptotic normality of $T_{4,1}$. In other words, it is sufficient to prove that

$$\sum_{j=1}^{n} \|\beta_{j,n,n'}\|_\infty^3 \longrightarrow 0, \tag{6.18}$$

when n and n' tend to the infinity. Recalling (6.16), we can rewrite

$$\beta_{j,n,n'} = \frac{\tilde{r}_{n,n'}}{n'n} \sum_{i=1}^{n'} \Big\{K_h(\mathbf{z}'_i - \mathbf{Z}_j) \frac{\overline{\boldsymbol{\psi}}(\mathbf{z}'_i)}{f^2_{\mathbf{Z}}(\mathbf{z}'_i)} \int K_h(\mathbf{z}'_i - \mathbf{z})\big(g^*(\mathbf{x}, \mathbf{X}_j) - \tau_i\big) f(\mathbf{x}, \mathbf{z})\, d\mathbf{x}\, d\mathbf{z} - \overline{\zeta}_i\Big\} := \frac{\tilde{r}_{n,n'}}{n'n} \sum_{i=1}^{n'} \gamma_{i,j},$$

where $\sup_i \overline{\zeta}_i := O(h^{2\alpha})$. Note that

$$\|\beta_{j,n,n'}\|_\infty^3 \leq p^3 \frac{\tilde{r}^3_{n,n'}}{(n')^3 n^3} \sum_{i_1,i_2,i_3=1}^{n'} \|\gamma_{i_1,j}\|_\infty \|\gamma_{i_2,j}\|_\infty \|\gamma_{i_3,j}\|_\infty.$$

The terms that that involve some products by the means $\overline{\zeta}_{i_k}$, $k=1,2,3$, are negligible and they may be forgotten here. For some constants Cst, this provides

$$\sum_{j=1}^{n}\mathbb{E}\Big[\|\beta_{j,n,n'}\|_\infty^3\Big] \leq \frac{Cst\,\tilde{r}_{n,n'}^3}{(n')^3n^3}\sum_{j=1}^{n}\sum_{i_1,i_2,i_3=1}^{n'}\frac{\|\overline{\boldsymbol{\psi}}\|_\infty(\mathbf{z}'_{i_1})\|\overline{\boldsymbol{\psi}}\|_\infty(\mathbf{z}'_{i_2})\|\overline{\boldsymbol{\psi}}\|_\infty(\mathbf{z}'_{i_3})}{f_{\mathbf{Z}}^2(\mathbf{z}'_{i_1})f_{\mathbf{Z}}^2(\mathbf{z}'_{i_2})f_{\mathbf{Z}}^2(\mathbf{z}'_{i_3})}$$
$$\times\quad \mathbb{E}\Big[\big|K_h(\mathbf{z}'_{i_1}-\mathbf{Z}_j)\int K_h(\mathbf{z}'_{i_1}-\mathbf{z}_1)\big(g^*(\mathbf{x}_1,\mathbf{X}_j)-\tau_{i_1}\big)f(\mathbf{x}_1,\mathbf{z}_1)\,d\mathbf{x}_1\,d\mathbf{z}_1\big|$$
$$\times\quad \big|K_h(\mathbf{z}'_{i_2}-\mathbf{Z}_j)\int K_h(\mathbf{z}'_{i_2}-\mathbf{z}_2)\big(g^*(\mathbf{x}_2,\mathbf{X}_j)-\tau_{i_2}\big)f(\mathbf{x}_2,\mathbf{z}_2)\,d\mathbf{x}_2\,d\mathbf{z}_2\big|$$
$$\times\quad \big|K_h(\mathbf{z}'_{i_3}-\mathbf{Z}_j)\int K_h(\mathbf{z}'_{i_3}-\mathbf{z}_3)\big(g^*(\mathbf{x}_3,\mathbf{X}_j)-\tau_{i_3}\big)f(\mathbf{x}_3,\mathbf{z}_3)\,d\mathbf{x}_3\,d\mathbf{z}_3\big|\Big].$$

By some now usual changes of variables, the latter expectations are zero when one of the three indices i_1,i_2 and i_3 is different from the others. Thus, the non-zero expectations are obtained when $i_1=i_2=i_3$. In the latter case, we get

$$\sum_{j=1}^{n}\|\beta_{j,n,n'}\|_\infty^3 \leq \frac{Cst\,\tilde{r}_{n,n'}^3}{(n')^3n^2}\sum_{i=1}^{n'}\frac{\|\overline{\boldsymbol{\psi}}\|_\infty^3(\mathbf{z}'_i)}{f_{\mathbf{Z}}^6(\mathbf{z}'_i)}$$
$$\times\quad \int|K|_h^3(\mathbf{z}'_i-\mathbf{z})\Big|\int K_h(\mathbf{z}'_i-\mathbf{z}_1)\big(g^*(\mathbf{x}_1,\mathbf{x})-\tau_i\big)f(\mathbf{x}_1,\mathbf{z}_1)\,d\mathbf{x}_1\,d\mathbf{z}_1\Big|^3 f(\mathbf{x},\mathbf{z})\,d\mathbf{x}\,d\mathbf{z}$$
$$\leq\quad \frac{Cst\,\tilde{r}_{n,n'}^3}{(n')^3n^2h^{2p}}\sum_{i=1}^{n'}\frac{\|\overline{\boldsymbol{\psi}}\|_\infty^3(\mathbf{z}'_i)}{f_{\mathbf{Z}}^6(\mathbf{z}'_i)}\int|K|^3(\mathbf{t})\Big|\int K(\mathbf{t}_1)\big(g^*(\mathbf{x}_1,\mathbf{x})-\tau_i\big)f(\mathbf{x}_1,\mathbf{z}'_i-h\mathbf{t}_1)\,d\mathbf{x}_1\,d\mathbf{t}_1\Big|^3$$
$$\times\quad f(\mathbf{x},\mathbf{z}'_i-h\mathbf{t})\,d\mathbf{x}\,d\mathbf{t}=O\Big(\frac{\tilde{r}_{n,n'}^3}{(n')^2n^2h^{2p}}\Big)=O\Big(\frac{1}{(nn'h^p)^{1/2}}\Big)=o(1).$$

Concerning the remainder terms $T_{4,2}$ and $T_{4,3}$, note that $\mathbb{E}[T_{4,2}]=\mathbb{E}[T_{4,3}]=0$. Moreover, since $\mathbb{E}[\zeta_{i,j,j}|j]$ is centered,

$$\mathbb{E}[T_{4,2}T_{4,2}^T]=\frac{4\tilde{r}_{n,n'}^2}{(n')^2n^4}\sum_{i_1,i_2=1}^{n'}\sum_{j=1}^{n}\mathbb{E}\Big[\mathbb{E}[\zeta_{i_1,j,j}|j]\mathbb{E}[\zeta_{i_2,j,j}^T|j]\Big].$$

When $i_1\neq i_2$, some usual changes of variables yield

$$\mathbb{E}\Big[\mathbb{E}[\zeta_{i_1,j,j}|j]\mathbb{E}[\zeta_{i_2,j,j}^T|j]\Big]=\frac{\overline{\boldsymbol{\psi}}(\mathbf{z}'_{i_1})\overline{\boldsymbol{\psi}}(\mathbf{z}'_{i_2})^T}{f_{\mathbf{Z}}^2(\mathbf{z}'_{i_1})f_{\mathbf{Z}}^2(\mathbf{z}'_{i_2})}\tau_{i_1}\tau_{i_2}$$
$$\times\quad\Big(\mathbb{E}\big[K_h^2(\mathbf{z}'_{i_1}-\mathbf{Z}_j)K_h^2(\mathbf{z}'_{i_2}-\mathbf{Z}_j)\big]-\mathbb{E}\big[K_h^2(\mathbf{z}'_{i_1}-\mathbf{Z}_j)\big]\mathbb{E}\big[K_h^2(\mathbf{z}'_{i_2}-\mathbf{Z}_j)\big]\Big)=O(h^{-2p}),$$

uniformly w.r.t. i. By a similar reasoning, we can prove that

$$\sup_i\mathbb{E}\Big[\mathbb{E}[\zeta_{i,j,j}|j]\mathbb{E}[\zeta_{i,j,j}^T|j]\Big]=O(h^{-3p}).$$

Therefore,

$$\mathbb{E}[T_{4,2}T_{4,2}^T]=O\Big(\frac{\tilde{r}_{n,n'}^2}{(n')^2n^4}\big((n')^2nh^{-2p}+n'nh^{-3p}\big)\Big)=O\big(\frac{n'}{n^2h^p}+\frac{1}{n^2h^{2p}}\big)=o(1).$$

Concerning $T_{4,3}$, this remainder term is centered and

$$\mathbb{E}[T_{4,3}T_{4,3}^T]=\frac{\tilde{r}_{n,n'}^2}{(n')^2n^4}\sum_{i_1,i_2=1}^{n'}\sum_{j_1,j_2=1}^{n}\sum_{j_3,j_4=1}^{n}\mathbb{E}\Big[\{\zeta_{i_1,j_1,j_2}-\mathbb{E}[\zeta_{i_1,j_1,j_2}|j_1]-\mathbb{E}[\zeta_{i_1,j_1,j_2}|j_2]\}$$
$$\times\quad\{\zeta_{i_2,j_3,j_4}-\mathbb{E}[\zeta_{i_2,j_3,j_4}|j_3]-\mathbb{E}[\zeta_{i_2,j_3,j_4}|j_4]\}^T\Big]. \tag{6.19}$$

The expectations on the latter r.h.s. are zero when $\{j_1, j_2\} \cap \{j_3, j_4\} = \emptyset$ due to independence and the fact that the terms $\zeta_{i,j,j'}$ are centered. Otherwise, there is at least an identity among the indices j_k, $k = 1, \ldots, 4$. For instance, assume $j_1 = j_3 = j$ and $j \neq j_2 \neq j_4$. Then,

$$\begin{aligned}
&\mathbb{E}\Big[\{\zeta_{i_1,j,j_2} - \mathbb{E}[\zeta_{i_1,j,j_2}|j] - \mathbb{E}[\zeta_{i_1,j,j_2}|j_2]\}\{\zeta_{i_2,j,j_4} - \mathbb{E}[\zeta_{i_2,j,j_4}|j] - \mathbb{E}[\zeta_{i_2,j,j_4}|j_4]\}^T\Big] \\
&= \mathbb{E}\Big[\{\zeta_{i_1,j,j_2} - \mathbb{E}[\zeta_{i_1,j,j_2}|j]\}\{\zeta_{i_2,j,j_4} - \mathbb{E}[\zeta_{i_2,j,j_4}|j]\}^T\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\{\zeta_{i_1,j,j_2} - \mathbb{E}[\zeta_{i_1,j,j_2}|j]\}\{\zeta_{i_2,j,j_4} - \mathbb{E}[\zeta_{i_2,j,j_4}|j]\}^T\Big|j\Big]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[\zeta_{i_1,j,j_2}\zeta_{i_2,j,j_4}^T|j\big]\Big] - \mathbb{E}\Big[\mathbb{E}\big[\zeta_{i_1,j,j_2}|j\big]\mathbb{E}\big[\zeta_{i_2,j,j_4}^T|j\big]\Big] = 0.
\end{aligned}$$

Due to the symmetry of the latter cross-products, all cases of a single identity among the j_k, $k = 1, \ldots, 4$, yield the same result. Therefore, we need (at least) two identities among them to obtain non zero covariances in the calculation of $\mathbb{E}[T_{4,3}T_{4,3}^T]$. Thus, let us assume that $j_1 = j_3$ and $j_2 = j_4$. Then, the corresponding terms in (6.19) is

$$\begin{aligned}
&\frac{\tilde{r}_{n,n'}^2}{(n')^2 n^4} \sum_{i_1,i_2=1}^{n'} \sum_{j_1,j_2=1}^{n} \mathbb{E}\Big[\{\zeta_{i_1,j_1,j_2} - \mathbb{E}[\zeta_{i_1,j_1,j_2}|j_1] - \mathbb{E}[\zeta_{i_1,j_1,j_2}|j_2]\}\{\zeta_{i_2,j_1,j_2} - \mathbb{E}[\zeta_{i_2,j_1,j_2}|j_1] - \mathbb{E}[\zeta_{i_2,j_1,j_2}|j_2]\}^T\Big] \\
&= \frac{\tilde{r}_{n,n'}^2}{(n')^2 n^4} \sum_{i_1,i_2=1}^{n'} \sum_{j_1,j_2=1}^{n} \Big(\mathbb{E}\Big[\zeta_{i_1,j_1,j_2}\zeta_{i_2,j_1,j_2}^T\Big] - 2\mathbb{E}\Big[\mathbb{E}[\zeta_{i_1,j_1,j_2}|j_1]\mathbb{E}[\zeta_{i_2,j_1,j_2}|j_1]^T\Big]\Big) \\
&=: v_{4,3,1} - v_{4,3,2}.
\end{aligned}$$

By now usual techniques, we get

$$\begin{aligned}
v_{4,3,1} &= \frac{\tilde{r}_{n,n'}^2}{(n')^2 n^4} \sum_{i_1,i_2=1}^{n'} \sum_{j_1,j_2=1}^{n} \mathbb{E}\Big[\zeta_{i_1,j_1,j_2}\zeta_{i_2,j_1,j_2}^T\Big] \simeq \frac{\tilde{r}_{n,n'}^2}{(n')^2 n^4} \sum_{i=1}^{n'} \sum_{j_1,j_2=1}^{n} \mathbb{E}\Big[\zeta_{i,j_1,j_2}\zeta_{i,j_1,j_2}^T\Big] \\
&\simeq \frac{\tilde{r}_{n,n'}^2}{(n')^2 n^2} \sum_{i=1}^{n'} \frac{\overline{\boldsymbol{\psi}}(\mathbf{z}_i')\overline{\boldsymbol{\psi}}^T(\mathbf{z}_i')}{f_{\mathbf{Z}}^4(\mathbf{z}_i')} \int K_h^2(\mathbf{z}_i' - \mathbf{z}_1) K_h^2(\mathbf{z}_i' - \mathbf{z}_2) \big(g^*(\mathbf{x}_1, \mathbf{x}_2) - \tau_i\big)^2 f(\mathbf{x}_1, \mathbf{z}_1) \\
&\times f(\mathbf{x}_2, \mathbf{z}_2)\, d\mathbf{x}_1\, d\mathbf{z}_1\, d\mathbf{x}_2\, d\mathbf{z}_2 = O\Big(\frac{\tilde{r}_{n,n'}^2}{n' n^2 h^{2p}}\Big) = O\big(\frac{1}{nh^p}\big) = o(1).
\end{aligned}$$

Moreover,

$$\begin{aligned}
v_{4,3,2} &= \frac{\tilde{r}_{n,n'}^2}{(n')^2 n^4} \sum_{i_1,i_2=1}^{n'} \sum_{j_1,j_2=1}^{n} \mathbb{E}\Big[\mathbb{E}[\zeta_{i_1,j_1,j_2}|j_1]\mathbb{E}[\zeta_{i_2,j_1,j_2}^T|j_1]\Big] \\
&\simeq \frac{\tilde{r}_{n,n'}^2}{(n')^2 n^4} \sum_{i=1}^{n'} \sum_{j_1,j_2=1}^{n} \mathbb{E}\Big[\mathbb{E}[\zeta_{i,j_1,j_2}|j_1]\mathbb{E}[\zeta_{i,j_1,j_2}^T|j_1]\Big] \\
&\simeq \frac{\tilde{r}_{n,n'}^2}{(n')^2 n^2} \sum_{i=1}^{n'} \frac{\overline{\boldsymbol{\psi}}(\mathbf{z}_i')\overline{\boldsymbol{\psi}}^T(\mathbf{z}_i')}{f_{\mathbf{Z}}^4(\mathbf{z}_i')} \int K_h^2(\mathbf{z}_i' - \mathbf{z}_1) K_h(\mathbf{z}_i' - \mathbf{z}_2) K_h(\mathbf{z}_i' - \mathbf{z}_3) \big(g^*(\mathbf{x}_1, \mathbf{x}_2) - \tau_i\big) \\
&\times \big(g^*(\mathbf{x}_1, \mathbf{x}_3) - \tau_i\big) f(\mathbf{x}_1, \mathbf{z}_1) f(\mathbf{x}_2, \mathbf{z}_2) f(\mathbf{x}_3, \mathbf{z}_3)\, d\mathbf{x}_1\, d\mathbf{z}_1\, d\mathbf{x}_2\, d\mathbf{z}_2\, d\mathbf{x}_3\, d\mathbf{z}_3 \\
&= O\Big(\frac{\tilde{r}_{n,n'}^2}{n' n^2 h^p}\Big) = O\big(\frac{1}{n}\big) = o(1).
\end{aligned}$$

Another case of two identities occurs when $j_1 = j_4$ and $j_2 = j_3$, but it can be dealt similarly. Then, we have proved that $\mathbb{E}[T_{4,3}T_{4,3}^T] = o(1)$ and $T_{4,3} = o_P(1)$.

6.8.3 Convergence of T_6 to 0

Replacing $\hat{f}_{\mathbf{Z}}$ in the definition of T_6 above by the normalized sum of the kernels, we get

$$
\begin{aligned}
T_6 = & \frac{\tilde{r}_{n,n'}}{n'n^2}\sum_{i=1}^{n'}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\frac{K_h(\mathbf{z}'_i-\mathbf{Z}_{j_1})K_h(\mathbf{z}'_i-\mathbf{Z}_{j_2})}{f^3_{\mathbf{Z}}(\mathbf{z}'_i)}\big(\mathbb{E}[\hat{f}_{\mathbf{Z}}(\mathbf{z}'_i)]-f_{\mathbf{Z}}(\mathbf{z}'_i)\big)\\
& \times\Big(g^*(\mathbf{X}_{j_1},\mathbf{X}_{j_2})-\mathbb{E}\big[g^*(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}'_i\big]\Big)\overline{\psi}(\mathbf{z}'_i)\\
& +\frac{\tilde{r}_{n,n'}}{n'n^3}\sum_{i=1}^{n'}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\sum_{j_3=1}^{n}\frac{K_h(\mathbf{z}'_i-\mathbf{Z}_{j_1})K_h(\mathbf{z}'_i-\mathbf{Z}_{j_2})}{f^3_{\mathbf{Z}}(\mathbf{z}'_i)}\big(K_h(\mathbf{z}'_i-\mathbf{Z}_{j_3})-\mathbb{E}[\hat{f}_{\mathbf{Z}}(\mathbf{z}'_i)]\big)\\
& \times\Big(g^*(\mathbf{X}_{j_1},\mathbf{X}_{j_2})-\mathbb{E}\big[g^*(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}'_i\big]\Big)\overline{\psi}(\mathbf{z}'_i)=:T_{6,1}+T_{6,2}.
\end{aligned}
$$

The first term $T_{6,1}$ is a bias term. By Assumptions 6.9.1-6.9.2,

$$
\sup_{i=1,\ldots,n'}\big|\mathbb{E}[\hat{f}_{\mathbf{Z}}(\mathbf{z}'_i)]-f_{\mathbf{Z}}(\mathbf{z}'_i)\big|=O(h^\alpha).
$$

The sum of the diagonal terms in $T_{6,1}$ is

$$
-\frac{\tilde{r}_{n,n'}}{n'n^2}\sum_{i=1}^{n'}\sum_{j=1}^{n}\frac{K_h^2(\mathbf{z}'_i-\mathbf{Z}_j)}{f^3_{\mathbf{Z}}(\mathbf{z}'_i)}\big(\mathbb{E}[\hat{f}_{\mathbf{Z}}(\mathbf{z}'_i)]-f_{\mathbf{Z}}(\mathbf{z}'_i)\big)\mathbb{E}\big[g^*(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}'_i\big]\overline{\psi}(\mathbf{z}'_i),
$$

that is $O_{\mathbb{P}}\big(\tilde{r}_{n,n'}h^\alpha/(nh^p)\big)$. The sum of the extra-diagonal terms in $T_{6,1}$ is the r.v.

$$
\begin{aligned}
\overline{T}_{6,1}:= & \frac{\tilde{r}_{n,n'}}{n'n^2}\sum_{i=1}^{n'}\sum_{1\le j_1\neq j_2\le n}\frac{K_h(\mathbf{z}'_i-\mathbf{Z}_{j_1})K_h(\mathbf{z}'_i-\mathbf{Z}_{j_2})}{f^3_{\mathbf{Z}}(\mathbf{z}'_i)}\big(\mathbb{E}[\hat{f}_{\mathbf{Z}}(\mathbf{z}'_i)]-f_{\mathbf{Z}}(\mathbf{z}'_i)\big)\\
& \times\Big(g^*(\mathbf{X}_{j_1},\mathbf{X}_{j_2})-\mathbb{E}\big[g^*(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}'_i\big]\Big)\overline{\psi}(\mathbf{z}'_i).
\end{aligned}
$$

Note that $\mathbf{z}\mapsto f_{\mathbf{Z}}(\mathbf{z})$ and $(\mathbf{z}_1,\mathbf{z}_2)\mapsto\mathbb{E}\big[g^*(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Z}_1=\mathbf{z}_1,\mathbf{Z}_2=\mathbf{z}_2\big]$ are α-times continuously differentiable on $\mathcal{Z}$ and $\mathcal{Z}^2$ respectively, because of Assumptions 6.9.2 and 6.9.4. By α-order Taylor expansions of such terms, they yield some factors h^α. It is easy to check that the expectation of $(\overline{T}_{6,1})^2$ is of order $\tilde{r}^2_{n,n'}h^{2\alpha}/(n^2h^{2p})$. Therefore,

$$
T_{6,1}=O_{\mathbb{P}}\Big(\frac{\tilde{r}_{n,n'}h^\alpha}{nh^p}\Big)=O_{\mathbb{P}}\Big(\frac{(n')^{1/2}h^\alpha}{\sqrt{nh^p}}\Big)=o_{\mathbb{P}}(1).
$$

Concerning $T_{6,2}$, we can assume that the indices j_1, j_2 and j_3 are pairwise distinct. Indeed, the cases of one or two identities among such indices can be easily dealt. They yield an upper bound that is $O_{\mathbb{P}}(\tilde{r}_{n,n'}h^\alpha/(nh^p))$ as above, and they are negligible. Once we remove such terms from the triple sums (indexed by (j_1,j_2,j_3)) defining $T_{6,2}$, we get the centered r.v. $\overline{T}_{6,2}$. Let us calculate the second moment of $\overline{T}_{6,2}$.

$$
\begin{aligned}
\mathbb{E}\big[\overline{T}^2_{6,2}\big]:= & \frac{nn'h^p}{n'^2n^6}\sum_{i_1=1}^{n'}\sum_{i_2=1}^{n'}\sum_{1\le j_1\neq j_2\neq j_3\le n}\ \sum_{1\le j_4\neq j_5\neq j_6\le n}\mathbb{E}\Bigg[\frac{K_h(\mathbf{z}'_{i_1}-\mathbf{Z}_{j_1})K_h(\mathbf{z}'_{i_1}-\mathbf{Z}_{j_2})}{f^3_{\mathbf{Z}}(\mathbf{z}'_{i_1})}\\
& \times\big(K_h(\mathbf{z}'_{i_1}-\mathbf{Z}_{j_3})-\mathbb{E}[\hat{f}_{\mathbf{Z}}(\mathbf{z}'_{i_1})]\big)\Big(g^*(\mathbf{X}_{j_1},\mathbf{X}_{j_2})-\mathbb{E}\big[g^*(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}'_{i_1}\big]\Big)\overline{\psi}(\mathbf{z}'_{i_1})\\
& \times\frac{K_h(\mathbf{z}'_{i_2}-\mathbf{Z}_{j_4})K_h(\mathbf{z}'_{i_2}-\mathbf{Z}_{j_5})}{f^3_{\mathbf{Z}}(\mathbf{z}'_{i_2})}\big(K_h(\mathbf{z}'_{i_2}-\mathbf{Z}_{j_6})-\mathbb{E}[\hat{f}_{\mathbf{Z}}(\mathbf{z}'_{i_2})]\big)\\
& \times\Big(g^*(\mathbf{X}_{j_4},\mathbf{X}_{j_5})-\mathbb{E}\big[g^*(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}'_{i_2}\big]\Big)\overline{\psi}(\mathbf{z}'_{i_2})^T\Bigg]\\
=: & \frac{nn'h^p}{n'^2n^6}\sum_{i_1,i_2=1}^{n'}\ \sum_{1\le j_1\neq j_2\neq j_3\le n}\ \sum_{1\le j_4\neq j_5\neq j_6\le n}E_{i_1,i_2,j_1-j_6}.
\end{aligned}
$$

When all the indices of the latter sums are different, the latter expectation is zero. Non zero terms above are obtained only when j_3 and j_6 are equal to some other indices. In the case $j_3 = j_6$ and no other identity among the indices, we obtain two extra factors h^α through α-order limited expansions of $(\mathbf{z}_1, \mathbf{z}_2) \mapsto \mathbb{E}\big[g^*(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{Z}_1 = \mathbf{z}_1, \mathbf{Z}_2 = \mathbf{z}_2\big]$. This yields an order $O(nn'h^{p+2\alpha}/(nh^p))$. When j_3 and j_6 are equal to two different indices ($j_3 = j_4$ and $j_6 = j_2$, e.g.), we lose another factor h^p but we still benefit from the two latter factors h^α. This yields an upper bound $O(nn'h^{p+2\alpha}/(n^2h^{2p})) = o(1)$. The other situations can be managed similarly. We get

$$\mathbb{E}\big[\overline{T}_{6,2}^2\big] = O\Big(\frac{nn'h^{p+2\alpha}}{nh^p}\Big) = o(1).$$

Globally, we obtain $T_6 \to 0$ in probability under Assumptions 6.3.1(ii)(a). □

6.8.4 Convergence of T_7 to 0

Since $\sup_{i=1,\ldots,n'} |\hat{f}_{\mathbf{Z}}(\mathbf{z}_i') - f_{\mathbf{Z}}(\mathbf{z}_i')| = o_{\mathbb{P}}(1)$, note that

$$\sup_{i=1,\ldots,n'} |T_{7,i}| \leq \frac{6}{f_{\mathbf{Z},min}^4} \sup_{i=1,\ldots,n'} |\hat{f}_{\mathbf{Z}}(\mathbf{z}_i') - f_{\mathbf{Z}}(\mathbf{z}_i')|^2,$$

with a probability arbitrarily close to one. Apply Lemma 6.19 with a fixed $t > 0$ and $\mathbf{z} = \mathbf{z}_i'$ for each $i = 1, \ldots, n'$

$$\mathbb{P}\bigg(\sup_{i=1,\ldots,n'} |T_{7,i}| \geq \frac{6}{f_{\mathbf{Z},min}^4}\left(\frac{C_{K,\alpha}h^\alpha}{\alpha!} + t\right)^2\bigg) \leq 2n' \exp\bigg(-\frac{nh^pt^2}{2f_{\mathbf{Z},max}\int K^2 + (2/3)C_K t}\bigg).$$

Set $t \propto h^{\alpha/2}$. Deduce $\sup_{i=1,\ldots,n'} |T_{7,i}| = O_{\mathbb{P}}(h^\alpha)$ since $nh^{p+\alpha}/\ln n' \to \infty$ by assumption. Then,

$$\begin{aligned}|T_7| \leq{}& \frac{\tilde{r}_{n,n'}}{n'n^2} \sup_i |T_{7,i}| \sum_{i=1}^{n'} |\overline{\psi}(\mathbf{z}_i')| \\ &\times \sum_{j_1=1}^{n}\sum_{j_2=1}^{n} |K|_h(\mathbf{z}_i' - \mathbf{Z}_{j_1})|K|_h(\mathbf{z}_i' - \mathbf{Z}_{j_2})\Big|g^*(\mathbf{X}_{j_1}, \mathbf{X}_{j_2}) - \mathbb{E}\big[g^*(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}_i'\big]\Big|.\end{aligned}$$

The expectation of the double sum is $O(h^\alpha)$, by an α-order limited expansion of $(\mathbf{z}_1, \mathbf{z}_2) \mapsto \mathbb{E}\big[g^*(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{Z}_1 = \mathbf{z}_1, \mathbf{Z}_2 = \mathbf{z}_2\big]$. Then, by Markov's inequality, we deduce

$$T_7 = O_{\mathbb{P}}(\tilde{r}_{n,n'} \sup_i |T_{7,i}|h^\alpha) = O_{\mathbb{P}}(\tilde{r}_{n,n'}h^{2\alpha}) = O_{\mathbb{P}}\big((n'nh^{p+4\alpha})^{1/2}\big),$$

and then $T_7 = o_{\mathbb{P}}(1)$ due to Assumption 6.3.1(ii)(a). □

6.8.5 Convergence of T_3 to 0

For every $\epsilon > 0$, by Markov's inequality,

$$\mathbb{P}(|T_3| > \epsilon) \leq \frac{C_{\Lambda''}\tilde{r}_{n,n'}}{2n'\epsilon} \sum_{i=1}^{n'} \mathbb{E}\big[\big(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}_i'} - \tau_{1,2|\mathbf{Z}=\mathbf{z}_i'}\big)^2\big]\boldsymbol{\psi}(\mathbf{z}_i').$$

An approximated calculation of $\mathbb{E}\big[\big(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}_i'} - \tau_{1,2|\mathbf{Z}=\mathbf{z}_i'}\big)^2\big]$ can be obtained following the steps of the proof of Lemma 6.24. Indeed, it can be easily seen that the order of magnitude of the latter expectation is the same as the variance of $U_{n,i}(g^*)$, and then of its Hájek projection $\hat{U}_{n,i}(g)$. Since the latter variance is $O((nh^p)^{-1})$, we get

$$\mathbb{P}(|T_3| > \epsilon) \leq B\frac{\tilde{r}_{n,n'}}{nh^p\epsilon},$$

for some constant B. Since $n'/(nh^p) \to 0$, we get $T_3 = o_{\mathbb{P}}(1)$, as claimed. □

6.9 Technical results concerning the first-step estimator

Three possible choices for g^* are given in [40]

$$g_1(\mathbf{X}_i, \mathbf{X}_j) := 4 \cdot \mathbb{1}\big\{X_{i,1} < X_{j,1}, X_{i,2} < X_{j,2}\big\} - 1,$$
$$g_2(\mathbf{X}_i, \mathbf{X}_j) := \mathbb{1}\big\{(X_{i,1} - X_{j,1}).(X_{i,2} - X_{j,2}) > 0\big\} - \mathbb{1}\big\{(X_{i,1} - X_{j,1}).(X_{i,2} - X_{j,2}) < 0\big\},$$
$$g_3(\mathbf{X}_i, \mathbf{X}_j) := 1 - 4 \cdot \mathbb{1}\big\{X_{i,1} < X_{j,1}, X_{i,2} > X_{j,2}\big\},$$

where $\mathbb{1}$ is the indicator function. In the following, we assume that we have chosen g^* as one of the g_k for a fixed $k \in \{1, 2, 3\}$.

Assumption 6.9.1. *The kernel K is bounded, and set $\|K\|_\infty =: C_K$. It is symmetrical and satisfies $\int K = 1$, $\int |K| < \infty$. This kernel is of order α for some integer $\alpha > 1$: for all $j = 1, \dots, \alpha - 1$ and every indices $i_1, \dots, i_j$ in $\{1, \dots, p\}$, $\int_{\mathbb{R}^p} K(\mathbf{u}) u_{i_1} \dots u_{i_j}\, d\mathbf{u} = 0$,.*

Assumption 6.9.2. *$f_{\mathbf{Z}}$ is α-times continuously differentiable and there exists a constant $C_{K,\alpha} > 0$ s.t., for all $\mathbf{z} \in \mathcal{Z}$,*

$$\int |K|(\mathbf{u}) \sum_{i_1, \dots, i_\alpha = 1}^{p} |u_{i_1} \dots u_{i_\alpha}| \sup_{t \in [0,1]} \Big| \frac{\partial^\alpha f_{\mathbf{Z}}}{\partial z_{i_1} \dots \partial z_{i_\alpha}} (\mathbf{z} + t\mathbf{u}) \Big|\, d\mathbf{u} \le C_{K,\alpha}.$$

Assumption 6.9.3. *There exist two positive constants $f_{\mathbf{Z},min}$ and $f_{\mathbf{Z},max}$ such that, for every $\mathbf{z} \in \mathcal{Z}$, $f_{\mathbf{Z},min} \le f_{\mathbf{Z}}(\mathbf{z}) \le f_{\mathbf{Z},max}$.*

Lemma 6.19. *Under Assumptions 6.9.1, 6.9.2 and 6.9.3, we have for any $t > 0$,*

$$\mathbb{P}\bigg(\big| \hat{f}_{\mathbf{Z}}(\mathbf{z}) - f_{\mathbf{Z}}(\mathbf{z}) \big| \ge \frac{C_{K,\alpha} h^\alpha}{\alpha!} + t \bigg) \le 2 \exp\bigg(- \frac{n h^p t^2}{2 f_{\mathbf{Z},max} \int K^2 + (2/3) C_K t} \bigg).$$

Lemma 6.20. *Under Assumptions 6.9.1-6.9.3 and if $C_{K,\alpha} h^\alpha / \alpha! < f_{\mathbf{Z},min}$, the estimator $\hat{f}_{\mathbf{Z}}(\mathbf{z})$ is strictly positive with a probability larger than*

$$1 - 2 \exp\Big(- n h^p \big(f_{\mathbf{Z},min} - C_{K,\alpha} h^\alpha / \alpha! \big)^2 / \big(2 f_{\mathbf{Z},max} \int K^2 + (2/3) C_K (f_{\mathbf{Z},min} - C_{K,\alpha} h^\alpha / \alpha!) \big) \Big).$$

Assumption 6.9.4. *For every $\mathbf{x} \in \mathbb{R}^2$, $\mathbf{z} \mapsto f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}, \mathbf{z})$ is differentiable almost everywhere up to the order α, $\mathbf{z} \in \mathcal{Z}$. For every $0 \le k \le \alpha$ and every $1 \le i_1, \dots, i_\alpha \le p$, let*

$$\mathcal{H}_{k,\vec{\imath}}(\mathbf{u}, \mathbf{v}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{z}) := \sup_{t \in [0,1]} \bigg| \frac{\partial^k f_{\mathbf{X},\mathbf{Z}}}{\partial z_{i_1} \dots \partial z_{i_k}} \Big(\mathbf{x}_1, \mathbf{z} + t\mathbf{u} \Big) \frac{\partial^{\alpha - k} f_{\mathbf{X},\mathbf{Z}}}{\partial z_{i_{k+1}} \dots \partial z_{i_\alpha}} \Big(\mathbf{x}_2, \mathbf{z} + t\mathbf{v} \Big) \bigg|,$$

denoting $\vec{\imath} = (i_1, \dots, i_\alpha)$. Assume that $\mathcal{H}_{k,\vec{\imath}}(\mathbf{u}, \mathbf{v}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{z})$ is integrable and there exists a finite constant $C_{\mathbf{X}\mathbf{Z},\alpha} > 0$, such that, for every $\mathbf{z} \in \mathcal{Z}$,

$$\int |K|(\mathbf{u}) |K|(\mathbf{v}) \sum_{k=0}^{\alpha} \binom{\alpha}{k} \sum_{i_1, \dots, i_\alpha = 1}^{p} \mathcal{H}_{k,\vec{\imath}}(\mathbf{u}, \mathbf{v}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{z}) |u_{i_1} \dots u_{i_k} v_{i_{k+1}} \dots v_{i_\alpha}|\, d\mathbf{u}\, d\mathbf{v}\, d\mathbf{x}_1\, d\mathbf{x}_2$$

is less than $C_{\mathbf{X}\mathbf{Z},\alpha}$.

Lemma 6.21 (Exponential bound for the estimated conditional Kendall's tau)**.** *Under Assumptions 6.9.1-6.9.4, for every $t > 0$ such that $C_{K,\alpha}h^\alpha/\alpha! + t \leq f_{\mathbf{Z},min}/2$ and every $t' > 0$, we have*

$$\mathbb{P}\left(|\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}}| > \frac{c_k}{f^2_{\mathbf{z}}(\mathbf{z})}\Big(\frac{C_{\mathbf{XZ},\alpha}h^\alpha}{\alpha!} + \frac{3f_{\mathbf{z}}(\mathbf{z})\int K^2}{2nh^p} + t'\Big) \times \left(1 + \frac{16f^2_{\mathbf{Z}}(\mathbf{z})}{f^3_{\mathbf{Z},min}}\Big(\frac{C_{K,\alpha}h^\alpha}{\alpha!} + t\Big)\right)\right)$$
$$\leq 2\exp\Big(-\frac{nh^pt^2}{2f_{\mathbf{Z},max}\int K^2 + (2/3)C_Kt}\Big) + 2\exp\Big(-\frac{(n-1)h^{2p}t'^2}{4f^2_{\mathbf{Z},max}(\int K^2)^2 + (8/3)C_K^2t'}\Big)$$
$$+ 2\exp\Big(-\frac{nh^p(f_{\mathbf{z}}(\mathbf{z}) - C_{\tilde{K},2}h^2)^2}{8f_{\mathbf{Z},max}\int \tilde{K}^2 + 4C_{\tilde{K}}(f_{\mathbf{z}}(\mathbf{z}) - C_{\tilde{K},2}h^2)/3}\Big),$$

with $c_1 := c_3 := 4$ and $c_2 := 2$.

Remark 6.22. *In Lemma 6.20 and 6.21, $f_{\mathbf{Z},min}$ can be replaced by $f_{\mathbf{Z}}(\mathbf{z})$. Moreover, when the support of K is included in $[-c, c]$ for some $c > 0$, $f_{\mathbf{Z},max}$ can be replaced by $\sup_{\tilde{\mathbf{z}}\in\mathcal{V}_{(\mathbf{z},\epsilon)}} f_{\mathbf{Z}}(\tilde{\mathbf{z}})$, denoting by $\mathcal{V}(\mathbf{z}, \epsilon)$ a closed ball of center $\mathbf{z}$ and any radius $\epsilon > 0$, when $n\,c < \epsilon$.*

Lemma 6.23 (Consistency)**.** *Under Assumption 6.9.1, if $nh^p_n \to \infty$, $\lim K(\mathbf{t})|\mathbf{t}|^p = 0$ when $|\mathbf{t}| \to \infty$, $f_{\mathbf{Z}}$ and $\mathbf{z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$ are continuous on $\mathcal{Z}$, then $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}$ tends to $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ in probability, when $n \to \infty$.*

To derive the asymptotic law of this estimator, we will assume:

Assumption 6.9.5. *(i) $nh^p_n \to \infty$ and $nh^{p+2\alpha}_n \to 0$; (ii) $K(\,\cdot\,)$ is compactly supported.*

Lemma 6.24 (Asymptotic normality)**.** *Assume 6.9.1, 6.9.4, 6.9.5, that the $\mathbf{z}'_i$ are distinct and that $f_{\mathbf{Z}}$ and $\mathbf{z} \mapsto f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}, \mathbf{z})$ are continuous on $\mathcal{Z}$, for every $\mathbf{x}$.*

Then, $(nh^p_n)^{1/2}\left(\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}'_i} - \tau_{1,2|\mathbf{Z}=\mathbf{z}'_i}\right)_{i=1,\ldots,n'} \xrightarrow{D} \mathcal{N}(0, \mathbb{H})$ as $n \to \infty$, where $\mathbb{H}$ is a $n' \times n'$ real matrix defined by

$$[\mathbb{H}]_{i,j} = \frac{4\int K^2\mathbb{1}_{\{\mathbf{z}'_i=\mathbf{z}'_j\}}}{f_{\mathbf{Z}}(\mathbf{z}'_i)}\big\{\mathbb{E}[\tilde{g}(\mathbf{X}_1, \mathbf{X})\tilde{g}(\mathbf{X}_2, \mathbf{X})|\mathbf{Z} = \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}'_i] - \tau^2_{1,2|\mathbf{Z}=\mathbf{z}'_i}\big\},$$

for every $1 \leq i, j \leq n'$, and $(\mathbf{X}, \mathbf{Z})$, $(\mathbf{X}_1, \mathbf{Z}_1)$, $(\mathbf{X}_2, \mathbf{Z}_2)$ are independent copies, where $\tilde{g}$ is the symmetrized version $\tilde{g}(\mathbf{x}_1, \mathbf{x}_2) := g^*(\mathbf{x}_1, \mathbf{x}_2) + g^*(\mathbf{x}_2, \mathbf{x}_1))/2$.

6.10 Estimation results for a particular sample

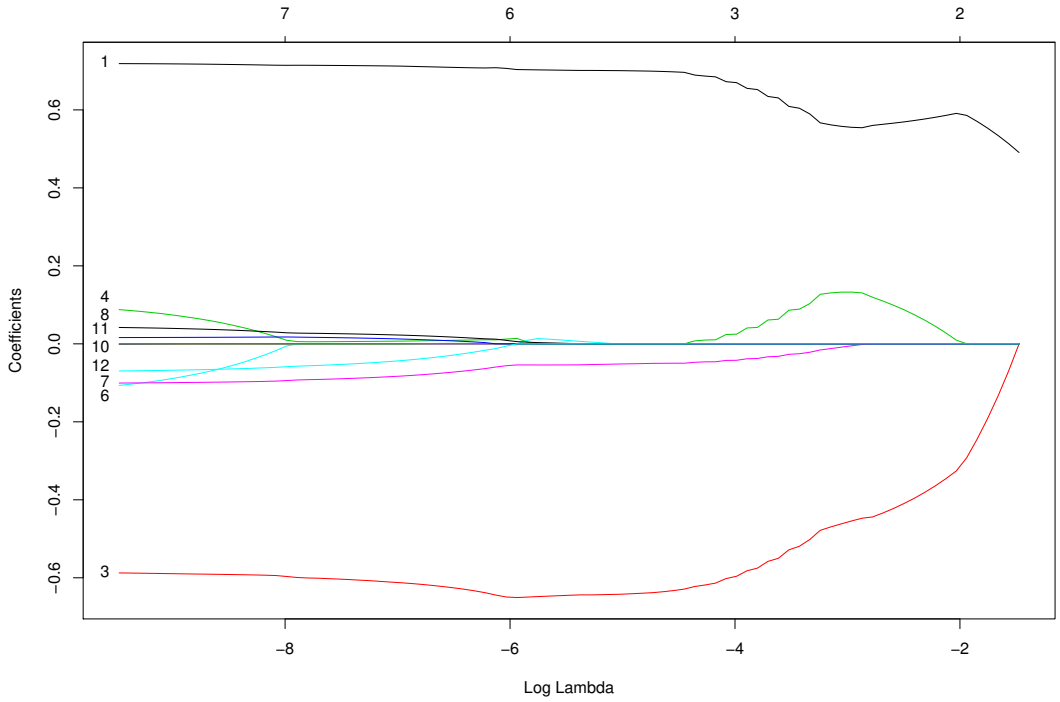


Figure 6.4: Evolution of the estimated non-zero coefficients as a function of the regularization parameter λ. The non-zero coefficients are $\beta_1 = 3/4$ and $\beta_3 = 3/4$. Note that the coefficients $\hat{\beta}_2$, $\hat{\beta}_5$ and $\hat{\beta}_9$ coefficients are always zero (and are not displayed).

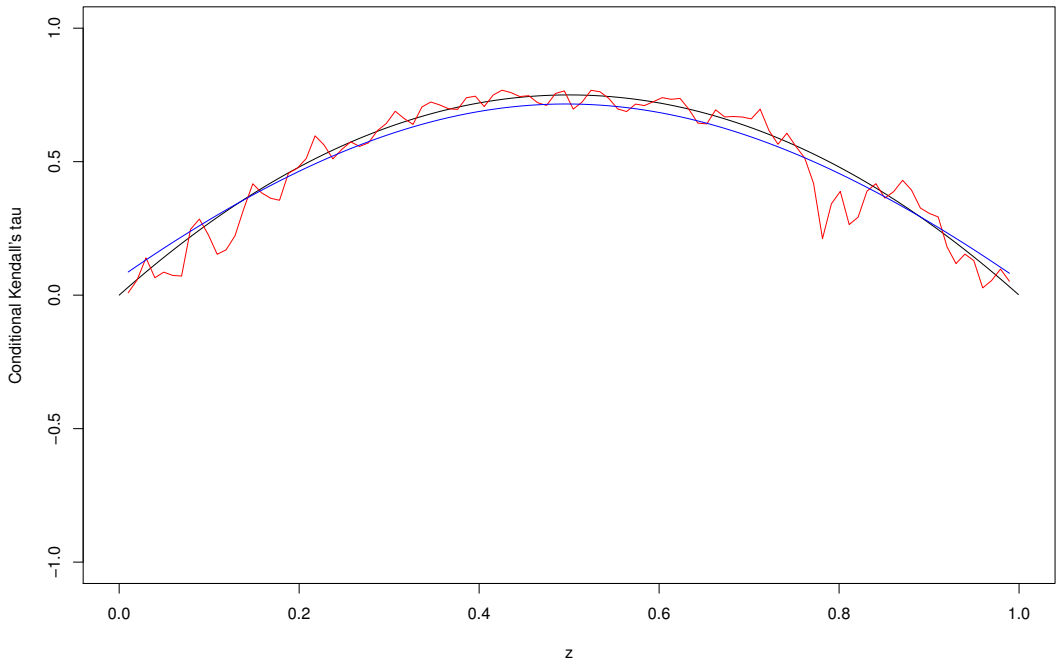


Figure 6.5: True conditional Kendall's tau $\tau_{1,2|Z=z}$ (black curve), estimated conditional Kendall's tau $\hat{\tau}_{1,2|Z=z}$ (red curve), and prediction $\Lambda^{(-1)}\big(\boldsymbol{\psi}(z)^T\hat{\beta}\big)$ (blue curve) as a function of z. For the blue curve, the regularization parameter is $2\hat{\lambda}^{cv} \simeq 0.034$ where $\hat{\lambda}^{cv}$ is selected by Algorithm 6.

Chapter 7

A classification point-of-view on conditional Kendall's tau

Abstract

We show how the problem of estimating conditional Kendall's tau can be rewritten as a classification task. Conditional Kendall's tau is a conditional dependence parameter that is a characteristic of a given pair of random variables. The goal is to predict whether the pair is concordant (value of 1) or discordant (value of -1) conditionally on some covariates. We prove the consistency and the asymptotic normality of a family of penalized approximate maximum likelihood estimators, including the equivalent of the logit and probit regressions in our framework. Then, we detail specific algorithms adapting usual machine learning techniques, including nearest neighbors, decision trees, random forests and neural networks, to the setting of the estimation of conditional Kendall's tau. Finite sample properties of these estimators and their sensitivities to each component of the data-generating process are assessed in a simulation study. Finally, we apply all these estimators to a dataset of European stock indices.

Keywords: Conditional Kendall's tau, conditional dependence measure, machine learning, classification task, stock indices.

Based on [41]: Derumigny, A., & Fermanian, J. D., A classification point-of-view about conditional Kendall's tau. *Computational Statistics & Data Analysis*, 135, 70-94, 2019.

7.1 Introduction

Beside linear correlations, most dependence measures between two random variables are functions of the underlying copula only: Spearman's rho, Kendall's tau, Blomqvist's beta, Gini's measure of association, etc. As a consequence, they are independent of the corresponding margins. This is seen as a positive point. See Joe [76], Nelsen [106], for instance, and, as a reminder, some basic definitions in 7.7. Such measures are well-known and widely used by practitioners. When some covariates are available, natural extensions of these tools can be defined, providing so-called "conditional" measures of dependence. In theory, it is sufficient to replace copulas by conditional copulas to obtain the "conditional version" of any dependence measure. Surprisingly, this simple and fruitful idea has not yet been

widely used in the literature. Nonetheless, in a series of papers, Gijbels et al. [63, 64, 65, 61] have popularized this approach, with a focus on conditional Kendall's tau and Spearman's rho. Note that conditional dependence measures have been invoked in different frameworks, often without any explicit link with conditional copulas: truncated data (Tsai [136], e.g.), multivariate dynamic models (Jondeau and Rockinger [78], Almeida and Czado [6], among others), vine structures (So and Yeung [127]), etc.

Now, let us introduce our key dependence measure: for each $\mathbf{z} \in \mathbb{R}^p$, the conditional Kendall's tau of a bivariate random vector $\mathbf{X} := (X_1, X_2)$ given some covariates $\mathbf{Z} = \mathbf{z}$ may be defined as

$$\tau_{1,2|\mathbf{Z}=\mathbf{z}} = \mathbb{P}\big((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) - \mathbb{P}\big((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) < 0 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big),$$

where $\mathbf{X}_1 = (X_{1,1}, X_{1,2})$ and $\mathbf{X}_2 = (X_{2,1}, X_{2,2})$ are two independent versions of $\mathbf{X}$. To simplify, we will assume that the law of $\mathbf{X}$ given $\mathbf{Z} = \mathbf{z}$ is continuous w.r.t. the Lebesgue measure, for every $\mathbf{z}$. This implies

$$\tau_{1,2|\mathbf{Z}=\mathbf{z}} = 2\,\mathbb{P}\big((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) - 1.$$

A conditional Kendall's tau belongs to the interval $[-1, 1]$ and reflects a positive ($\tau_{1,2|\mathbf{Z}=\mathbf{z}} > 0$) or negative ($\tau_{1,2|\mathbf{Z}=\mathbf{z}} < 0$) dependence between X_1 and X_2, given $\mathbf{Z} = \mathbf{z}$. Unlike correlations, this measure has the advantage of being always defined, even if some X_k, $k = 1, 2$, has no finite second moments (when it follows a Cauchy distribution, for example).

Some estimators of conditional Kendall's tau have already been proposed in the literature, either as a by-product of the estimation of conditional copulas - see Gijbels et al. [63] and Fermanian and Lopez [50] - or directly, as in Derumigny and Fermanian [39, 40]. Nonetheless, to the best of our knowledge, nobody has yet noticed the relationship between conditional Kendall's tau and classification methods.

Let us explain this simple idea. Denote $W := 2 \times \mathbb{1}\{(X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0\} - 1$ and

$$\mathbb{P}\big((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) = \mathbb{P}\big(W = 1 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) =: p(\mathbf{z}).$$

Actually, the prediction of concordance/discordance among pairs of observations $(\mathbf{X}_1, \mathbf{X}_2)$ given $\mathbf{Z}$ can be seen as a classification task of such pairs. If a model is able to evaluate the conditional probability of observing concordant pairs of observations, then it is able to evaluate conditional Kendall's tau, and the former quantity is one of the outputs of most classification techniques. Therefore, most classifiers can potentially be invoked (for example linear classifiers, decision trees, random forests, neural networks and so on [54]), but applied here to pairs of observations.

Indeed, for every $1 \leq i, j \leq n$, $i \neq j$, define $W_{(i,j)}$ as

$$W_{(i,j)} := 2 \times \mathbb{1}\{(X_{j,1} - X_{i,1})(X_{j,2} - X_{i,2}) > 0\} - 1 = \begin{cases} 1 & \text{if } (i,j) \text{ is a concordant pair,} \\ -1 & \text{if } (i,j) \text{ is a discordant pair.} \end{cases} \tag{7.1}$$

A classification technique will allocate a given couple (i, j) into one of the two categories $\{1, -1\}$ (or "concordant versus discordant", equivalently), with a certain probability, given the value of the common covariate $\mathbf{Z}$.

Section 7.2 introduces a general regression-type approach for the estimation of conditional Kendall's tau. Some asymptotic results of consistency and asymptotic normality are stated. In Section 7.3, we explain how some machine learning techniques can be adapted to deal with our particular framework, and we detail the corresponding algorithms. A small simulation study compares the small-sample properties of all these algorithms in Section 7.4. In Section 7.5, these techniques are applied to European stock market data. We evaluate to what extent the dependence between pairs of European stock indices may evolve with respect to different covariates. All proofs have been postponed into appendices.

7.2 Regression-type approach

Typically, a regression-type model based on conditional Kendall's tau may be written as

$$\tau_{1,2|\mathbf{Z}=\mathbf{z}} = g_0(\mathbf{z}, \beta^*), \ \ \forall \mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^p, \tag{7.2}$$

for some finite dimensional parameter $\beta^* \in \mathbb{R}^{p'}$ and some function g_0. As a particular case, a single-index approach would be

$$\tau_{1,2|\mathbf{Z}=\mathbf{z}} = g(\boldsymbol{\psi}(\mathbf{z})^T\beta^*), \ \ \forall \mathbf{z} \in \mathcal{Z}, \tag{7.3}$$

where $\boldsymbol{\psi} : \mathbb{R}^p \to \mathbb{R}^{p'}$ is known, and g may be known (parametric model) or not (semi-parametric model), as in [39]. In this section, we propose an inference procedure of β^* under (7.3) when the link function g is analytically known. This procedure will be based on the signs of pairs only, and not on the specific values of the vectors $\mathbf{X}_i$. Then, since inference will be based on the observations of $W \in \{1, -1\}$, our model belongs to the family of limited-dependent variable methods. One difficulty will arise from the pointwise conditioning events $\mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}$, that will necessitate localization techniques. Actually, we will consider couples of observations $\mathbf{X}_i$ and $\mathbf{X}_j$ for which the associate covariates are close to a given value $\mathbf{z}$. Indeed, the relationship (7.3) does not define the dependence levels between every couple $(\mathbf{X}_i, \mathbf{Z}_i)$ and $(\mathbf{X}_j, \mathbf{Z}_j)$, $i \neq j$, but only between those that share the same covariate. If the variables $\mathbf{Z}$ were discrete, we would consider a subset of couples such that $\mathbf{Z}_i = \mathbf{Z}_j$. In our case of continuous variables $\mathbf{Z}$ (see below), the latter event does not occur almost surely, and some smoothing/localization techniques have to be invoked.

Let K be a p-dimensional kernel and (h_n) be a bandwidth sequence. The bandwidth will simply be denoted by h and we set $K_h(\mathbf{z}) = K(\mathbf{z}/h)/h^p$. The log-likelihood associated to the observation $(W_{(i,j)}, \mathbf{Z}_i, \mathbf{Z}_j)$ given $\mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}$ is

$$\ell_\beta(W_{(i,j)}, \mathbf{z}) := \left(\frac{1+W_{(i,j)}}{2}\right) \log \mathbb{P}_\beta\Big(W_{(i,j)} = 1 \Big| \mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}\Big) + \left(\frac{1-W_{(i,j)}}{2}\right) \log \mathbb{P}_\beta\Big(W_{(i,j)} = -1 \Big| \mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}\Big).$$

In practice, when the underlying law of $\mathbf{Z}$ is continuous, there is virtually no couple for which $\mathbf{Z}_i = \mathbf{Z}_j$. Therefore, we will consider a localized "approximated" log-likelihood, based on $(W_{(i,j)}, \mathbf{Z}_i, \mathbf{Z}_j)$ for all pairs (i, j), $i \neq j$. It will be defined as the double sum

$$L_n(\beta) := \frac{1}{n(n-1)} \sum_{i,j;i\neq j} K_h(\mathbf{Z}_i - \mathbf{Z}_j)\ell_\beta(W_{(i,j)}, \tilde{\mathbf{Z}}_{i,j}),$$

for any choice of $\tilde{\mathbf{Z}}_{i,j}$ that belongs to a neighborhood of $\mathbf{Z}_i$ or $\mathbf{Z}_j$. We will assume that K is a compactly supported p-dimensional kernel of order $m \geq 2$.

The most obvious choices would be to select $\tilde{\mathbf{Z}}_{i,j}$ among $\{\mathbf{Z}_i, \mathbf{Z}_j, (\mathbf{Z}_i + \mathbf{Z}_j)/2\}$. Here, we propose

$$\begin{aligned} L_n(\beta) &:= \frac{1}{n(n-1)} \sum_{i,j;i\neq j} K_h(\mathbf{Z}_i - \mathbf{Z}_j)\ell_\beta(W_{(i,j)}, \mathbf{Z}_i) \\ &= \frac{1}{n(n-1)} \sum_{i,j;i\neq j} K_h(\mathbf{Z}_i - \mathbf{Z}_j) \left\{ \left(\frac{1+W_{(i,j)}}{2}\right) \log\left(\frac{1}{2} + \frac{1}{2} g(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)\right) \right. \\ &+ \left. \left(\frac{1-W_{(i,j)}}{2}\right) \log\left(\frac{1}{2} - \frac{1}{2} g(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)\right) \right\}, \end{aligned}$$

under (7.3). We can therefore derive an estimator of β^* based on the maximization of the latter function, with a ℓ_1 penalty (Lasso-type estimator), as

$$\hat{\beta} := \arg\max_{\beta \in \mathbb{R}^{p'}} L_n(\beta) - \lambda_n |\beta|_1, \tag{7.4}$$

where λ_n (also simply denoted as λ) is a tuning parameter to be chosen. Note that $L_n(\beta)$ is not really a likelihood function since the observations $(W_{(i,j)}, \mathbf{Z}_i, \mathbf{Z}_j)$ for every couple (i,j), $i \neq j$, are not mutually independent.

If $K \geq 0$, the objective function is a concave function of β if it satisfies

$$\delta g''(t)(1+\delta g)(t) \leq (g')^2(t), \ \forall t, \tag{7.5}$$

for $\delta \in \{1, -1\}$. When $\beta \mapsto L_n(\beta)$ is concave, the penalized criterion above is concave too and the calculation of $\hat{\beta}$ can be led in practical terms through convex optimization routines, even with a large number of regressors ($p' \gg 1$). Since this will be our framework, we will show that (7.5) holds for some usual classification techniques. When it is not the case, we have to rely on other optimization schemes and to avoid considering too many regressors.

Moreover, note that, when g is odd (i.e. $g(-t) = -g(t)$), the estimator simply becomes

$$\hat{\beta} := \arg\max_{\beta \in \mathbb{R}^{p'}} \frac{1}{n(n-1)} \sum_{i,j;i\neq j} K_h(\mathbf{Z}_i - \mathbf{Z}_j) \log\left(\frac{1}{2} + \frac{1}{2} g(W_{(i,j)} \boldsymbol{\psi}(\mathbf{Z}_i)^T \beta)\right) - \lambda |\beta|_1. \tag{7.6}$$

The implementation of an algorithm to solve problem (7.4) or its simplified version (7.6) may seem difficult due to the non-differentiability of the l_1 norm. Nevertheless, as in the case of the ordinary Lasso, it can be solved in a very efficient way using the Alternative Direction Method of Multipliers (ADMM) for general l_1 minimization, following [24, Section 6.1]. More precisely, assume $L_n(\beta)$ is a concave and differentiable function of β (this is the case in both Examples 7.1 and 7.2). Then the optimization task (7.4) can be rewritten as finding the solution $(\mathbf{x}, \mathbf{z}) \in \mathbb{R}^{2p'}$ of

$$\begin{cases} \text{minimize } f(\mathbf{x}) + g(\mathbf{z}) \\ \text{subject to } \mathbf{x} - \mathbf{z} = 0, \end{cases} \tag{7.7}$$

where $f(\mathbf{x}) := -L_n(\mathbf{x})$ and $g(\mathbf{z}) := \lambda_n |\mathbf{z}|_1$. The solution is given by iterating the following algorithm, denoting by $\mathbf{u} \in \mathbb{R}^{p'}$ the dual variable of the problem (7.7) and by $\rho > 0$ the step size (similarly to the usual gradient descent algorithm),

$$\begin{aligned} \mathbf{x}^{k+1} &:= \arg\min_{\mathbf{x}} \left(f(\mathbf{x}) + (\rho/2)||\mathbf{x} - \mathbf{z}^k + \mathbf{u}^k||_2^2\right), \\ \mathbf{z}^{k+1} &:= S_{\lambda_n/\rho}(\mathbf{x}^{k+1} + \mathbf{u}^k), \\ \mathbf{u}^{k+1} &:= \mathbf{u}^k + \mathbf{x}^{k+1} - \mathbf{z}^{k+1}, \end{aligned}$$

where for any $\kappa > 0$, S_κ is the element-wise soft thresholding operator, i.e. for each component $S_\kappa(a) := (1 - \kappa/|a|)_+ \times a$, for $a \neq 0$, and $S_\kappa(0) := 0$. Note that we have reduced the non-differentiable problem (7.4) into a sequence of differentiable optimization steps for $\mathbf{x}$, and the computation of the proximal operator S_κ for the $\mathbf{z}$-updates. We refer to [110] for a detailed presentation about proximal operators and their use in optimization. ADMM can also be adapted for large-scale data, using standard libraries and frameworks for parallel computing such as MPI, MapReduce and Hadoop, see [24] for more details about the implementation of such methods.

Example 7.1 (Logit)**.** *If we choose the Fisher transform $g(t) = (e^t - 1)/(e^t + 1)$, then g is odd and the optimization program becomes*

$$\hat{\beta} := \arg\max_{\beta \in \mathbb{R}^{p'}} \frac{1}{n(n-1)} \sum_{i,j;i\neq j} K_h(\mathbf{Z}_i - \mathbf{Z}_j) \log\left(logit(W_{(i,j)} \boldsymbol{\psi}(\mathbf{Z}_i)^T \beta)\right) - \lambda |\beta|_1,$$

where the so-called logit link function is defined by $logit(x) = e^x/(1+e^x)$*. Therefore* $\hat{\beta}$ *can be seen as the maximizer of the log-likelihood of a weighted logistic regression with independent realizations of an explained variable* $W_{(i,j)}$*, given some explanatory variables* $\mathbf{Z}_i$*. On a practical side, when* $K \geq 0$*, the* β*-criterion is concave. This allows to use the existing software and optimization routines of logistic regression without many changes.*

Example 7.2 (Probit)**.** *Similarly, choosing* $g(t) = 2\Phi(t) - 1$*, where* Φ *denotes the cdf of the standard normal distribution, yieds the equivalent of a (weighted) probit regression. Indeed, this function* g *is odd, (7.6) applies in this case and our criterion in (7.4) is concave wrt* β*.*

Let us assume that a family of models or some statistical procedure allow the calculation of the functional link $g(\epsilon\psi(\mathbf{z})^T\beta)$ and then $p(\mathbf{z})$, for any $\mathbf{z}$, $\epsilon \in \{-1, 1\}$ and any given value β: logit, probit, regression trees, neural networks, etc. Then, we can estimate the "true" parameter β^* by $\hat{\beta}$, as given by (7.4), in practical terms.

Now, we state the asymptotic properties of $\hat{\beta}$, under the assumption that $\beta \mapsto L_n(\beta)$ is concave. To this goal, we introduce some notation.

For any $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^p$, denote

$$p(\mathbf{x}, \mathbf{y}) := \mathbb{P}_{\beta^*}((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0 | \mathbf{Z}_1 = \mathbf{x}, \mathbf{Z}_2 = \mathbf{y}).$$

The latter expectations are calculated when the underlying parameter is assumed to be the true value β^*. Note that $p(\mathbf{x}) := p(\mathbf{x}, \mathbf{x})$ and $2p(\mathbf{z}) - 1 = \tau_{1,2|\mathbf{Z}=\mathbf{z}}$. Moreover, for any $\mathbf{x}, \mathbf{y}$ and $\mathbf{z} \in \mathbb{R}^p$, set

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) := \mathbb{P}_{\beta^*}((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0, (X_{3,1} - X_{1,1})(X_{3,2} - X_{1,2}) > 0 | \mathbf{Z}_1 = \mathbf{x}, \mathbf{Z}_2 = \mathbf{y}, \mathbf{Z}_3 = \mathbf{z}).$$

This is the conditional probability that $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ are concordant, given their respective covariates. Denote, for any $\beta \in \mathbb{R}^{p'}$,

$$\phi(\mathbf{x}, \mathbf{y}, \beta) := p(\mathbf{x}, \mathbf{y}) \log\left(q(\mathbf{x}, \beta)\right) + (1 - p(\mathbf{x}, \mathbf{y})) \log\left(1 - q(\mathbf{x}, \beta)\right), \; q(\mathbf{x}, \beta) := 1/2 + g(\boldsymbol{\psi}(\mathbf{x})^T\beta)/2.$$

Note that $q(\mathbf{z}, \beta^*) = p(\mathbf{z})$. Finally, for any real function f and $\varepsilon > 0$, denote by f_ε the function $x \mapsto \sup_{t,\, |x-t|<\varepsilon} |f(t)|$.

Regularity assumption R0: The density $f_{\mathbf{Z}}$ of $\mathbf{Z}$ is assumed to be m-times continuously differentiable. Moreover, the functions $\phi(\mathbf{x}, \cdot, \beta)$ and $q(\cdot, \beta)$ are continuous for any $\mathbf{x} \in \mathcal{Z}$ and any $\beta \in \mathbb{R}^{p'}$. To simplify, $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \mapsto p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ will be continuous on $\mathcal{Z}^3$.

Theorem 7.3. *Under R0, (7.14) and (7.15) in 7.8, if* $\lambda_n \to \lambda_\infty$ *and* $n^2h^p \to \infty$ *when* $n \to \infty$*, if the true model is given by (7.3) and* $\beta \mapsto L_n(\beta)$ *is concave, then the solution* $\hat{\beta}$ *of (7.4) tends in probability towards* $\beta^{**} := \arg\max_\beta L_\infty(\beta) - \lambda_\infty|\beta|_1$*, where*

$$L_\infty(\beta) := \int \phi(\mathbf{z}, \mathbf{z}, \beta) f_{\mathbf{Z}}^2(\mathbf{z})\, d\mathbf{z}.$$

In particular, when $\lambda_\infty = 0$, the estimator $\hat{\beta}$ tends to $\arg\max_\beta L_\infty(\beta) = \beta^*$, because $\phi(\mathbf{z}, \mathbf{z}, \beta)$ is the expected log-likelihood associated to $W_{(1,2)}$ given $\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}$. Thus, for every $\mathbf{z}$, the latter quantity is maximal when $\beta = \beta^*$.

Theorem 7.4. *Under the conditions of Theorem 7.3 and some additional conditions of regularity in 7.9 (notably (7.16), (7.17), (7.18) and (7.19)), if* $n^{1/2}\lambda_n \to \mu$ *and* $nh^p \to \infty$ *when* $n \to \infty$*, then* $n^{1/2}(\hat{\beta} - \beta^*)$ *weakly tends to*

$$\mathbf{u}^* = \arg\min_{\mathbf{u}\in\mathbb{R}^p} \mathbb{W}(\beta^*)\mathbf{u} + \frac{1}{2}\mathbf{u}^T\mathbb{H}(\beta^*)\mathbf{u} - \mu\sum_{k;\beta_k^*=0}|u_k| - \mu\sum_{k;\beta_k^*\neq 0} sign(\beta_k^*)u_k,$$

where $\mathbb{W}(\beta^*) \sim \mathcal{N}(0_p, \Sigma_{\beta^*})$*,* $\Sigma_{\beta^*} = \int \partial_\beta\phi(\mathbf{z},\mathbf{z},\beta^*)\partial_\beta\phi(\mathbf{z},\mathbf{z},\beta^*)^T f_{\mathbf{Z}}^3(\mathbf{z})\,d\mathbf{z}$ *and*

$$\mathbb{H}(\beta^*) = \int \partial^2_{\beta,\beta^T}\phi(\mathbf{z},\mathbf{z},\beta^*) f_{\mathbf{Z}}^2(\mathbf{z})\,d\mathbf{z}.$$

Remark 7.5. *All the previous results and those of the next sections are based on the kernel-weighted log-likelihood criterion* $L_n(\beta)$*, and then on the choice of the bandwidth* h*. We have not tried to find an "optimal" smoothing parameter* h*. This task is outside the scope of this paper and is left for further research. Instead, we have preferred to rely on the usual rule-of-thumb (Scott [122]), even if, strictly speaking, it is relevant only for kernel estimators of densities. Nonetheless, we have not empirically found an "excessive sensitivity" of our simulation results w.r.t.* h*.*

7.3 Classification algorithms and conditional Kendall's tau

In the latter section, we have studied a localized likelihood procedure to estimate β^* under (7.3), when we can explicitly write (and code) the link function g. This may be seen as a restrictive approach, because it is far from obvious to guess the right functional form of g. To improve the level of flexibility of our conditional Kendall's tau model, we recall the estimation of $\tau_{1,2|\mathbf{z}}$ is similar to the evaluation of $\mathbb{P}\big(W_{(1,2)} = 1\big|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big)$, i.e. the probability $p(\mathbf{z})$ of classifying the couple $(1,2)$ into one of two categories (concordant or discordant), given a common value $\mathbf{z}$ of their covariates. Formally, the answer of such a question can be directly yielded by some classification algorithms. This is the topic of this section. Therefore, instead of estimating an assumed parametric model by penalization, as in (7.4), a classification algorithm will "automatically" evaluate $p(\mathbf{z})$ by $\hat{p}(\mathbf{z})$. An estimator of the conditional Kendall's tau will simply be $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} := 2\hat{p}(\mathbf{z}) - 1$.

Now, we show how different classification algorithms can be used and adapted to the estimation of $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ in practice. The first step is to transform the dataset $\mathcal{D} = (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n} \in \big(\mathbb{R}^{2+p}\big)$, called the *initial dataset*, into an object $\tilde{\mathcal{D}}$, that will be called the *dataset of pairs* (see Algorithm 7). Each element of this dataset of pairs is indexed by an integer $k \in \{1,\dots,n(n-1)/2\}$, which corresponds to any (unordered) pair (i,j), $i \neq j$, of observations in the initial dataset.

For any pair of observations, we compute the associated covariate $\tilde{Z}_k$ which is just the average of the two covariates $\mathbf{Z}_i$ and $\mathbf{Z}_j$ (contrary to Section 7.2 where we have chosen $\mathbf{Z}_i$). Note that we want $\mathbf{Z}_i$ and $\mathbf{Z}_j$ to be close to each other, so that the pair (i,j) is relevant. This means that a weight variable V_k is defined for any pair. It is related to the proximity between $\mathbf{Z}_i$ and $\mathbf{Z}_j$. Obviously, if $V_k = 0$ then the corresponding pair is not kept, finally. This selection induces also a computational benefit, by reducing the size of the dataset and the computation time. For example, suppose that $n = 4000$. Then, up to around 8×10^6 possible pairs can be constructed but only a small group of them (around 10^4 or 10^5 pairs, typically) will be relevant. The others are pairs for which the covariates are considered too far apart. Note that, in order to increase the proportion of k such that the weight V_k is zero, it is sufficient to use compactly supported kernels. For instance, for any arbitrary p-dimensional kernel K, we can consider $\tilde{K}(\mathbf{z}) := \gamma K(\mathbf{z})\mathbb{1}\{\|\mathbf{z}\|_\infty \leq 1\}$, with some normalizing constant γ so that $\int \tilde{K} = 1$.

Algorithm 7: Algorithm for creating the dataset of pairs from the initial dataset.

Input: Initial dataset $\mathcal{D} = (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n} \in \left(\mathbb{R}^{2+p}\right)^n$;
$k \leftarrow 0$;
for $i \leftarrow 1$ **to** $(n-1)$ **do**
  for $j \leftarrow (i+1)$ **to** n **do**
    $\tilde{\mathbf{Z}}_k \leftarrow (\mathbf{Z}_i + \mathbf{Z}_j)/2$;
    $W_k \leftarrow W_{(i,j)}$ as defined in Equation (7.1) ;
    $V_k \leftarrow K_h(\mathbf{Z}_i - \mathbf{Z}_j)$;
    $k \leftarrow k+1$;
  end
end
Define $\mathcal{K} := \{k : V_k > 0\}$;
Output: A dataset of pairs $\tilde{\mathcal{D}} := (W_k, \tilde{\mathbf{Z}}_k, V_k)_{k\in\mathcal{K}} \in \left(\{-1,1\} \times \mathbb{R}^p \times \mathbb{R}_+\right)^{n(n-1)/2}$.

7.3.1 The case of probit and logit classifiers

With the new dataset $\tilde{\mathcal{D}}$, we can virtually apply any classification method to predict the concordance value W_k of the pair k, given the covariate $\tilde{\mathbf{Z}}_k$ and the weight V_k. The logit and probit models yield some of the oldest and easiest methods in classification. They have straightforward adapted versions in our case: see Algorithm 8. These weighted penalized GLM procedures are estimated using the R package `ordinalNet` [143]. Note that we are still estimating $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ under the parametric model given by (7.3). The tuning parameter λ can be chosen using a generalization of Algorithm 2 in Derumigny and Fermanian [39]. The chosen λ is the one which minimizes the cross-validation criterion,

$$CV(\lambda) := \sum_{k=1}^{N} d\left(\mathbf{z} \mapsto \hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}} \,;\, \mathbf{z} \mapsto g\big(\boldsymbol{\psi}(\mathbf{z})^T \hat{\beta}^{(\lambda,-k)}\big)\right),$$

where $d(\cdot\,;\,\cdot)$ is a distance on a space of bounded functions of $\mathbf{z}$, for example the distance generated by the L_2 norm, $\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\cdot}$ is an estimator of Kendall's tau using the dataset $\mathcal{D}_k$, $\hat{\beta}^{(\lambda,-k)}$ is estimated on the dataset $\mathcal{D}\backslash\mathcal{D}_k$ using the tuning parameter λ, and the initial dataset $\mathcal{D}$ has been separated at random in N subsets $\mathcal{D}_1, \dots, \mathcal{D}_N$ of equal size.

Algorithm 8: Estimation of the conditional Kendall's tau $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ using a logit (resp. probit) regression.

Input: A dataset of pairs $\tilde{\mathcal{D}} := (W_k, \tilde{\mathbf{Z}}_k, V_k)_{k\in\mathcal{K}}$
Input: A point $\mathbf{z} \in \mathcal{Z}$, a function $\boldsymbol{\psi}$ and a penalty level λ;
Compute the usual weighted penalized logit (resp. probit) estimator $\hat{\beta}$ on the dataset $(W_k, \boldsymbol{\psi}(\tilde{\mathbf{Z}}_k), V_k)_{k\in\mathcal{K}}$ with a tuning parameter λ ;
Output: An estimator $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} := (e^{\boldsymbol{\psi}(\mathbf{z})^T\hat{\beta}} - 1)/(e^{\boldsymbol{\psi}(\mathbf{z})^T\hat{\beta}} + 1)$
(resp. $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} := 2\Phi(\boldsymbol{\psi}(\mathbf{z})^T\hat{\beta}) - 1$).

7.3.2 Decision trees and random forests

Now, let us discuss how partition-based methods can be used for the estimation of the conditional Kendall's tau. Strictly speaking, such techniques are parametric: the relationship (7.2) implicitly applies,

but for some complex untractable function g. And the parameter β^* is related to some covariate thresholds, typically. Nonetheless, a classical decision tree can be directly trained on the weighted dataset $\tilde{\mathcal{D}}$. We use the R package `tree` by Ripley [117], following Breiman et al. [25]. Therefore, the application of decision trees to our framework is straightforward, and does not require any special adaptation, contrary to random forests. And the tree procedure allows the calculation of the probability of observing a concordant pair, given any common value of $\mathbf{Z}$.

In a classical classification setting, random forests are techniques of aggregation of decision trees that are built on a subset of samples and subsets of variables. More precisely, a typical random forest algorithm is the following: sample 80% of the rows of the dataset (without replacement), and 80% of the explanatory variables; estimate a tree on this, and repeat this procedure a certain number of times, with different sub-samples every time. In our framework, it is not clear at which level subsampling should take place.

The easiest solution would be to directly plug-in the dataset of pairs $\tilde{\mathcal{D}}$ into a classical random forest algorithm, but it does not obviously lead to the best solution. For comparison, we detail this solution in Algorithm 9. We propose now an improvement on Algorithm 9. Indeed, noting that aggregation of trees is useless if all trees are identical, it seems that the more variability in the input of the trees, the better. Following this idea, we have noticed that the observations in the dataset of pairs are not independent. Influence of this lack of independence is discussed in a general setting in Section 7.3.5. For example, the pair $(1,2)$ is usually not independent of the pair $(1,3)$, because they both share the first observation $(X_{1,1}, X_{1,2}, \mathbf{Z}_1)$. Therefore, to increase the diversity of inputs in the different trees, we suggest to lead a first sampling $\mathcal{S}_j$ on the initial dataset, and then to build a dataset of pairs on the sampled observations $\mathcal{D}_j := (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i\in\mathcal{S}_j}$ (see Algorithm 10). As a matter of fact, if for example the first observation does not belong to the sample $\mathcal{S}_j$, then the dataset $\mathcal{D}_j$ and the estimated tree $\mathcal{T}_j$ become both independent of this first observation $(X_{1,1}, X_{1,2}, \mathbf{Z}_1)$. This independence property makes the trees less dependent, and significantly improves the performance in our results compared to the original Algorithm 9.

Algorithm 9: Random forests un-adapted for the estimation of the conditional Kendall's tau

Input: Initial dataset $\mathcal{D} = (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n} \in \left(\mathbb{R}^{2+p}\right)^n$;

Compute the dataset of pairs $\tilde{\mathcal{D}}$ using Algorithm 7 on $\mathcal{D}$;

for $j \leftarrow 1$ **to** N_{tree} **do**

- Sample a set $\mathcal{S}_j \subset \{1,\dots,n(n-1)/2\}$ without replacement ;
- Compute the dataset of pairs $\tilde{\mathcal{D}}_j = (W_k, \tilde{Z}_k, V_k)_{k\in\mathcal{S}_j}$ using observations from $\tilde{\mathcal{D}}$;
- Sample a set $\mathcal{S}'_j \subset \{1,\dots,p'\}$ without replacement ;
- Estimate a classification tree $\mathcal{T}_j$ on the dataset $(W_k, (\psi_l(\tilde{Z}_k))_{l\in\mathcal{S}'_j}, V_k)_{k\in\mathcal{S}_j}$;

end

Output: An estimator $\hat{\tau}_{1,2|\mathbf{Z}=\cdot} := N_{tree}^{-1} \sum_{j=1}^{N_{tree}} \mathcal{T}_j(\cdot)$.

Algorithm 10: Random forests adapted for the estimation of the conditional Kendall's tau

Input: Initial dataset $\mathcal{D} = (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n} \in \left(\mathbb{R}^{2+p}\right)^n$;
for $j \leftarrow 1$ **to** N_{tree} **do**
  Sample a set $\mathcal{S}_j \subset \{1, \dots, n\}$ without replacement ;
  $\mathcal{D}_j \leftarrow (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i \in \mathcal{S}_j}$;
  Compute the dataset of pairs $\tilde{\mathcal{D}}_j = (W_k, \tilde{Z}_k, V_k)_{k \in \mathcal{K}_j}$ using Algorithm 7 on $\mathcal{D}_j$, providing $\mathcal{K}_j$;
  Sample a set $\mathcal{S}'_j \subset \{1, \dots, p'\}$ without replacement ;
  Estimate a classification tree $\mathcal{T}_j$ on the dataset $(W_k, (\psi_l(\tilde{Z}_k))_{l \in \mathcal{S}'_j}, V_k)_{k \in \mathcal{K}_j}$;
end
Output: An estimator $\hat{\tau}_{1,2|\mathbf{Z}=\cdot} := N_{tree}^{-1} \sum_{j=1}^{N_{tree}} \mathcal{T}_j(\cdot)$.

7.3.3 Nearest neighbors

The nearest neighbors also provide a very popular classification algorithm and can be directly used on the dataset $\tilde{\mathcal{D}}$ (see Algorithm 11). Here, we no longer assume (7.3) or even (7.2), and we live in a nonparametric framework. A pretty difficult problem is to choose a convenient number of nearest neighbors. As usual in nonparametric statistics, we must find a compromise between variance (tendancy to undersmooth, i.e. to choose a too small N) and bias (tendancy to oversmooth, i.e. to choose a too big N). Moreover, in our case, with $n(n-1)/2$ possible pairs, choosing a right value for N can be challenging. Indeed, in the usual (i.i.d.) nearest neighbor framework, the asymptotically optimal N is a power of the sample size. Here, this is different because there are three potential sample sizes: n, if we consider there are fundamentally n sources of randomness, $n(n-1)/2$ by considering that the new sample has a cardinality equal to the number of pairs, or even $|\mathcal{K}|$ that is random and depends on h. Thus, our problem is to choose a "relevant formula" for N based on the "convenient" sample size.

Algorithm 11: Estimation of the conditional Kendall's tau $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ using nearest neighbors.

Input: A dataset of pairs $\tilde{\mathcal{D}} := (W_k, \tilde{\mathbf{Z}}_k, V_k)_{k \in \mathcal{K}}$
Input: A point $\mathbf{z} \in \mathcal{Z}$, a number N of nearest neighbors and a distance d on $\mathbb{R}^{p'}$;
$\mathcal{K}_{\mathbf{z}} \leftarrow \arg\min_{E \subset \mathcal{K}, |E|=N} \left(\sum_{k \in E} d\big(\boldsymbol{\psi}(\mathbf{z}), \boldsymbol{\psi}(\tilde{\mathbf{Z}}_k)\big) \right)$;
Output: An estimator $\hat{\tau}^{(N)}_{1,2|\mathbf{Z}=\mathbf{z}} := \big(\sum_{k \in \mathcal{K}_{\mathbf{z}}} V_k W_k \big) / \sum_{k \in \mathcal{K}_{\mathbf{z}}} V_k$.

In applications, one might not be interested in the value of the conditional Kendall's tau at only one point, but also in the whole function $\mathbf{z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$. The goodness of this estimation is linked to the underlying density $f_{\mathbf{Z}}$ of $\mathbf{Z}$: the estimation can be made more precise in regions where $f_{\mathbf{Z}}$ is high, allowing to use a higher number of neighbors with close covariates. At the opposite, in some regions where $f_{\mathbf{Z}}$ is low, a smaller N should be used. Note that, in general, $f_{\mathbf{Z}}$ is unknown and its estimation may be difficult as well, due to the curse of dimensionality. Therefore, it is highly desirable to build a local number of neighbors $N(\mathbf{z})$. Such a local choice $N(\mathbf{z})$ will help to avoid both under- and over-smoothing in all parts of the space $\mathcal{Z}$.

Cross-validation techniques are widely use for the choice of tuning parameters, but might not be here the best solution as one would like to find a local choice of N. This problem has similarities with classical non-parametric regression. We propose to use a procedure inspired by Lepski's method for choosing the bandwidth [95], once adapted to our setting. Lepski's method is built on a simple principle: when two

non-parametric estimators are close, the best is the smoothest. When two non-parametric estimators are far apart, the best is the least smooth. Let $(\mathcal{Z}_i)_{i\in\mathcal{I}}$ be a partition of $\mathcal{Z}$. The goal will be to choose the best estimator on each $\mathcal{Z}_i$, which corresponds to the choice of a local number of nearest neighbors N_i. This procedure is called "local" since the diameters of the $\mathcal{Z}_i$ will be small. For example, if $p = 1$ and $\mathcal{Z}$ is a bounded interval then the $\mathcal{Z}_i$ can be chosen as small intervals. We denote by $\mathcal{N} \subset \mathbb{N}$ the finite set of possible numbers of neighbors. Following Lepski's approach, we choose $\mathcal{N}$ as a geometric progression, i.e. $\mathcal{N} = \{\lfloor a_1 \times a_2^i \rfloor,\, i = 1, \dots, i_{max}\}$ for some constants $a_1, a_2 > 0$, where $\lfloor x \rfloor$ denotes the integer part of a real x.

To measure how far the estimators are from each other, we introduce a distance d_i, $i \in \mathcal{I}$. As our estimators of conditional Kendall's tau are bounded (between -1 and 1) and measurable, several choices are possible. In applications, we will use

$$d_i(f,g) = \left(\frac{1}{j_{max}} \sum_{j=1}^{j_{max}} \left[(f(\mathbf{z}_{i,j}) - g(\mathbf{z}_{i,j}))/M \right]^2 \right)^{1/2}, \mathbf{z}_{i,1}, \dots, \mathbf{z}_{i,j_{max}} \in \mathcal{Z}_i, \tag{7.8}$$

where M is a normalization factor independent of i and the subsets $\mathbf{z}_{i,1}, \dots, \mathbf{z}_{i,j_{max}}$ are arbitrarily chosen in $\mathcal{Z}_i$, $i = 1, \dots, i_{max}$. We will use $M = (\max - \min)\{\hat{\tau}^{(N)}_{1,2|\mathbf{Z}=\mathbf{z}}, N \in \mathcal{N}, \mathbf{z} \in \mathcal{Z}\}$. Indeed, in the classical nonparametric regression model $Y = f(X) + \varepsilon$, with an unknown function f, M should be replaced by the standard deviation of the noise ε. In our case, we can define a (pseudo-)noise $\xi_{\mathbf{z},N} := \hat{\tau}^{(N)}_{1,2|\mathbf{Z}=\mathbf{z}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}}$, but it is unknown in practice and its distribution is complicated. Therefore M serves as a proxy of the amplitude of the variations in the estimated conditional Kendall's tau. This normalization by M ensures a kind of adaptativity of the estimation.

Algorithm 12: Lepski's method for a local choice of the number of nearest neighbors, and the corresponding estimator of the conditional Kendall's tau.

Input: A set $\mathcal{N} \subset \mathbb{N}$ of possible numbers of nearest neighbors and the corresponding estimates $\hat{\tau}^{(N)}_{1,2|\mathbf{Z}=\cdot}$ given by Algorithm 11, for all $N \in \mathcal{N}$;

Input: A partition $(\mathcal{Z}_i)_{i\in\mathcal{I}}$ of $\mathcal{Z}$ and a distance d_i on a space of bounded measurable real functions defined on $\mathcal{Z}_i$, for every $i \in \mathcal{I}$;

foreach $i \in \mathcal{I}$ **do**

$$S_i \leftarrow \left\{ N \in \mathcal{N} : d_i\left(\hat{\tau}^{(N)}_{1,2|\mathbf{Z}=\cdot}, \hat{\tau}^{(N')}_{1,2|\mathbf{Z}=\cdot}\right) \le A\sqrt{(1/N')\log(\max(\mathcal{N})/N')}, \forall N' \in \mathcal{N} \cap [1, N] \right\}; \tag{7.9}$$

$N_i \leftarrow \max(S_i)$;

end

Output: An estimator $\mathbf{z} \mapsto \hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} := \sum_{i\in\mathcal{I}} \mathbb{1}\{\mathbf{z} \in \mathcal{Z}_i\} \hat{\tau}^{(N_i)}_{1,2|\mathbf{Z}=\mathbf{z}}$.

We have observed that the sensitivity to $\mathcal{N}$ is not too large, if it is chosen in a reasonable way, for example between 5 or 10 possibilities. When $\mathbf{Z}$ is univariate, a simple partition $(\mathcal{Z}_i)_{i\in\mathcal{I}}$ can be given by the deciles of $\mathbf{Z}$. We choose $A = 1$ for simplicity since we believe there is no procedure for choosing it. A statistician who would like to play with the smoothness of the result is free to adjust the constant A, using an expert knowledge of the situation. Finally, the $\mathbf{z}_{i,j}$ can be chosen as quantiles of $\mathbf{Z}$, or as a regular grid on each $\mathcal{Z}_i$.

7.3.4 Neural networks

Nowadays, neural networks have become very popular with a wide range of applications. In classification problems, a neural network can be seen as an estimator that depends on some parameters, but in a very flexible and complex way. For every input $\mathbf{z}$, it yields the probability of belonging to any class. In our framework, we will train a network on the dataset of pairs $\tilde{\mathcal{D}}$. It is well-known that most neural networks do not induce convex programs, and the outputs therefore depend on some initial parameter values. One strategy is to independently train networks with different starting parameter values, that may be randomly chosen, for example.

This method of using independent estimators (conditionally on the initial sample $\mathcal{D}$) and then aggregating them is related to the random forest approach of the previous section and the discussion therein. Therefore, the same techniques are relevant and we have noticed an improvement in terms of performance by using an adapted version of Algorithm 10. More precisely, we fix a number of neural networks. For each neural network, we sample without replacement a part of the initial dataset from which the corresponding dataset of pairs is constructed and used as a training set. In order to improve stability, we aggregate the predictions of the different neural networks by using their median as the final predicted Kendall's tau. There is a trade-off between computation time and accuracy: a larger number of networks should improve the accuracy while taking obviously a longer time to be trained. The precise choice of the best architecture of the network is a complicated task, which is left for future research. As we are looking for functions $\mathbf{z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$ which are smooth almost everywhere and easy to interpret in applications, we choose a simple architecture with $N_{nnet} = 10$ neural networks, each having a single hidden layer of 3 neurons. Besides, bigger networks seem to deteriorate the performance of this estimator, see Section 7.4.6.

Algorithm 13: Neural networks with median bagging, adapted for the estimation of the conditional Kendall's tau

Input: Initial dataset $\mathcal{D} = (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n} \in \left(\mathbb{R}^{2+p}\right)^n$;
for $j \leftarrow 1$ **to** N_{nnet} **do**
  Sample a set $\mathcal{S}_j \subset \{1, \dots, n\}$ without replacement ;
  $\mathcal{D}_j \leftarrow (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i \in \mathcal{S}_j}$;
  Compute the dataset of pairs $\tilde{\mathcal{D}}_j = (W_k, \tilde{\mathbf{Z}}_k, V_k)_{k \in \mathcal{K}_j}$ using Algorithm 7 on $\mathcal{D}_j$, providing $\mathcal{K}_j$;
  Estimate a neural net $\mathfrak{N}_j$ on the dataset $(W_k, \boldsymbol{\psi}(\tilde{\mathbf{Z}}_k), V_k)_{k \in \mathcal{K}_j}$;
end
Output: An estimator $\hat{\tau}_{1,2|\mathbf{Z}=\cdot} := Median\{\mathfrak{N}_j(\cdot), j = 1, \dots, N_{nnet}\}$.

7.3.5 Lack of independence and its influence on the proposed algorithms

The machine learning methods that are discussed in this section were all designed for i.i.d. data. But it is easy to see that some observations in the dataset of pairs $\tilde{\mathcal{D}}$ will not be independent. Indeed, assume that the observations in the original dataset $(X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n}$ are i.i.d., to simplify. The pair $(i = 1, j = 2)$ and the pair $(i = 1, j = 3)$ both involve the first observation $(X_{1,1}, X_{1,2}, \mathbf{Z}_1)$, and therefore are not independent. This is a theoretical problem, but numerical results in Section 7.4 show that this does not often seem to be a problem in practice.

As far as the logit and probit are concerned, it was proved in the previous Section 7.2 that they are related to a family of estimators that can use $\tilde{\mathcal{D}}$ "as is". They yield consistent and asymptotically normal estimates, nonetheless, if the specification is correct. It is likely that the other methods presented here enjoy similar properties and are also largely unaffected by dependence between pairs. Note that, if all observations in $\mathcal{D}$ are identically distributed, then the observations in $\tilde{\mathcal{D}}$ are identically distributed as well. This is favourable to our methods.

Concerning the dependence inside $\tilde{\mathcal{D}}$, we will show that it is not too strong. For example, the pairs $(1,2)$ and $(1,3)$ are not independent, but the pairs $(1,2)$ and $(3,4)$ are indeed independent. This means that there is still "a large proportion of" independence left in $\tilde{\mathcal{D}}$. Formally, if two distinct pairs are randomly chosen in $\tilde{\mathcal{D}}$, the probability that they are really independent is high. Indeed, there are $N_{tot} := n(n-1)(n(n-1)-2)/8$ couples of distinct pairs. Beside, the number N_{ind} of couples of pairs which are independent is $N_{ind} := n(n-1)(n-2)(n-3)/8$. The factor $1/8$ appears in both N_{tot} and N_{ind} since we can always switch the two observations in the first pair, in the second pair, and switch the two pairs (every 4-tuple is counted $2^3 = 8$ times). It is easy to see that $N_{ind}/N_{tot} = 1 - O(1/n)$ as $n \to \infty$.

This means that the pairs are "almost all" independent from each other, as $n \to \infty$. In other words, the dependence between two pairs become negligible with averages. That is the reason why the machine learning methods used will perform well if the original dataset $\mathcal{D}$ is large enough. If the original dataset $\mathcal{D}$ is not i.i.d., for example as observations of a time series, we conjecture that such methods will work in a similar way as long as dependence is not too strong, for example if the data-generating process satisfies some usual assumptions, see Remark 7.6.

Whenever bootstrap, subsetting, resampling, or cross-validation is led on these classification-based estimators, we advise to perform them on the original dataset $\mathcal{D}$ rather than on the dataset of pairs $\tilde{\mathcal{D}}$, as we did in Sections 7.3.1, 7.3.2 and 7.3.4. This seems to yield a good improvement in performance. An example is given by the difference between Algorithms 9 and 10. This can be simply summed up as *"do the resampling on the original dataset $\mathcal{D}$, not on the transformed dataset $\tilde{\mathcal{D}}$"*. Nevertheless, a complete study and justification of this general principle is beyond the scope of this paper and is left for future work.

7.4 Simulation study

In this section, we have studied the relative performances of our estimators by simulation. For a given model and a given method of estimation, we sample 100 different experiments, and estimate the model for each sample. We fix the sample size as $n = 3000$. We remark that, for a given dimension $p > 0$ of $\mathbf{Z}$ and a given support $\mathcal{Z}$ of $\mathbf{Z}$, we have different "blocks" of the model which can be chosen in an independent way:

(i) the law $\mathbb{P}_{\mathbf{Z}}$ of $\mathbf{Z}$,

(ii) the function $\mathbf{z} \in \mathcal{Z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$,

(iii) the (conditional) copula family $(C_\tau)_{\tau\in(0,1)}$ or $(C_\tau)_{\tau\in(-1,1)}$ of $(X_1, X_2)|\mathbf{Z} = \mathbf{z}$, indexed by its conditional Kendall's tau - for example the Gaussian, Student, Clayton, Gumbel, etc, copula families. Such a family can also depend on $\mathbf{Z}$: for example, think of a Student copula with varying degrees of freedom -,

(iv) the conditional margins $X_1|\mathbf{Z}$ and $X_2|\mathbf{Z}$,

(v) the choice of the functions ψ_i, for $i = 1, \dots, p'$,

(vi) the choice of the estimator $\hat{\tau}_{1,2|\mathbf{Z}=\cdot}$.

Our so-called "reference setting" will be defined as $p = 1$, $\mathcal{Z} = [0,1]$ and (i) $\mathbb{P}_Z = \mathcal{U}_{[0;1]}$; (ii) $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = 3z(1-z)$; (iii) $(C_\tau)_{\tau\in(0,1)}$ is the Gaussian Copula family ; (iv) $\mathbb{P}_{X_1|Z=z} = \mathbb{P}_{X_2|Z=z} = \mathcal{N}(z,1)$. For each tested model, the performance of the estimator will be evaluated by the mean integrated ℓ_2 error. With obvious notation, it will be estimated as

$$Err := \mathbb{E}\left[\int_{\mathcal{Z}} (\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}})^2 d\mathbf{z}\right] \approx \frac{1}{N_{simu} N_{points}} \sum_{i=1}^{N_{simu}} \sum_{j=1}^{N_{points}} (\hat{\tau}^{(i)}_{1,2|\mathbf{Z}=\mathbf{z}^{(j)}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}^{(j)}})^2, \quad (7.10)$$

where N_{simu}, N_{points} are positive integers, $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(N_{points})}$ are fixed points in $\mathcal{Z}$, and $\hat{\tau}^{(i)}_{1,2|\mathbf{Z}=\mathbf{z}^{(j)}}$ is the estimated conditional Kendall's tau at point $\mathbf{z}^{(j)}$ trained on data from the i-th simulation. We choose $N_{simu} := 100$ experiments, and in this reference setting, the integral is discretized with $N_{points} := 100$ equispaced points on the segment $[0.01, 0.99]$, to avoid numerical problems at the boundaries.

In the following simulations, "logit" and "probit" refer to Algorithm 8. "Tree" refers to the application of the method `tree()` of package `tree` by Ripley [117] on the dataset $\tilde{\mathcal{D}}$ produced by Algorithm 7. "Random forests" refers to Algorithm 10. "Nearest neighbors" refers to the adapted version using Algorithm 11, once aggregated using Algorithm 12. Finally "Neural networks" refers to Algorithm 13. Such specifications are now part of our "reference setting".

7.4.1 Choice of the functions $\{\psi_i\}, i = 1, \dots, p'$.

We consider six different choices of $\boldsymbol{\psi}$, that are

1. No transformation, i.e. $\psi_1^{(1)}(z) = z$.
2. Polynomials of degree lower than 4: $\psi_i^{(2)}(z) = 2^{-i+1}(z-0.5)^{i-1}$ for $i = 1, \dots, 5$.
3. Polynomials of degree lower than 10: $\psi_i^{(3)}(z) = 2^{-i+1}(z-0.5)^{i-1}$ for $i = 1, \dots, 11$.
4. Fourier basis of order 2 with an intercept: $\psi_1^{(4)}(z) = 1$, $\psi_{2i}^{(4)}(z) = \cos(2\pi i z)$ and $\psi_{2i+1}^{(4)}(z) = \sin(2\pi i z)$ for $i = 1, 2$.
5. Fourier basis of order 5 with an intercept: $\psi_1^{(5)}(z) = 1$, $\psi_{2i}^{(5)}(z) = \cos(2\pi i z)$ and $\psi_{2i+1}^{(5)}(z) = \sin(2\pi i z)$ for $i = 1, \dots, 5$.
6. Concatenation of $\boldsymbol{\psi}^{(2)}$ and $\boldsymbol{\psi}^{(4)}$, which will be denoted by $\boldsymbol{\psi}^{(6)}$.

For each of the choices of $\boldsymbol{\psi}$ above, and each estimator, we compute the criterion (7.10). The results are displayed in the following Table 7.1.

With the choice of $\boldsymbol{\psi}^{(6)}$, logit and probit methods provide the best results. This good performance deteriorates with other choices of $\boldsymbol{\psi}$, especially when the model is misspecified. Neural networks provide the best results with $\boldsymbol{\psi}^{(1)}$, and their performance declines when further transformations of $\mathbf{z}$ are introduced in $\boldsymbol{\psi}$. Nearest neighbors have nearly the best behavior with $\boldsymbol{\psi}^{(1)}$, and it does not seem that other transformations can significantly increase its performance. On the contrary, for trees and random forests, it seems that bigger families $\boldsymbol{\psi}$ can yield improvements over $\boldsymbol{\psi}^{(1)}$.

From now on, we will choose $\boldsymbol{\psi}^{(6)}$ for the methods *logit*, *probit*, *tree* and *random forests*. Indeed, for these methods, this choice of $\boldsymbol{\psi}$ yields nearly the lowest error criterion and presents the advantages

Chosen ψ	Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
$\psi^{(1)}$	48.1	48.1	7.5	4.89	2.26	0.561
$\psi^{(2)}$	0.721	0.554	4.28	3.28	2.26	1.32
$\psi^{(3)}$	0.663	0.528	4.13	3.41	2.23	1.73
$\psi^{(4)}$	1.41	1.45	4.73	14.2	2.72	1.74
$\psi^{(5)}$	1.05	1.06	4.76	10.3	2.79	2.67
$\psi^{(6)}$	0.456	0.434	4.57	3.15	2.64	3.87

Table 7.1: Error criterion (7.10) for each choice of ψ and each method, multiplied by 1000.

of proposing various shapes, which will help to combine the performances of both polynomials and oscillating functions. On the contrary, for the methods *nearest neighbors* and *neural networks*, we choose $\psi^{(1)}$ as adding new functions does not seem to increase the performance of both of these methods. Figure 7.1 displays a comparison of the different methods on a typical simulated sample.

7.4.2 Comparing different copulas families

Now, we keep the reference setting and we change only its part (iii), i.e. the functional form of the conditional copula. The results are displayed in Table 7.3. We observe that such choice of a parametric copula families has nearly no effect on the performance of the estimators. Nonetheless, with the Student copula (either with fixed or variable degrees of freedom), most estimators have slightly worse performances than with other copulas. This can be explained by the fact that this copula allows asymptotic dependence, i.e. a strong tail association.

Copula family	Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
Gaussian	0.456	0.434	4.57	3.15	2.26	0.561
Student 4 df	0.549	0.515	4.54	3.28	2.87	0.753
Student $(2+1/z)$ df	0.531	0.518	4.66	3.23	2.82	0.805
Clayton	0.498	0.472	4.52	3.36	2.67	0.742
Gumbel	0.45	0.431	4.56	3.23	2.66	0.775
Frank	0.448	0.42	4.5	3.28	2.13	0.615

Table 7.3: Error criterion (7.10) for each copula family and each method, multiplied by 1000.

7.4.3 Comparing different conditional margins

In this subsection, we still start from the reference setting and we change only its part (iv), i.e. the functional form of the conditional margins $(X_1|Z)$ and $(X_2|Z)$. We consider the following alternatives:

1. $\mathbb{P}_{X_1|Z=z} = \mathbb{P}_{X_2|Z=z} = \mathcal{N}(z,1)$ (as in the reference case).
2. $\mathbb{P}_{X_1|Z=z} = \mathcal{N}(\cos(10\pi z),1)$; $\mathbb{P}_{X_2|Z=z} = \mathcal{N}(z,1)$. The idea is to make X_1 oscillate fast enough so that the algorithms will have difficulties to localize concordant and discordant pairs ;
3. $\mathbb{P}_{X_1|Z=z} = Exp(|z|)$; $\mathbb{P}_{X_2|Z=z} = \mathcal{U}_{[z,z+1]}$. This choice allows to see how estimation is affected by changes in the conditional support of (X_1, X_2) given $\mathbf{Z} = \mathbf{z}$;

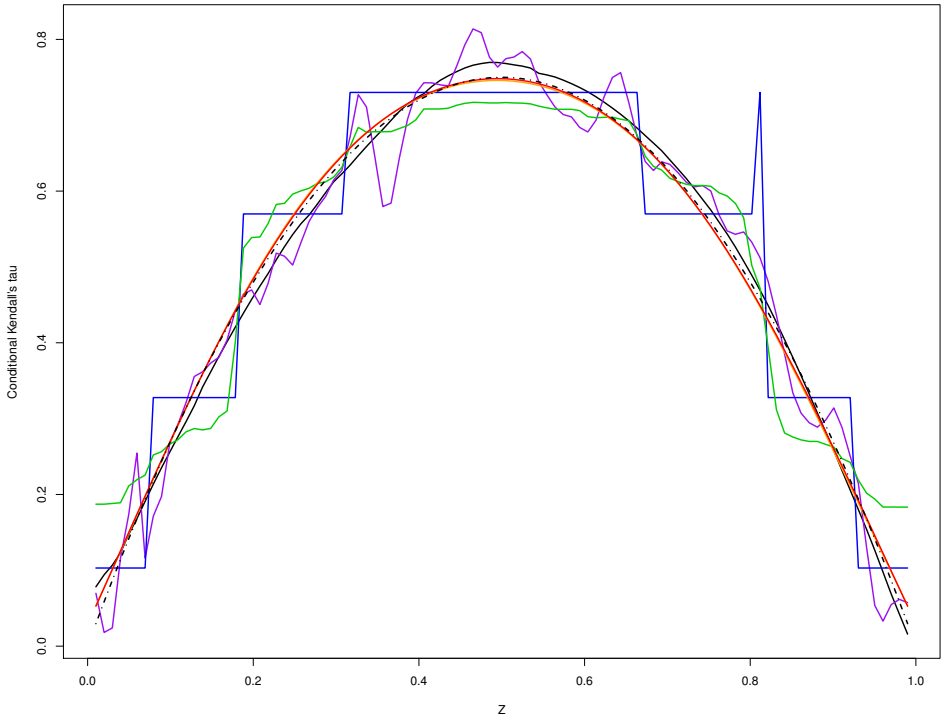


Figure 7.1: An example of the estimation of the conditional Kendall's tau using different estimation methods (see Table 7.2 below). The black dash-dotted curve is the true conditional Kendall's tau that has been used in the simulation experiment.

Method	Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
Algorithm	Algorithm 8	Algorithm 8	`tree()` of [117]	Algorithm 10	Algorithms 11-12	Algorithm 13
Color	orange	red	blue	green	purple	black

Table 7.2: Summary of available estimation methods for the estimation of the conditional Kendall's tau and corresponding algorithm and curve color.

For each estimator, we state in the second line of the above table the algorithm used to compute it, and in the third line the color of the corresponding curve on Figures 7.1 to 7.13. For example, the estimator "probit" is computed using Algorithm 8 and corresponds to the red curves.

4. $\mathbb{P}_{X_1|Z=z} = \mathcal{N}(0, z^2)$; $\mathbb{P}_{X_2|Z=z} = \mathcal{U}_{[0,|z|]}$. Then, we will see how estimation is affected by changes in the conditional variance of (X_1, X_2) given $\mathbf{Z} = \mathbf{z}$.

Setting	Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
1.	0.456	0.434	4.57	3.15	2.26	0.561
2.	0.809	0.818	4.65	3.72	2.65	0.838
3.	1.15	1.12	5.29	4.21	3.57	1.32
4.	0.493	0.471	4.43	3.44	2.54	0.662

Table 7.4: Error criterion (7.10) for each choice of conditional margins and each method, multiplied by 1000.

In a similar way as in the previous section, the results of these experiments, as displayed in Table 7.4 show that changes in terms of conditional marginal distributions generally have a mild impact on the overall performance of the estimators. Moreover, such changes have no effect on the ranking between estimators: the *logit* and *probit* methods are always the best, followed by the *neural networks*, the *nearest neighbors*, and the *random forests* are behind (in this order). The estimator *Tree* shows the lowest performance, but note that it also has the lowest computation time.

7.4.4 Comparing different forms for the conditional Kendall's tau

In this part, we keep the reference setting, but we change only its part (ii), i.e. the functional form of the conditional Kendall's tau itself. We consider the following choices:

1. $f_1(z) := 0.9 - 0.8\,\mathbb{1}\{z \geq 0.5\}$,
2. $f_2(z) := 3z(1-z)$,
3. $f_3(z) := 0.5 + 0.4\sin(4\pi z)$,
4. $f_4(z) := 0.1 + 1.6z\mathbb{1}\{z < 0.5\} + 1.6(z-0.5)\mathbb{1}\{u \geq 0.5\}$.

The results are presented in Table 7.5. If the estimated model is close to be well-specified, the best methods are parametric, i.e. the *logit* and *probit* regressions. In all the other cases, *neural networks* seem to perform very well. There appears a compromise between a minimization of the error and a minimization of the computation time. We refer to Table 7.8 for a quantitative comparison of the performance of such methods in terms of computation time as a function of the sample size n.

Setting	Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
f_1	11.2	11.6	4.12	4.03	3.89	1.48
f_2	0.456	0.434	4.57	3.15	2.26	0.561
f_3	3.77	3.22	5.95	4.76	2.35	2.17
f_4	12.8	12.8	16.8	10	3.71	1.97

Table 7.5: Error criterion (7.10) for different Kendall's tau models and each estimation method, multiplied by 1000.

7.4.5 Higher dimensional settings

In the previous sections, we had chosen a univariate vector $\mathbf{Z}$, i.e. $p = 1$. Since this may sound a bit restrictive, we would like to obtain some finite-sample results in dimension $p = 2$. Note that the latter dimension cannot be too high because of the curse of dimensionality linked with the necessary kernel smoothing (done in Algorithm 7 when creating the dataset of pairs). We also choose a simple dictionary $\boldsymbol{\psi}$ of functions, which consists of the two projections on the coordinates of $\mathbf{Z}$. The performance of the estimators is still be assessed by the approximate error criterion (7.10). The corresponding $\mathbf{z}^{(j)}$ are chosen as a grid of 400 points equispaced on the square $[0.01, 0.99]^2$.

In this framework, we first choose block (iii) of the model : the conditional copula of X_1 and X_2 given $\mathbf{Z}$ will be Gaussian, and block (iv) : $\mathbb{P}_{X_1|\mathbf{Z}=\mathbf{z}} = \mathbb{P}_{X_2|\mathbf{Z}=\mathbf{z}} = \mathcal{N}(z_1, 1)$. We will try different combinations for the remaining blocks (i) and (ii), as described as follows:

(1) $Z_1 \sim \mathcal{N}(0,1)$, $Z_2 \sim \mathcal{U}_{[-1,1]}$, and the copula of (Z_1, Z_2) is Gaussian with a Kendall's tau equal to 0.5. Moreover, $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = z_2 \tanh(z_1)$. This model is interesting because the function $\mathbf{z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$ will be far away from a linear function of $\boldsymbol{\psi}(\mathbf{z})$, and machine learning techniques should work better than logistic/probit regressions.

(2) We keep the same model as previously, but by setting $g(\tau_{1,2|\mathbf{Z}=\mathbf{z}}) = z_1 + z_2$, using the function g in Example 7.1 so that we recover the parametric setting of Section 7.2.

(3) $Z_1 \sim Exp(1)$, $Z_2 \sim \mathcal{N}(0,1)$ and both variables are independent. Set $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = \exp(-z_1|z_2|)$. Again, we have a misspecified nonlinear model that is far away from logit/probit models.

The results are given in Table 7.6. With the exception of the well-specified setting (2), the logit model performs worse than non-parametric methods. In all these settings, neural networks show better performances than all other methods, followed by nearest neighbors and tree-based methods. Finally, parametric methods are the worst, especially under misspecification of the model.

Setting	Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
(1)	35.5	35.5	9.63	11.7	6.72	2.21
(2)	0.433	0.681	10.9	5.85	4.33	0.848
(3)	17.8	17.2	5.72	9.79	1.84	1.36

Table 7.6: Error criterion (7.10) for each setting with 2-dimensional $\mathbf{Z}$ random vectors and each method, multiplied by 1000.

7.4.6 Choice of the number of neurons in the one-dimensional reference setting

We consider networks with different numbers of neurons, and study their performance, both statistically and computationally. The results are displayed in the following Table 7.7. We observe that increasing the number of neurons only seems to deteriorate the performance of the method.

7.4.7 Influence of the sample size n

In our one-dimensional reference setting, we fix all the parameters except n. For a grid of values of n we evaluate the performance of our estimators.

Number of neurons	3	5	10	30
Criteria	0.561	0.808	1.47	1.45
Time (s)	234	429	607	5.29e+03

Table 7.7: Error criterion (7.10) multiplied by 1000, and average computation time in seconds for each architecture of the neural networks.

		Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
$n = 1000$	Criteria	1.58	1.52	5.85	4.45	4.01	2.01
	Time (s)	59.6	156	0.215	8.11	5.04	30.6
$n = 2000$	Criteria	0.666	0.64	4.9	3.39	2.95	1.79
	Time (s)	192	489	0.99	35.9	17.1	85.3
$n = 3000$	Criteria	0.456	0.434	4.57	3.15	2.26	0.561
	Time (s)	414	1010	2.37	87	36.9	234
$n = 5000$	Criteria	0.275	0.253	3.77	3.05	1.69	0.791
	Time (s)	957	2420	6.37	218	111	461
$n = 8000$	Criteria	0.22	0.204	3.6	3.39	1.27	0.225
	Time (s)	2178	5480	15.2	499	290	1268

Table 7.8: Error criterion (7.10) multiplied by 1000 and computation time in seconds for each method and each choice of n.

We observe that, for most methods, the computation time increases and the error criterion decreases when the sample size increases. We note that the number of pairs is $O(n(n-1))$ (at most) and, therefore, the computation time should increase as $O(n^2)$, which is coherent with the results of Table 7.8. The relative order of the performances does not seem to change with the sample size n: the same methods are the best ones with small or large n. Note that we have not tried to find an "optimal" fine-tuning of the parameters for each method and each choice of n. Indeed, finding optimal choices of tuning parameters is not an easy task (in a theoretical and practical sense). More accurate analysis are left for future research.

7.4.8 Influence of the lack of independence

In Section 7.3.5, we explain some theoretical considerations about the lack of independence in the dataset $\tilde{\mathcal{D}}$ and some consequences. The following simulation experiment complements this analysis with some empirical results.

Indeed, using Algorithm 7, we note that pairs of observations are not independent, and therefore, the elements of the dataset of pairs $\tilde{\mathcal{D}}$ are not independent from each other in general. This could damage the performance of our methods, compared to a situation where all elements would be independent. We now consider such a situation, in order to compare the performance of the methods in both cases. Note that the cardinality of $\tilde{\mathcal{D}}$ is $n(n-1)/2$. Therefore, we will compare the two following settings:

1. Reference situation: fix $n = 3000$, simulate n independent copies $\mathcal{D}_n := (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n}$,

construct the dataset of pairs $\tilde{\mathcal{D}}_n$ using Algorithm 7. Use the estimators on the training set $\tilde{\mathcal{D}}_n$.

2. Independent situation: fix $n = 3000$, simulate $n(n-1) \simeq 9,000,000$ independent copies $\mathcal{D}_{n(n-1)} := (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n(n-1)}$. Create the dataset of consecutive pairs $\overline{\mathcal{D}}_{n(n-1)}$ on this sample using Algorithm 14. This means that we use only consecutive pairs, i.e. (1,2), (3,4), (5,6), and so on. Use the estimators on the training set $\overline{\mathcal{D}}_{n(n-1)}$.

Algorithm 14: Algorithm for creating the dataset of consecutive pairs from the initial dataset.

Input: Initial dataset $\mathcal{D} = (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n} \in \left(\mathbb{R}^{2+p}\right)^n$;
for $k \leftarrow 1$ **to** $\lfloor n \rfloor / 2$ **do**
  $i, j \leftarrow 2k-1, 2k$;
  $\tilde{\mathbf{Z}}_k \leftarrow (\mathbf{Z}_i + \mathbf{Z}_j)/2$;
  $W_k \leftarrow W_{(i,j)}$ as defined in Equation (7.1) ;
  $V_k \leftarrow K_h(\mathbf{Z}_i - \mathbf{Z}_j)$;
end
Define $\mathcal{K} := \{k : V_k > 0\}$;
Output: A dataset of pairs $\overline{\mathcal{D}} := (W_k, \tilde{\mathbf{Z}}_k, V_k)_{k \in \mathcal{K}} \in \left(\{-1,1\} \times \mathbb{R}^p \times \mathbb{R}_+\right)^{\lfloor n \rfloor / 2}$.

Note that, by construction, the cardinalities of $\overline{\mathcal{D}}_{n(n-1)}$ and $\tilde{\mathcal{D}}_n$ are the same, i.e. both have exactly $n(n-1)/2$ pairs. This is the reason why we chose to simulate $n(n-1)$ points in the independent situation, so that these two numbers of pairs can match. Note that the elements in $\overline{\mathcal{D}}$ are independent from each other by construction while some elements in $\tilde{\mathcal{D}}$ may not be independent from each other in general. We can now compare the performances of the estimators trained on $\overline{\mathcal{D}}_{n(n-1)}$ and on $\tilde{\mathcal{D}}_n$ using the criterion (7.10). Some results are given in Table 7.9. Note that the simulation of the each $(X_{i,1}, X_{i,2}, \mathbf{Z}_i)$ is still made under the previous one-dimensional "reference setting".

	Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
Independent	0.127	0.114	3.02	2.52	0.12	0.0363
Not independent	0.456	0.434	4.57	3.15	2.26	0.561

Table 7.9: Error criterion (7.10) multiplied by 1000 for each method and each situation. "Independent" means the independent situation with $\overline{\mathcal{D}}_{n(n-1)}$, and "Not independent" means the reference situation with $\tilde{\mathcal{D}}_n$.

As expected, all estimators show a better performance in the independent situation. Nonetheless, the independent situation has been simulated using $n(n-1) \simeq 9,000,000$ points whereas the reference situation uses only $n = 3,000$ points. Even if the numbers of pairs in both experiments are the same, the sample size of the dataset was much larger in the independent situation. This means that there is more information available, and explains also why the independent situation has a better performance: it just uses more data. Such a huge sample may not be available in practice though.

Nevertheless, the original procedure costs $O(n^2)$, which can be large for very large values of n. In this case, it is always to possible to restrict oneself to consecutive pairs, with a cost of only $O(n)$. Such a procedure is possible if the dataset is very large and Algorithm 14 can be seen as an alternative

to Algorithm 7 where only consecutive pairs are used. This would lower the computation cost at the expense of precision.

7.5 Applications to financial data

In this section, we study the changes of the conditional dependence between the daily returns of MSCI stock indices during two periods: the European debt crisis (from 18 March 2009 to 26 August 2012) and the after-crisis period (26 August 2012 to 2 March 2018). We will consider the couples (Germany, France), (Germany, Denmark), (Germany, Greece), respectively denoted by $(X_1, X_2), (X_1, X_3), (X_1, X_4)$. We will separately consider two choices of conditioning variables $\mathbf{Z}$:

- a proxy variable for the intraday volatility $\sigma := (High - Low)/Close$, where $High$ denotes the maximum daily value of the Eurostoxx index, Low denotes its minimum and $Close$ is the index value at the end of the corresponding trading day.
- a proxy of so-called "implied volatility moves" $\Delta\sigma^I$. It will record the daily variations of the EuroStoxx 50 Volatility Index, whose quotes are available at `https://www.stoxx.com/index-details?symbol=V2TX`: $\Delta\sigma_i^I := V2TX(i) - V2TX(i-1)$ for each trading day i. The EuroStoxx 50 Volatility Index $V2TX$ measures the levels of future volatility, as anticipated by the market through option prices.

Note that, for a given couple, the levels of the estimated conditional Kendall's tau are different (in general) for different conditioning variables. Indeed, the unconditional Kendall's tau $\tau_{1,2}$, the average conditional Kendall's tau with respect to σ, which is $\mathbb{E}_\sigma\left[\tau_{1,2|\sigma}\right]$ and the average conditional Kendall's tau with respect to $\Delta\sigma^I$, which is $\mathbb{E}_{\Delta\sigma^I}\left[\tau_{1,2|\Delta\sigma^I}\right]$ have no reason to be equal.

Both conditioning variables σ and $\Delta\sigma^I$ are of dimension 1. For each method and each conditioning variable, we will use the "best" choice of ψ as determined from the simulations in Section 7.4.1, that is $\psi^{(6)}$ for the methods *logit*, *probit*, *tree* and *random forests* and $\psi^{(1)}$ for the methods *Nearest neighbors* and *neural networks*. On the following figures, the matching between colors and corresponding estimators still follows Table 7.2.

Remark 7.6. *It is well-known that sequences of asset returns are not i.i.d. In particular, their volatilities are time-dependent, as in GARCH-type or stochastic volatility models. Moreover, the tail behavior of their distributions is significantly varying, due to some periods of market stress. Several families of models (switching regime models, jumps, etc) have tried to capture such stylized facts. We conjecture that such temporal dependencies will not affect our results too much. Indeed, dependence will be mitigated by considering all possible couples of random vectors, independently of their dates. It is easy to go one step beyond, for instance by keeping only the couples of returns indexed by i and j when $|i-j| > m$, for some "reasonably chosen" threshold m ($m = 20$, e.g.). In every case, it is highly likely that our inference procedures are still consistent and asymptotically normal, for most types of dependence between successive observations (mixing processes, weak dependence, m-dependence, mixingales, etc), even if the asymptotic variances are different from ours.*

7.5.1 Conditional dependence with respect to the Eurostoxx's volatility proxy σ

We will first consider the conditioning events given by σ, the proxy variable for the market intraday volatility. The results are displayed on Figures 7.2 to 7.7. Intuitively, dependence should tend to increase with

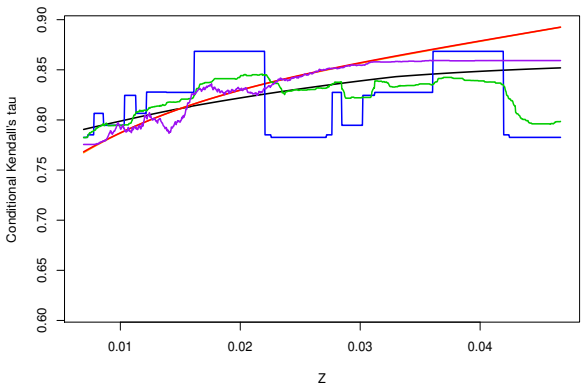


Figure 7.2: Conditional Kendall's tau between (X_1, X_2) given σ during the European debt crisis

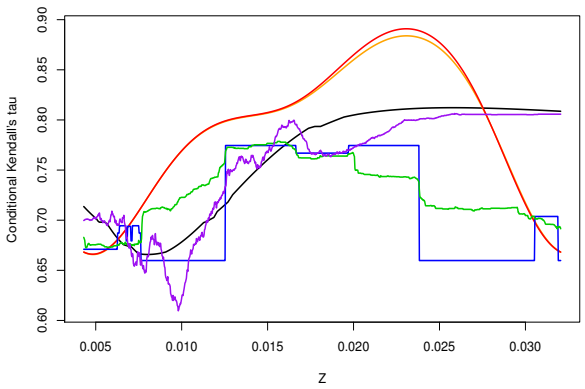


Figure 7.3: Conditional Kendall's tau between (X_1, X_2) given σ during the After-crisis period

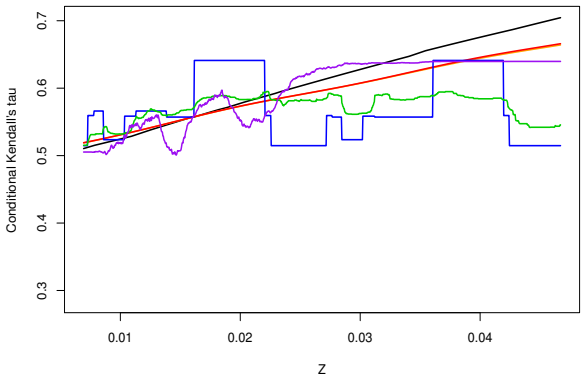


Figure 7.4: Conditional Kendall's tau between (X_1, X_3) given σ during the European debt crisis

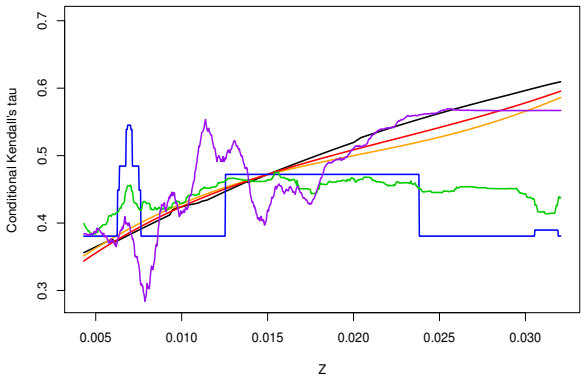


Figure 7.5: Conditional Kendall's tau between (X_1, X_3) given σ during the After-crisis period

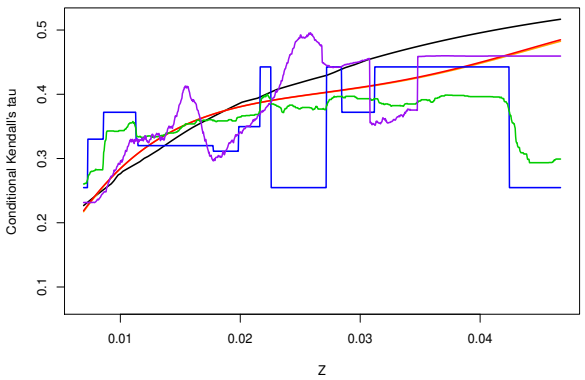


Figure 7.6: Conditional Kendall's tau between (X_1, X_4) given σ during the European debt crisis

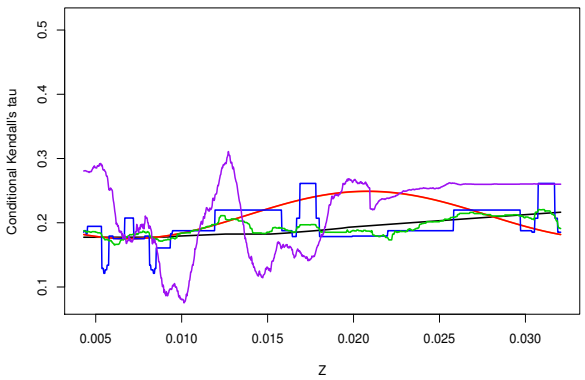


Figure 7.7: Conditional Kendall's tau between (X_1, X_4) given σ during the After-crisis period

market volatility: when "bad news" are announced, they are source of stress for most dealers, especially inside the Eurozone that brings together economically connected countries. This phenomenon should be particularly sensitive during the European debt crisis, because a lot of such "bad news" were related to the Eurozone itself (economic/financial news of public debts in several European countries). Let us see whether this is the case.

On most figures, the estimated conditional Kendall's tau seems to exhibit some kind of concavity. The behavior of these functions can be roughly broken down into two main regimes:

1. The "moderate" volatility regime (also called the "normal regime") in the sense that the volatility stay mild, say in the lower half of its range. In this normal regime, conditional Kendall's tau is an increasing function of volatility. This is coherent with most empirical research where it is shown that dependence increases with volatility.

2. The high volatility regime: this is a "stressed regime" where σ lies in the upper half of its range. In this less frequent regime, the influence of the European volatility σ on the conditional Kendall's tau appears to be less clear: the estimators become more "fluctuating" and more different from each other, as a consequence of the small number of observations in most stressed regimes.

During the European debt crisis (see Figures 7.2, 7.4 and 7.6), the three couples seem to exhibit the same shape of conditional dependence with respect to σ, even if their average levels are different. These similarities can be a little bit surprising considering that the economic situations of the corresponding countries are different. It can be conjectured that the heterogeneity in the "mean" levels of conditional dependence is sufficient to reflect this diversity of situations. In this perspective, the increasing pattern of conditional dependence w.r.t. the "volatility" would be a pure characteristic of that period, regardless of the chosen pair of European countries. Indeed, we have observed this pattern for most couples of European countries in the Eurozone. An explanation might be that investors were focusing on the same international news, for example, about the future of the Eurozone, and, therefore, they were reacting in a similar way, irrespective of the country.

For each couple of countries, conditional Kendall's tau is nearly always lower during the After-crisis period than during the European debt crisis. Apparently, in the After-crisis period, factors and events that are specific to each country attract more attention from investors than during the crisis, which results in lower dependence. In this context, the shapes of conditional dependence are no longer similar for different couples. In particular, the conditional Kendall's tau between German and French returns show a significant increase during the low volatility regime and a decrease during the high volatility regime: see Figure 7.3. The conditional dependence between the German and the Danish returns is also increasing during the low volatility regime, but in the high volatility, their conditional Kendall's tau seems to be rather constant, even increasing according to the nearest neighbors and the neural networks estimators. Concerning Figure 7.7, we do not seem any clear tendency. It is likely that σ has almost no impact on the conditional dependence between the German and Greek stock index returns.

7.5.2 Conditional dependence with respect to the variations $\Delta\sigma^I$ of the Eurostoxx's implied volatility index

The implied volatility is computed using option prices. In this sense, this financial quantity reflects investors' anticipation of future uncertainty. When important events happen, investors most often update their anticipations, which results in a change of implied volatilities. This change, denoted by $\Delta\sigma^I$ may be

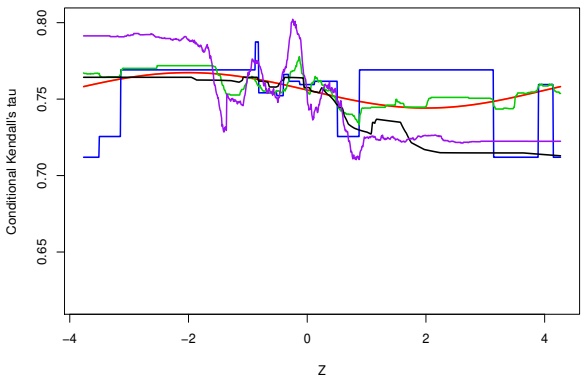


Figure 7.8: Conditional Kendall's tau between (X_1, X_2) given $\Delta\sigma^I$ during the European debt crisis

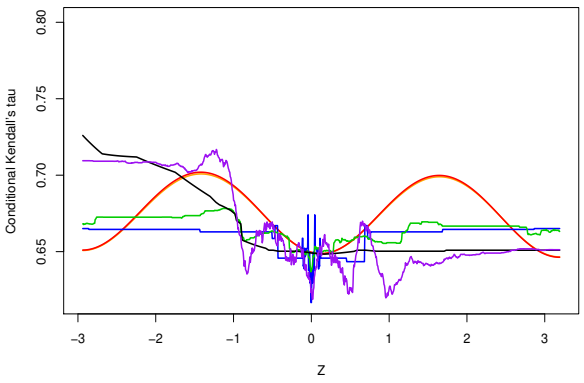


Figure 7.9: Conditional Kendall's tau between (X_1, X_2) given $\Delta\sigma^I$ during the After-crisis period

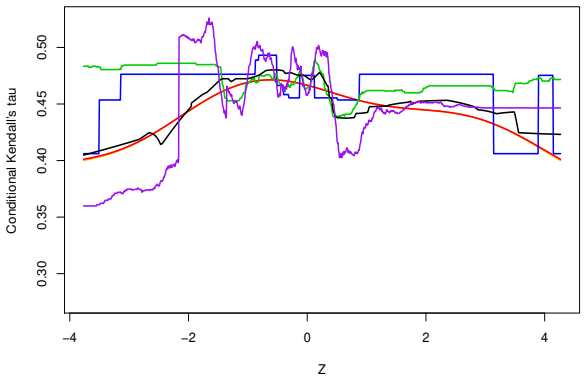


Figure 7.10: Conditional Kendall's tau between (X_1, X_3) given $\Delta\sigma^I$ during the European debt crisis

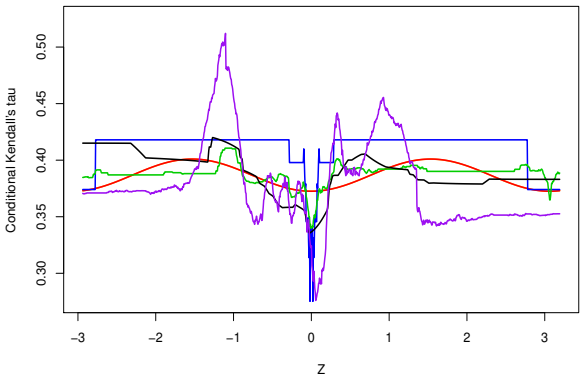


Figure 7.11: Conditional Kendall's tau between (X_1, X_3) given $\Delta\sigma^I$ during the After-crisis period

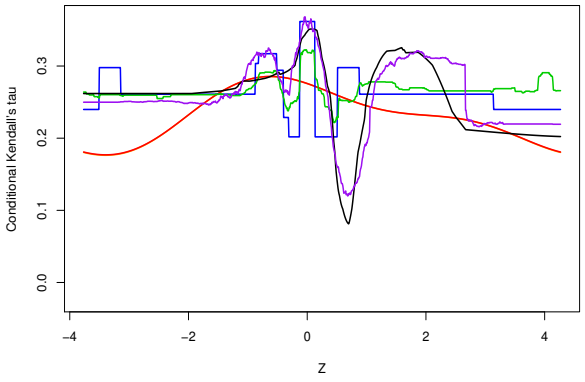


Figure 7.12: Conditional Kendall's tau between (X_1, X_4) given $\Delta\sigma^I$ during the European debt crisis

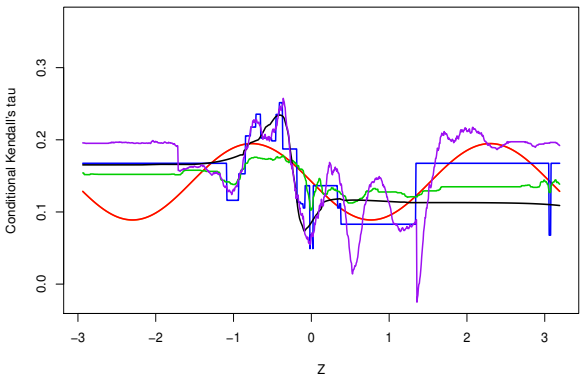


Figure 7.13: Conditional Kendall's tau between (X_1, X_4) given $\Delta\sigma^I$ during the After-crisis period

linked to variations of the conditional dependence between stock returns of different countries. Figures 7.8 to 7.13 illustrate the variations of the conditional Kendall's tau between couples of stock returns with respect to the conditioning variable $\Delta\sigma^I$ during the two periods we study.

For each couple, the levels of the conditional Kendall's tau are higher during the European debt crisis than during the after-crisis period. This is coherent with our conclusions in the previous subsection. But here, conditional Kendall's taus look like concave functions of $\Delta\sigma^I$ during the crisis, while they exhibit "double bumps" features after the crisis. During the crisis, when $\Delta\sigma^I$ is small in absolute value, implied volatilities do not change much and the dependence is in general higher than during big changes of the market implied volatility, i.e. when $|\Delta\sigma^I|$ is high (see Figures 7.10 and 7.12).

One exception is the couple (France, Germany), for which the conditional Kendall's tau is roughly a decreasing function of $\Delta\sigma^I$ during the crisis. France and Germany are close countries and have strong economic relationships, but Germany is seen as a country in the "center of Europe" while France share a lot of similarities with countries of the periphery (in the South of Europe). Indeed, during the crisis, when implied volatility decreases (corresponding to a negative value of $\Delta\sigma^I$), the dependence is higher, which can be interpreted as investors seeing the two countries as close. On the contrary, when the market implied volatility increases, there are concerns in the market about the robustness of Eurozone and investors raise doubts about southern European countries - including France - which tend to decrease the conditional Kendall's tau between French and German returns.

After the crisis, the couples (Germany, France), and (Germany, Denmark) revert to a more usual shape of conditional dependence: when volatility does not change much, conditional Kendall's tau is low ; when volatility changes much, conditional Kendall's tau is higher, reflecting more stressed situations. In this period, an exception is the couple (Germany, Greece), whose conditional Kendall's tau has a particular shape, that looks like the one of the couple (Germany, France) during the crisis. This is coherent with the fact that, in stressed situations, when volatility increases, investors sometimes remembers that Greece still has a fragile economy, which results in a lower conditional Kendall's tau. But three estimators suggest that, when volatility increases very much, conditional Kendall's tau between Germany and Greece increases again, following the classical tendencies that we had already observed.

7.6 Conclusion

In a parametric setting, we have proposed a localized log-likelihood method to estimate conditional Kendall's tau. When the link function is analytically tractable and explicit, it is then possible to code and optimize the full penalized criterion. The consistency and the asymptotic normality of such estimators have been stated. In particular, this is the case for logit or probit-type link functions. We noticed that evaluating a Kendall's tau is equivalent to evaluating a probability of being classified as a concordant pair. Therefore, most classification procedures can be adapted to directly estimate conditional Kendall's tau. Classification trees, random forests, nearest neighbors and neural networks have been discussed. They generally provide more flexible parametric models than previously.

We note that multiple trade-offs arise when choosing one of these methods, as displayed in Table 7.10. Depending on the requirements of the situation, statisticians can choose some algorithms that best match their needs. To summarize, trees and random forests methods are the fastest ones, but exhibit the lowest performances. Parametric methods such as the logit and probit may perform very well under some "simple" functional forms of g and ψ, but they deteriorate quickly when the true underlying

Method	Performance in the sense of (7.10)	Computation time	Interpretation	Tuning parameters	
				Number	Difficulty of choice
Logit / Probit (well-specified)	Best	Very Slow	Yes	1	Easy
Logit / Probit (mis-specified)	Low	Very Slow	Possible	1	Easy
Tree	Average	Very Fast	Possible	3 (see [117])	Average
Random forests	Good	Average	No	at least 4	Average
Nearest neighbors	Very Good	Fast	No	at least 5	Complicated
Neural network	Excellent	Slow	No	at least 2	Complicated

Table 7.10: Strengths and weaknesses of the proposed estimation procedures

model departs from their parametric specification. Note that they also show the longest computation time. Nonetheless, interpretability of the coefficient β can be useful in applications. Even if the model is misspecified, it can still be seen as an estimation of the best approximation of $\mathbf{z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$ on the functional space generated by ψ. Nearest neighbors methods are average in terms of computation time as well as performance. Neural networks are the slowest of all our nonparametric methods, but they behave nearly uniformly the best ones in term of prediction. Finally, we have evaluated these different methods on several empirical illustrations.

7.7 Some basic definitions about copulas

Here, we recall the main concepts around copulas and conditional copulas. First, a d-dimensional copula is a cdf on $[0,1]^d$ whose margins are uniform distributions. Sklar's theorem states that, for any d-dimensional distributions H, whose marginal cdfs' are denoted as $F_1, \ldots, F_d$, there exists a copula C s.t.

$$H(x_1, \ldots, x_d) = C\big(F_1(x_1), \ldots, F_d(x_d)\big), \tag{7.11}$$

for every $(x_1, \ldots, x_d) \in \mathbb{R}^d$. If the law of H is continuous, the latter C is unique, and it is called *the copula* associated to H. Inversely, for a given copula and some univariate cdfs' F_k, $k = 1, \ldots, d$, Equation (7.11) defines a d-dimensional cdf H.

The latter concept of copula is similarly related to any random vector $\mathbf{X}$ whose cdf is H, and there is no ambiguity by using the same term. Copulas are invariant w.r.t. strictly increasing transforms of the margins X_k, $k = 1, \ldots, d$. They provide very practical tools for modeling complex and/or highly dimensional distributions in a flexible way, by splitting the task into two parts: the specification of the marginal distributions on one side, and the specification of the copula on the other side. Therefore, a copula can be seen as a function that describes the dependence between the components of $\mathbf{X}$, independently of the marginal distributions. Several popular dependence measures are functionals of the underlying copula only: Kendall's tau, Spearman's rho, Blomqvist coefficient, etc. The classical textbooks by Joe [76] or Nelsen [106] provide numerous and detailed results.

Numerous parametric families of copulas have been proposed in the literature: Gaussian, Student, Archimedean, Marshall-Olkin, extreme-value, etc. Several inference methods have been adapted to evaluate an underlying copula, possible without estimating the marginal cdfs' (Canonical Maximum Like-

lihood). See Cherubini and Luciano [29] for details. Nonparametric methods have been developed too, since the seminal papers of Deheuvels [33, 34] about empirical copula processes.

Second, conditional copulas have been formally introduced by Patton [112, 111]. They are rather straightforward extensions of the latter concepts, when dealing with conditional distributions. Formally, for a given sigma-algebra $\mathcal{F}$, let $H(\cdot|\mathcal{F})$ (resp. $F_k(\cdot|\mathcal{F})$) be the conditional distribution of $\mathbf{X}$ (resp. X_k, $k=1,\ldots,d$) given $\mathcal{F}$. The "conditional version of" Sklar's theorem now states that there exists a random copula $C(\cdot|\mathcal{F})$ s.t.

$$H(x_1,\ldots,x_d|\mathcal{F}) = C\big(F_1(x_1|\mathcal{F}),\ldots,F_d(x_d|\mathcal{F})|\mathcal{F}\big), \text{ a.e.} \tag{7.12}$$

for every $(x_1,\ldots,x_d)\in\mathbb{R}^d$. If the law of $H(\cdot|\mathcal{F})$ is continuous, the latter $C(\cdot|\mathcal{F})$ is unique, and it is called *the conditional copula* associated to $H(\cdot|\mathcal{F})$, given $\mathcal{F}$. Inversely, given $\mathcal{F}$, a conditional copula $C(\cdot|\mathcal{F})$ and some univariate cdfs' $F_k(\cdot|\mathcal{F})$, $k=1,\ldots,d$, Equation (7.12) defines a d-dimensional conditional cdf $H(\cdot|\mathcal{F})$. See Fermanian and Wegkamp [52] for extensions of the latter concepts.

7.8 Proof of Theorem 7.3

Simple calculations provide: if $i\neq j$ and under (7.3),

$$\begin{aligned}
\mathbb{E}[L_n(\beta)] &= \mathbb{E}\left[K_h(\mathbf{Z}_i-\mathbf{Z}_j)\ell_\beta(W_{(i,j)},\mathbf{Z}_i)\right] = \mathbb{E}\left[K_h(\mathbf{Z}_i-\mathbf{Z}_j)\mathbb{E}[\ell_\beta(W_{(i,j)},\mathbf{Z}_i)|\mathbf{Z}_i,\mathbf{Z}_j]\right]\\
&= \mathbb{E}\left[K_h(\mathbf{Z}_i-\mathbf{Z}_j)\left(p(\mathbf{Z}_i,\mathbf{Z}_j)\log\left(\frac{1}{2}+\frac{1}{2}g(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)\right)+(1-p(\mathbf{Z}_i,\mathbf{Z}_j))\log\left(\frac{1}{2}-\frac{1}{2}g(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)\right)\right)\right]\\
&= \mathbb{E}\left[K_h(\mathbf{Z}_i-\mathbf{Z}_j)\phi(\mathbf{Z}_i,\mathbf{Z}_j,\beta)\right]\\
&= \mathbb{E}\left[\int K(\mathbf{t})\phi(\mathbf{Z}_i,\mathbf{Z}_i-h\mathbf{t},\beta)f_{\mathbf{Z}}(\mathbf{Z}_i-h\mathbf{t})\,d\mathbf{t}\right],
\end{aligned}$$

that tends to $\mathbb{E}\left[\phi(\mathbf{Z}_i,\mathbf{Z}_i,\beta)f_{\mathbf{Z}}(\mathbf{Z}_i)\right]=L_\infty(\beta)$ when $n\to\infty$, if $\int\left(\phi(\mathbf{z},\cdot,\beta)f_{\mathbf{Z}}(\cdot)\right)_\varepsilon(\mathbf{z})f_{\mathbf{z}}(\mathbf{z})\,d\mathbf{z}<\infty$, for some $\varepsilon>0$ (invoke the dominated convergence Theorem and the compact support of K).

Now, let us prove that, for any β, $L_n(\beta)$ tends towards $L_\infty(\beta)$ in probability, when $n\to\infty$. It is sufficient to prove that the variance of $L_n(\beta)$ tends to zero.

$$\begin{aligned}
\mathbb{E}\Big[\Big(L_n(\beta)-\mathbb{E}[L_n(\beta)]\Big)^2\Big] &= \frac{1}{n^2(n-1)^2}\sum_{i_1,j_1;i_1\neq j_1}\sum_{i_2,j_2;i_2\neq j_2}\\
&\quad\Big(\mathbb{E}\left[K_h(\mathbf{Z}_{i_1}-\mathbf{Z}_{j_1})K_h(\mathbf{Z}_{i_2}-\mathbf{Z}_{j_2})\ell_\beta(W_{(i_1,j_1)},\mathbf{Z}_{i_1})\ell_\beta(W_{(i_2,j_2)},\mathbf{Z}_{i_2})\right]-\mathbb{E}[L_n(\beta)]^2\Big)\\
&=: \frac{1}{n^2(n-1)^2}\sum_{i_1,j_1;i_1\neq j_1}\sum_{i_2,j_2;i_2\neq j_2} v_{i_1,j_1,i_2,j_2},
\end{aligned}$$

with obvious notation. Obviously, v_{i_1,j_1,i_2,j_2} is zero when i_1 and j_1 are not equal to i_2 nor j_2. At the opposite, in the case of equalities between some of these four indices, we get non-zero terms.

To be specific, when $i_1=i_2=i$, and $j_1\neq j_2$, we have

$$\begin{aligned}
v_{i,j_1,i,j_2} &= \mathbb{E}\left[K_h(\mathbf{Z}_i-\mathbf{Z}_{j_1})K_h(\mathbf{Z}_i-\mathbf{Z}_{j_2})\ell_\beta(W_{(i,j_1)},\mathbf{Z}_i)\ell_\beta(W_{(i,j_2)},\mathbf{Z}_i)\right]-\mathbb{E}[L_n(\beta)]^2\\
&= \mathbb{E}\left[\int K(\mathbf{x})K(\mathbf{y})A(\mathbf{Z}_i,\mathbf{Z}_i-h\mathbf{x},\mathbf{Z}_i-h\mathbf{y})f_{\mathbf{Z}}(\mathbf{Z}_i-h\mathbf{x})f_{\mathbf{Z}}(\mathbf{Z}_i-h\mathbf{y})\,d\mathbf{x}\,d\mathbf{y}\right]-\mathbb{E}[L_n(\beta)]^2,
\end{aligned}$$

by setting

$$\begin{aligned}
A(\mathbf{x},\mathbf{y},\mathbf{z}) &:= \mathbb{E}\left[\ell_\beta(W_{(i,j_1)},\mathbf{Z}_i)\ell_\beta(W_{(i,j_2)},\mathbf{Z}_i)|\mathbf{Z}_i=\mathbf{x},\mathbf{Z}_{j_1}=\mathbf{y},\mathbf{Z}_{j_2}=\mathbf{z}\right]\\
&= p(\mathbf{x},\mathbf{y},\mathbf{z})\log^2 q(\mathbf{x},\beta)+(p(\mathbf{x},\mathbf{y})-p(\mathbf{x},\mathbf{y},\mathbf{z}))\log q(\mathbf{x},\beta)\log(1-q(\mathbf{x},\beta))\\
&+ (p(\mathbf{x},\mathbf{z})-p(\mathbf{x},\mathbf{y},\mathbf{z}))\log q(\mathbf{x},\beta)\log(1-q(\mathbf{x},\beta))+(1-p(\mathbf{x},\mathbf{y})-p(\mathbf{x},\mathbf{z})+p(\mathbf{x},\mathbf{y},\mathbf{z}))\log^2(1-q(\mathbf{x},\beta)).
\end{aligned}$$

If $\int A(\mathbf{z},\cdot,\cdot)_\varepsilon(\mathbf{z},\mathbf{z})f^2_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z}<\infty$ for some $\varepsilon>0$, then v_{i,j_1,i,j_2} tends to a constant when $n\to\infty$ (independently of the choice of such indices).

A similar analysis can be led for the other terms. When $i_1=j_2$ and $j_1\neq i_2$, we get

$$\begin{aligned}v_{i_1,j_1,i_2,i_1}&=\mathbb{E}\left[K_h(\mathbf{Z}_{i_1}-\mathbf{Z}_{j_1})K_h(\mathbf{Z}_{i_2}-\mathbf{Z}_{i_1})\ell_\beta(W_{(i_1,j_1)},\mathbf{Z}_{i_1})\ell_\beta(W_{(i_2,i_1)},\mathbf{Z}_{i_2})\right]-\mathbb{E}[L_n(\beta)]^2\\&=\mathbb{E}\left[\int K(\mathbf{x})K(\mathbf{y})B(\mathbf{Z}_{i_1},\mathbf{Z}_{i_1}-h\mathbf{x},\mathbf{Z}_{i_1}+h\mathbf{y})f_{\mathbf{Z}}(\mathbf{Z}_{i_1}-h\mathbf{x})f_{\mathbf{Z}}(\mathbf{Z}_{i_1}+h\mathbf{y})\,d\mathbf{x}\,d\mathbf{y}\right]-\mathbb{E}[L_n(\beta)]^2,\end{aligned}$$

by setting

$$\begin{aligned}B(\mathbf{x},\mathbf{y},\mathbf{z})&:=\mathbb{E}\left[\ell_\beta(W_{(i_1,j_1)},\mathbf{Z}_{i_1})\ell_\beta(W_{(i_2,i_1)},\mathbf{Z}_{i_2})|\mathbf{Z}_{i_1}=\mathbf{x},\mathbf{Z}_{j_1}=\mathbf{y},\mathbf{Z}_{i_2}=\mathbf{z}\right]\\&=p(\mathbf{x},\mathbf{y},\mathbf{z})\log q(\mathbf{x},\beta)\log q(\mathbf{z},\beta)+(p(\mathbf{x},\mathbf{y})-p(\mathbf{x},\mathbf{y},\mathbf{z}))\log q(\mathbf{x},\beta)\log(1-q(\mathbf{z},\beta))\\&+(p(\mathbf{x},\mathbf{z})-p(\mathbf{x},\mathbf{y},\mathbf{z}))\log q(\mathbf{z},\beta)\log(1-q(\mathbf{x},\beta))\\&+(1-p(\mathbf{x},\mathbf{y})-p(\mathbf{x},\mathbf{z})+p(\mathbf{x},\mathbf{y},\mathbf{z}))\log(1-q(\mathbf{x},\beta))\log(1-q(\mathbf{z},\beta)).\end{aligned}$$

If $\int B(\mathbf{z},\cdot,\cdot)_\varepsilon(\mathbf{z},\mathbf{z})f^2_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z}<\infty$, then v_{i_1,j_1,i,i_1} tends to a constant when $n\to\infty$.

When $j_1=j_2=j$ and $i_1\neq i_2$, we obtain

$$\begin{aligned}v_{i_1,j_1,i_2,j_2}&=\mathbb{E}\left[K_h(\mathbf{Z}_{i_1}-\mathbf{Z}_j)K_h(\mathbf{Z}_{i_2}-\mathbf{Z}_j)\ell_\beta(W_{(i_1,j)},\mathbf{Z}_{i_1})\ell_\beta(W_{(i_2,j)},\mathbf{Z}_{i_2})\right]-\mathbb{E}[L_n(\beta)]^2\\&=\mathbb{E}\left[\int K(\mathbf{x})K(\mathbf{y})C(\mathbf{Z}_j+h\mathbf{x},\mathbf{Z}_j,\mathbf{Z}_j+h\mathbf{y})f_{\mathbf{Z}}(\mathbf{Z}_j+h\mathbf{x})f_{\mathbf{Z}}(\mathbf{Z}_j+h\mathbf{y})\,d\mathbf{x}\,d\mathbf{y}\right]-\mathbb{E}[L_n(\beta)]^2,\end{aligned}$$

by setting

$$\begin{aligned}C(\mathbf{x},\mathbf{y},\mathbf{z})&:=\mathbb{E}\left[\ell_\beta(W_{(i_1,j)},\mathbf{Z}_{i_1})\ell_\beta(W_{(i_2,j)},\mathbf{Z}_{i_2})|\mathbf{Z}_{i_1}=\mathbf{x},\mathbf{Z}_j=\mathbf{y},\mathbf{Z}_{i_2}=\mathbf{z}\right]\\&=p(\mathbf{x},\mathbf{y},\mathbf{z})\log q(\mathbf{x},\beta)\log q(\mathbf{z},\beta)+(p(\mathbf{x},\mathbf{y})-p(\mathbf{x},\mathbf{y},\mathbf{z}))\log q(\mathbf{x},\beta)\log(1-q(\mathbf{z},\beta))\\&+(p(\mathbf{y},\mathbf{z})-p(\mathbf{x},\mathbf{y},\mathbf{z}))\log q(\mathbf{z},\beta)\log(1-q(\mathbf{x},\beta))\\&+(1-p(\mathbf{x},\mathbf{y})-p(\mathbf{y},\mathbf{z})+p(\mathbf{x},\mathbf{y},\mathbf{z}))\log(1-q(\mathbf{x},\beta))\log(1-q(\mathbf{z},\beta)).\end{aligned}$$

If $\int C(\cdot,\mathbf{z},\cdot)_\varepsilon(\mathbf{z},\mathbf{z})f^2_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z}<\infty$, then $v_{i_1,j,i_2,j}$ tends to a constant when $n\to\infty$.

When $j_1=i_2$ and $i_1\neq j_2$:

$$\begin{aligned}v_{i_1,j_1,j_1,j_2}&=\mathbb{E}\left[K_h(\mathbf{Z}_{i_1}-\mathbf{Z}_{j_1})K_h(\mathbf{Z}_{j_1}-\mathbf{Z}_{j_2})\ell_\beta(W_{(i_1,j_1)},\mathbf{Z}_{i_1})\ell_\beta(W_{(j_1,j_2)},\mathbf{Z}_{j_1})\right]-\mathbb{E}[L_n(\beta)]^2\\&=\mathbb{E}\left[\int K(\mathbf{x})K(\mathbf{y})D(\mathbf{Z}_{j_1}+h\mathbf{x},\mathbf{Z}_{j_1},\mathbf{Z}_{j_1}-h\mathbf{y})f_{\mathbf{Z}}(\mathbf{Z}_{j_1}+h\mathbf{x})f_{\mathbf{Z}}(\mathbf{Z}_{j_1}-h\mathbf{y})\,d\mathbf{x}\,d\mathbf{y}\right]-\mathbb{E}[L_n(\beta)]^2,\end{aligned}$$

by setting

$$\begin{aligned}D(\mathbf{x},\mathbf{y},\mathbf{z})&:=\mathbb{E}\left[\ell_\beta(W_{(i_1,j_1)},\mathbf{Z}_{i_1})\ell_\beta(W_{(j_1,j_2)},\mathbf{Z}_{j_1})|\mathbf{Z}_{i_1}=\mathbf{x},\mathbf{Z}_{j_1}=\mathbf{y},\mathbf{Z}_{j_2}=\mathbf{z}\right]\\&=p(\mathbf{x},\mathbf{y},\mathbf{z})\log q(\mathbf{x},\beta)\log q(\mathbf{y},\beta)+(p(\mathbf{x},\mathbf{y})-p(\mathbf{x},\mathbf{y},\mathbf{z}))\log q(\mathbf{x},\beta)\log(1-q(\mathbf{y},\beta))\\&+(p(\mathbf{y},\mathbf{z})-p(\mathbf{x},\mathbf{y},\mathbf{z}))\log q(\mathbf{y},\beta)\log(1-q(\mathbf{x},\beta))\\&+(1-p(\mathbf{x},\mathbf{y})-p(\mathbf{y},\mathbf{z})+p(\mathbf{x},\mathbf{y},\mathbf{z}))\log(1-q(\mathbf{x},\beta))\log(1-q(\mathbf{y},\beta)).\end{aligned}$$

If $\int D(\cdot,\mathbf{z},\cdot)_\varepsilon(\mathbf{z},\mathbf{z})f^2_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z}<\infty$, then v_{i_1,j_1,j_1,j_2} tends to a constant when $n\to\infty$.

There are two cases of two equalities. If $i_1=i_2=i$ and $j_1=j_2=j$, this yields

$$\begin{aligned}v_{i,j,i,j}&=\mathbb{E}\left[K_h(\mathbf{Z}_i-\mathbf{Z}_j)^2\ell^2_\beta(W_{(i,j)},\mathbf{Z}_i)\right]-\mathbb{E}[L_n(\beta)]^2\\&=h^{-p}\mathbb{E}\left[\int K(\mathbf{x})^2E(\mathbf{Z}_i,\mathbf{Z}_i-h\mathbf{x})f_{\mathbf{Z}}(\mathbf{Z}_i-h\mathbf{x})\,d\mathbf{x}\right]-\mathbb{E}[L_n(\beta)]^2,\end{aligned}$$

by setting

$$\begin{aligned} E(\mathbf{x},\mathbf{y}) &:= \mathbb{E}\left[\ell_\beta^2(W_{(i,j)},\mathbf{Z}_i)|\mathbf{Z}_i=\mathbf{x},\mathbf{Z}_j=\mathbf{y}\right] \\ &= p(\mathbf{x},\mathbf{y})\log^2 q(\mathbf{x},\beta)+(1-p(\mathbf{x},\mathbf{y}))\log^2(1-q(\mathbf{x},\beta)). \end{aligned}$$

If $\int E(\mathbf{z},\cdot)_\varepsilon(\mathbf{z})f_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z}<\infty$, then $h^p v_{i,j,i,j}$ tends to a constant when $n\to\infty$.

Finally, if $i_1=j_2$ and $j_1=i_2$, we get

$$\begin{aligned} v_{i_1,j_1,i_1,j_1} &= \mathbb{E}\left[K_h(\mathbf{Z}_{i_1}-\mathbf{Z}_{j_1})^2\ell_\beta(W_{(i_1,j_1)},\mathbf{Z}_{i_1})\ell_\beta(W_{(j_1,i_1)},\mathbf{Z}_{j_1})\right]-\mathbb{E}[L_n(\beta)]^2 \\ &= h^{-p}\mathbb{E}\left[\int K(\mathbf{x})^2F(\mathbf{Z}_i,\mathbf{Z}_i-h\mathbf{x})f_{\mathbf{Z}}(\mathbf{Z}_i-h\mathbf{x})\,d\mathbf{x}\right]-\mathbb{E}[L_n(\beta)]^2, \end{aligned}$$

by setting

$$\begin{aligned} F(\mathbf{x},\mathbf{y}) &:= \mathbb{E}\left[\ell_\beta(W_{(i_1,j_1)},\mathbf{Z}_{i_1})\ell_\beta(W_{(j_1,i_1)},\mathbf{Z}_{j_1})|\mathbf{Z}_{i_1}=\mathbf{x},\mathbf{Z}_{j_1}=\mathbf{y}\right] \\ &= p(\mathbf{x},\mathbf{y})\log q(\mathbf{x},\beta)\log q(\mathbf{y},\beta)+(1-p(\mathbf{x},\mathbf{y}))\log(1-q(\mathbf{x},\beta))\log(1-q(\mathbf{y},\beta)). \end{aligned}$$

If $\int F(\mathbf{z},\cdot)_\varepsilon(\mathbf{z})f_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z}<\infty$, then $h^p v_{i_1,j_1,j_1,i_1}$ tends to a constant when $n\to\infty$.

Summarizing the previous terms, we have obtained $Var(L_n(\beta))=O(n^{-1}+n^{-2}h^{-p})$, that tends to zero pointwise, when $n^2h^p\to\infty$. We deduce $L_n(\beta)-L_\infty(\beta)=L_n(\beta)-\mathbb{E}[L_n(\beta)]+\mathbb{E}[L_n(\beta)]-L_\infty(\beta)=o_P(1)$. Since $L_n(\cdot)$ and $L_\infty(\cdot)$ are concave, invoking the convexity lemma of Geyer [59] (see Knight and Fu [82], alternatively), the maximizer $\hat\beta$ of L_n tends in probability towards the maximizer of L_∞. □

We summarize the latter technical assumptions that are sufficient to obtain the consistency of $\hat\beta$: for some $\varepsilon>0$,

$$\int \left(\phi(\mathbf{z},\cdot,\beta)f_{\mathbf{Z}}(\cdot)\right)_\varepsilon(\mathbf{z})f_{\mathbf{z}}(\mathbf{z})\,d\mathbf{z}<\infty, \tag{7.13}$$

$$\int \Big(A(\mathbf{z},\cdot,\cdot)_\varepsilon(\mathbf{z},\mathbf{z})+B(\mathbf{z},\cdot,\cdot)_\varepsilon(\mathbf{z},\mathbf{z})+C(\cdot,\mathbf{z},\cdot)_\varepsilon(\mathbf{z},\mathbf{z})+D(\cdot,\mathbf{z},\cdot)_\varepsilon(\mathbf{z},\mathbf{z})\Big)f^2_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z}<\infty, \tag{7.14}$$

$$\int \Big(\phi(\mathbf{z},\cdot,\beta)_\varepsilon(\mathbf{z})+E(\mathbf{z},\cdot)_\varepsilon(\mathbf{z})+F(\mathbf{z},\cdot)_\varepsilon(\mathbf{z})\Big)f_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z}<\infty. \tag{7.15}$$

7.9 Proof of Theorem 7.4

Set $\mathbf{u}:=\sqrt{n}(\beta-\beta^*)$ and $\hat{\mathbf{u}}:=\sqrt{n}(\hat\beta-\beta^*)$. Obviously,

$$\begin{aligned} \hat{\mathbf{u}} &= \arg\max_{\mathbf{u}\in\mathbb{R}^{p'}} L_n(\beta^*+n^{-1/2}\mathbf{u})-\lambda_n|\beta^*+n^{-1/2}\mathbf{u}|_1, \\ &= \arg\max_{\mathbf{u}\in\mathbb{R}^{p'}} nL_n(\beta^*+n^{-1/2}\mathbf{u})-nL_n(\beta^*)-n\lambda_n\big\{|\beta^*+n^{-1/2}\mathbf{u}|_1-|\beta^*|_1\big\}. \end{aligned}$$

Note that

$$\begin{aligned} n\lambda_n|\beta^*+n^{-1/2}\mathbf{u}|_1-n|\beta^*|_1 &= n^{1/2}\lambda_n\sum_{k;\beta_k^*=0}|u_k|+n^{1/2}\lambda_n\sum_{k;\beta_k^*\neq 0}sign(\beta_k^*)u_k \\ &\longrightarrow \mu\sum_{k;\beta_k^*=0}|u_k|+\mu\sum_{k;\beta_k^*\neq 0}sign(\beta_k^*)u_k, \end{aligned}$$

when $n\to\infty$. Moreover,

$$nL_n(\beta^*+n^{-1/2}\mathbf{u})-nL_n(\beta^*)=n^{1/2}\dot L_n(\beta^*).\mathbf{u}+\frac{1}{2}\mathbf{u}^T\ddot L_n(\overline\beta)\mathbf{u}+\frac{1}{6\sqrt{n}}\dddot L_n(\overline\beta).\mathbf{u}^{(3)},$$

for some (random) $\overline\beta$ s.t. $|\beta^*-\overline\beta|<|\beta^*-\beta|$. We will successively prove that

(i) $n^{1/2}\dot{L}_n(\beta^*)$ weakly tends to a Gaussian random vector $\mathbb{W}$, $\mathbb{W} \sim \mathcal{N}(0_p, \Sigma_{\beta^*})$;

(ii) $\ddot{L}_n(\beta^*)$ tends in probability towards a constant matrix $\mathbb{H}(\beta^*)$;

(iii) $\dddot{L}_n(\overline{\beta})$ is $O_P(1)$.

Then, $\hat{\mathbf{u}} = \arg\max_{\mathbf{u}} \mathcal{L}_n(\mathbf{u})$, where $\mathcal{L}_n(\mathbf{u})$ weakly tends to

$$\mathcal{L}_\infty(\mathbf{u}) := \mathbb{W}.\mathbf{u} + \frac{1}{2}\mathbf{u}^T\mathbb{H}(\beta^*)\mathbf{u} - \mu\sum_{k;\beta_k^*=0}|u_k| - \mu\sum_{k;\beta_k^*\neq 0} sign(\beta_k^*)u_k,$$

that is concave. The result will follow, applying Theorem 1 in Kato [79].

First, let us prove (i), i.e. the asymptotic normality of $n^{1/2}\dot{L}_n(\beta)$ for any given parameter β. Consider the centered criterion

$$M_n(\beta) := \dot{L}_n(\beta) - \mathbb{E}[\dot{L}_n(\beta)] = \frac{1}{n(n-1)}\sum_{i,j;i\neq j}\ell_{ij}(\beta),$$

where $\ell_{ij} := K_h(\mathbf{Z}_i - \mathbf{Z}_j)\partial_\beta\ell_\beta(W_{(i,j)}, \mathbf{Z}_i) - \mathbb{E}[\dot{L}_n(\beta)]$. We symmetrize the localized likelihood:

$$M_n(\beta) = \frac{1}{2n(n-1)}\sum_{i,j;i\neq j} M_{ij}(\beta),$$

where $M_{ij}(\beta)$ (or simply M_{ij}) is $\ell_{ij}(\beta) + \ell_{ji}(\beta)$. Note that $M_{ij} = M_{ji}$ and that $\mathbb{E}[M_{ij}] = 0$.

By the dominated convergence theorem and a change of variable, we easily check that $\mathbb{E}[\dot{L}_n(\beta)] = \partial_\beta L_\infty(\beta) + o(1)$ if, for some $\varepsilon > 0$, we have $\int \left(\partial_\beta\phi(\mathbf{z},\cdot,\beta)f_{\mathbf{Z}}(\cdot)\right)_\varepsilon(\mathbf{z})f_{\mathbf{z}}(\mathbf{z})\,d\mathbf{z} < \infty$. Moreover, by simple calculations, we get, if $i \neq j$,

$$\begin{aligned}
\mathbb{E}[M_n|\mathbf{Z}_i] &= \frac{1}{2n}\mathbb{E}[M_{ij} + M_{ji}|\mathbf{Z}_i] = \frac{1}{n}\mathbb{E}[M_{ij}|\mathbf{Z}_i] \\
&= \frac{1}{n}\int\{K_h(\mathbf{Z}_i - \mathbf{z})\partial_\beta\phi(\mathbf{Z}_i, \mathbf{z}, \beta) + K_h(\mathbf{z} - \mathbf{Z}_i)\partial_\beta\phi(\mathbf{z}, \mathbf{Z}_i, \beta)\}f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} - \frac{2}{n}\mathbb{E}[\dot{L}_n(\beta)] \\
&= \frac{1}{n}\int K(\mathbf{t})\{\partial_\beta\phi(\mathbf{Z}_i, \mathbf{Z}_i - h\mathbf{t}, \beta)f_{\mathbf{Z}}(\mathbf{Z}_i - h\mathbf{t}) + \partial_\beta\phi(\mathbf{Z}_i + h\mathbf{t}, \mathbf{Z}_i, \beta)f_{\mathbf{Z}}(\mathbf{Z}_i + h\mathbf{t})\}\,d\mathbf{t} - \frac{2}{n}\mathbb{E}[\dot{L}_n(\beta)] \\
&= \frac{2}{n}\partial_\beta\phi(\mathbf{Z}_i, \mathbf{Z}_i, \beta)f_{\mathbf{Z}}(\mathbf{Z}_i) - \frac{2}{n}\dot{L}_\infty(\beta) + o(n^{-1}) + r_{n,i},
\end{aligned}$$

where, by a m-order limited expansion, we obtain

$$\|r_{n,i}\| \leq \frac{Cst.h^m\int|K|}{nm!}\|(f_{\mathbf{Z}}(\cdot)\partial_\beta\phi(\mathbf{Z}_i,\cdot,\beta))^{(m)} + (f_{\mathbf{Z}}(\cdot)\partial_\beta\phi(\cdot,\mathbf{Z}_i,\beta))^{(m)}\|_\varepsilon(\mathbf{Z}_i),$$

for any norm $\|\cdot\|$ on $\mathbb{R}^p$ and a positive constant Cst. We deduce that $n^{1/2}\sum_{i=1}^n \mathbb{E}[M_n|\mathbf{Z}_i]$ is asymptotically normal by invoking the usual CLT, under condition (7.17) below. To be specific, the Hájek projection of M_n is

$$\frac{\sqrt{n}}{2}\sum_{i=1}^n \mathbb{E}[M_n|\mathbf{Z}_i] = \frac{1}{\sqrt{n}}\sum_{i=1}^n\{\partial_\beta\phi(\mathbf{Z}_i, \mathbf{Z}_i, \beta)f_{\mathbf{Z}}(\mathbf{Z}_i) - \dot{L}_\infty(\beta)\} + o_P(1) \rightsquigarrow \mathcal{N}(0, \Sigma_\beta), \text{ with}$$

$$\Sigma_\beta := \int \partial_\beta\phi(\mathbf{z}, \mathbf{z}, \beta)\partial_\beta\phi(\mathbf{z}, \mathbf{z}, \beta)^T f_{\mathbf{Z}}^3(\mathbf{z})\,d\mathbf{z} - \dot{L}_\infty(\beta)\dot{L}_\infty(\beta)^T.$$

Note that that $\dot{L}_\infty(\beta^*) = 0$.

Now consider the "remainder term" $\Delta_n := M_n(\beta) - \sum_{i=1}^n \mathbb{E}[M_n|\mathbf{Z}_i]/2$. Since $\mathbb{E}[M_n|\mathbf{Z}_i] = n^{-1}\mathbb{E}[M_{ij}|\mathbf{Z}_i]$, we deduce

$$\Delta_n = M_n(\beta) - \frac{1}{2}\sum_{i=1}^n \mathbb{E}[M_n|\mathbf{Z}_i] = \frac{1}{2n(n-1)}\sum_{i,j;i\neq j}\{M_{ij} - \mathbb{E}[M_{ij}|\mathbf{Z}_i]\}.$$

It is relatively easy to prove that Δ_n is negligible, i.e. $\Delta_n = o_P(n^{-1/2})$. Indeed, let us prove that the variance of $n^{1/2}\Delta_n$ tends to zero with n:

$$Var(n^{1/2}\Delta_n) = n\mathbb{E}[\Delta_n\Delta_n^T] = \frac{1}{4n(n-1)^2}\sum_{i_1,j_1;i_1\neq j_1}\sum_{i_2,j_2;i_2\neq j_2}\delta(i_1,i_2,j_1,j_2),$$

$$\delta(i_1,i_2,j_1,j_2) = \mathbb{E}\left[\{M_{i_1j_1} - \mathbb{E}[M_{i_1j_1}|\mathbf{Z}_{i_1}]\}\times\{M_{i_2j_2} - \mathbb{E}[M_{i_2j_2}|\mathbf{Z}_{i_2}]\}^T\right].$$

If there is no identity among the indices (i_1,j_1,i_2,j_2), with $i_1\neq j_1$ and $i_2\neq j_2$, then $\delta(i_1,i_2,j_1,j_2)$ is zero. Moreover, this is still the case when there is only a single identity. For instance, assume $i_1 = i_2 = i$ and $j_1\neq j_2$. Then,

$$\begin{aligned}\delta(i,i,j_1,j_2) &= \mathbb{E}\left[\{M_{ij_1} - \mathbb{E}[M_{ij_1}|\mathbf{Z}_i]\}\times\{M_{ij_2} - \mathbb{E}[M_{ij_2}|\mathbf{Z}_i]\}^T\right]\\ &= \mathbb{E}\left[\{M_{ij_1} - \mathbb{E}[M_{ij_1}|\mathbf{Z}_i]\}\times\mathbb{E}[\{M_{ij_2} - \mathbb{E}[M_{ij_2}|\mathbf{Z}_i]\}^T|\mathbf{Z}_i,\mathbf{Z}_{j_1}]\right]\\ &= \mathbb{E}\left[\{M_{ij_1} - \mathbb{E}[M_{ij_1}|\mathbf{Z}_i]\}\times 0\right] = 0.\end{aligned}$$

The other terms for which a single identity between the indices can be managed similarly.

At the opposite, non-zero terms appear when $i_1 = i_2 = i$ and $j_1 = j_2 = j$. In this case, we obtain

$$\delta(i,i,j,j) = \mathbb{E}\left[\{M_{ij} - \mathbb{E}[M_{ij}|\mathbf{Z}_i]\}\{M_{ij} - \mathbb{E}[M_{ij}|\mathbf{Z}_i]\}^T\right] = \mathbb{E}[M_{ij}M_{ij}^T] - \mathbb{E}[\mathbb{E}[M_{ij}|\mathbf{Z}_i]\mathbb{E}[M_{ij}|\mathbf{Z}_i]^T].$$

By a usual change of variable and by symmetry, we get

$$\begin{aligned}\mathbb{E}[M_{ij}M_{ij}^T] &= 2\mathbb{E}\left[\{\ell_{ij}(\beta)\ell_{ij}(\beta)^T + \ell_{ij}(\beta)\ell_{ji}(\beta)^T\}\right]\\ &= 2\mathbb{E}\left[K_h^2(\mathbf{Z}_i-\mathbf{Z}_j)\partial_\beta\ell_\beta(W_{(i,j)},\mathbf{Z}_i)\partial_\beta\ell_\beta(W_{(i,j)},\mathbf{Z}_i)^T\right.\\ &+ \left. K_h(\mathbf{Z}_i-\mathbf{Z}_j)K_h(\mathbf{Z}_j-\mathbf{Z}_i)\partial_\beta\ell_\beta(W_{(i,j)},\mathbf{Z}_i)\partial_\beta\ell_\beta(W_{(j,i)},\mathbf{Z}_j)^T\right] + O(1)\\ &= 2h^{-p}\mathbb{E}\left[\int\left(K^2(\mathbf{x})H_1(\mathbf{Z}_i,\mathbf{Z}_i-h\mathbf{x}) + K(\mathbf{x})K(-\mathbf{x})H_2(\mathbf{Z}_i,\mathbf{Z}_i-h\mathbf{x})\right)f_{\mathbf{Z}}(\mathbf{Z}_i-h\mathbf{x})\,d\mathbf{x}\right] + O(1)\\ &= 2h^{-p}\int\left(K^2(\mathbf{x})H_1(\mathbf{z},\mathbf{z}-h\mathbf{x}) + K(\mathbf{x})K(-\mathbf{x})H_2(\mathbf{z},\mathbf{z}-h\mathbf{x})\right)f_{\mathbf{Z}}(\mathbf{z}-h\mathbf{x})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{x}\,d\mathbf{z} + O(1),\end{aligned}$$

by setting

$$\begin{aligned}H_1(\mathbf{x},\mathbf{y}) &= \mathbb{E}\left[\partial_\beta\ell_\beta(W_{(i,j)},\mathbf{Z}_i)\partial_\beta\ell_\beta(W_{(i,j)},\mathbf{Z}_i)^T|\mathbf{Z}_i=\mathbf{x},\mathbf{Z}_j=\mathbf{y}\right]\\ &= p(\mathbf{x},\mathbf{y})\frac{\boldsymbol{\psi}(\mathbf{x})\boldsymbol{\psi}(\mathbf{x})^T g'(\boldsymbol{\psi}(\mathbf{x})^T\beta)^2}{(1+g(\boldsymbol{\psi}(\mathbf{x})^T\beta))^2} + (1-p(\mathbf{x},\mathbf{y}))\frac{\boldsymbol{\psi}(\mathbf{x})\boldsymbol{\psi}(\mathbf{x})^T g'(\boldsymbol{\psi}(\mathbf{x})^T\beta)^2}{(1-g(\boldsymbol{\psi}(\mathbf{x})^T\beta))^2},\ \text{and}\end{aligned}$$

$$\begin{aligned}H_2(\mathbf{x},\mathbf{y}) &= \mathbb{E}\left[\partial_\beta\ell_\beta(W_{(i,j)},\mathbf{Z}_i)\partial_\beta\ell_\beta(W_{(j,i)},\mathbf{Z}_j)^T|\mathbf{Z}_i=\mathbf{x},\mathbf{Z}_j=\mathbf{y}\right]\\ &= p(\mathbf{x},\mathbf{y})\frac{\boldsymbol{\psi}(\mathbf{x})\boldsymbol{\psi}(\mathbf{y})^T g'(\boldsymbol{\psi}(\mathbf{x})^T\beta)g'(\boldsymbol{\psi}(\mathbf{y})^T\beta)}{(1+g(\boldsymbol{\psi}(\mathbf{x})^T\beta))(1+g(\boldsymbol{\psi}(\mathbf{y})^T\beta))} + (1-p(\mathbf{x},\mathbf{y}))\frac{\boldsymbol{\psi}(\mathbf{x})\boldsymbol{\psi}(\mathbf{y})^T g'(\boldsymbol{\psi}(\mathbf{x})^T\beta)g'(\boldsymbol{\psi}(\mathbf{y})^T\beta)}{(1-g(\boldsymbol{\psi}(\mathbf{x})^T\beta))(1+g(\boldsymbol{\psi}(\mathbf{y})^T\beta))}.\end{aligned}$$

Therefore, $\mathbb{E}[M_{ij}M_{ij}^T]$ is $O(h^{-p})$, if $\int(\|H_1\|+\|H_2\|)_\varepsilon(\mathbf{z},\cdot)f_{\mathbf{Z}}(\cdot)_\varepsilon(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty$.

The last possible case providing non-zero $\delta(i_1,i_2,j_1,j_2)$ is $i_1 = j_2 = i$ and $j_1 = i_2 = j$. Then, we obtain

$$\delta(i,j,j,i) = \mathbb{E}\left[\{M_{ij} - \mathbb{E}[M_{ij}|\mathbf{Z}_i]\}\times\{M_{ji} - \mathbb{E}[M_{ji}|\mathbf{Z}_j]\}^T\right] = \delta(i,i,j,j),$$

due to the symmetry of M_{ij}. Thus, we have proved that $Var(n^{1/2}\Delta_n) = O(n^{-1}h^{-p})) = o(1)$, which implies

$$n^{1/2}M_n(\beta) = \frac{n^{1/2}}{2}\sum_{i=1}^n\mathbb{E}[M_n|\mathbf{Z}_i] + o_P(1).$$

We deduce that $n^{1/2}M_n(\beta)$ weakly tends towards the Gaussian random vector $\mathcal{N}(0_p, \Sigma_\beta)$ for any β. When $\beta = \beta^*$, $\partial_\beta L_\infty(\beta^*) = 0$, and this yields (i).

Second, let us deal with (ii) above. It is easy to prove that $\ddot{L}_n(\beta^*)$ tends to $\mathbb{H}(\beta^*)$ in probability, when n tends to the infinity. Indeed, the arguments are exactly the same as in 7.8, where we have proved that $L_n(\beta^*)$ is convergent in probability. We only have to replace $\ell_\beta(\cdot,\cdot)$ by its second derivatives w.r.t. β. To save space, the specific derivations of such conditions of regularity are left to the reader: simply replace the functions A, B,...,F of 7.8 by their second derivatives w.r.t. β, taken at $\beta = \beta^*$, and rewrite (7.14) and (7.15).

Third, to prove (iii), it is sufficient to state that $\mathbb{E}[\|\dddot{L}_n(\beta)\|]$ is bounded from above, uniformly w.r.t. β in a small neighborhood of β^*. By derivation, we get, for every indices a, b, c in $\{1, \ldots, p'\}$,

$$\begin{aligned}
&\mathbb{E}[|\frac{\partial^3}{\partial\beta_a\partial\beta_b\partial\beta_c}L_n(\beta)|] \\
&\quad\leq \quad Cst \times \mathbb{E}\Big[|K|_h(\mathbf{Z}_i - \mathbf{Z}_j)\big\{p(\mathbf{Z}_i,\mathbf{Z}_j)H(1,\mathbf{Z}_i,\beta,a,b,c,) + (1-p(\mathbf{Z}_i,\mathbf{Z}_j))H(-1,\mathbf{Z}_i,\beta,a,b,c)\big\}\Big],
\end{aligned}$$

where, for every $\delta \in \{1,-1\}$,

$$H(\delta,\mathbf{Z}_i,\beta,a,b,c) = \left(\frac{|g'(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)|^3}{|1+\delta g(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)|^3} + \frac{|g'g''(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)|}{|1+\delta g(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)|^2} + \frac{|g'''(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)|}{|1+\delta g(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)|}\right)|\boldsymbol{\psi}(\mathbf{Z}_i)_a\boldsymbol{\psi}(\mathbf{Z}_i)_b\boldsymbol{\psi}(\mathbf{Z}_i)_c|.$$

and Cst denotes a real constant that depend on g and (a,b,c) only. Therefore, it is sufficient to assume that

$$\int |K|(\mathbf{t})\Big(p(\,\mathbf{z}-h\mathbf{t})H(1,\mathbf{z},\beta,a,b,c) + (1-p(\mathbf{z},\mathbf{z}-h\mathbf{t}))H(-1,\mathbf{z},\beta,a,b,c)\Big)f_{\mathbf{Z}}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z}-h\mathbf{t})\,d\mathbf{t}\,d\mathbf{z} < \infty.$$

This is guaranteed by Assumption (7.19) below. Then, under the latter assumption, (iii) is stated and this finishes the proof. □

For convenience, let us gather the main technical assumptions that have been requested to prove Theorem 7.4: for some $\varepsilon > 0$,

$$\int \big(\partial_\beta\phi(\mathbf{z},\cdot,\beta)f_{\mathbf{Z}}(\cdot)\big)_{\varepsilon}(\mathbf{z})f_{\mathbf{z}}(\mathbf{z})\,d\mathbf{z} < \infty. \tag{7.16}$$

$$\mathbb{E}\left[\|(f_{\mathbf{Z}}(\cdot)\partial_\beta\phi(\mathbf{Z}_i,\cdot,\beta))^{(m)} + (f_{\mathbf{Z}}(\cdot)\partial_\beta\phi(\cdot,\mathbf{Z}_i,\beta))^{(m)}\|_\varepsilon(\mathbf{Z}_i)\right] < \infty. \tag{7.17}$$

$$\int (\|H_1\| + \|H_2\|)_\varepsilon(\mathbf{z},\cdot)f_{\mathbf{Z}}(\cdot)_\varepsilon(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty. \tag{7.18}$$

For every indices $(a,b,c) \in \{1,\ldots,p'\}$ and for $\mathcal{V}(\beta^*)$, some (small) neighborhood around β^*,

$$\sup_{\beta\in\mathcal{V}(\beta^*)}\int\left((p(\mathbf{z},\cdot)f_{\mathbf{Z}}(\cdot))_\varepsilon(\mathbf{z})H(1,\mathbf{z},\beta,a,b,c) + ((1-p)(\mathbf{z},\cdot)f_{\mathbf{Z}}(\cdot))_\varepsilon(\mathbf{z})H(-1,\mathbf{z},\beta,a,b,c,)\right)f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty. \tag{7.19}$$

Remark 7.7. *Note that* $||p(\cdot,\cdot)||_\infty \leq 1$. *If* g *and its derivatives are bounded, Condition (7.19) is satisfied if*

$$\sup_{\beta\in\mathcal{V}(\beta^*)}\sup_{\delta\in\{-1,1\}}\int \|\boldsymbol{\psi}(\mathbf{z})\|^3|1+\delta g(\boldsymbol{\psi}(\mathbf{z})^T\beta)|^{-3}f_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty.$$

Acknowledgments

This work is supported by the Labex Ecodec under the grant ANR-11-LABEX-0047 from the French Agence Nationale de la Recherche. The first author thanks Solt Kovács for inspiring ideas and a discussion which lead to this article.

Part III

Other topics in inference

Chapter 8

Estimation of a regular conditional functional by conditional U-statistic regression

Abstract

U-statistics are a large class of estimators, generalizing the empirical mean of a random variable $\mathbf{X}$ to sums over every k-tuple of distinct observations of $\mathbf{X}$. They can be used to estimate a regular functional $\theta(\mathbb{P}_{\mathbf{X}})$ of the law of $\mathbf{X}$. When a covariate $\mathbf{Z}$ is available, a conditional U-statistic describes the effect of $\mathbf{z}$ on the conditional law of $\mathbf{X}$ given $\mathbf{Z} = \mathbf{z}$. Conditional U-statistics can therefore be used to estimate a regular conditional functional $\theta(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\cdot})$. We give non-asymptotic bounds for conditional U-statistics and review asymptotic results. Then, assuming a parametric model of the conditional functional of interest, we propose a regression-type estimator based on conditional U-statistics. Its theoretical properties are derived, first in a non-asymptotic framework and then in two different asymptotic regimes. Some examples are given to illustrate our methods.

Keywords: kernel smoothing, regression-type models, penalized estimation.

Based on [36]: Derumigny, A., Estimation of a regular conditional functional by conditional U-statistics regression. *Arxiv preprint*, arXiv:1903.10914, 2019.

8.1 Introduction

Let $\mathbf{X}$ be a random variable with values in a measurable space $(\mathcal{X}, \mathcal{A})$, and denote by $\mathbb{P}_{\mathbf{X}}$ its law. A natural application is $\mathcal{X} = \mathbb{R}^{p_{\mathbf{X}}}$, for a fixed dimension $p_{\mathbf{X}} > 0$. Often, we are interested in estimating a regular functional $\theta(\mathbb{P}_{\mathbf{X}})$ of the law of $\mathbf{X}$, of the form

$$\theta(\mathbb{P}_{\mathbf{X}}) = \mathbb{E}\big[g(\mathbf{X}_1, \dots, \mathbf{X}_k)\big] = \int g(\mathbf{x}_1, \dots, \mathbf{x}_k) d\mathbb{P}_{\mathbf{X}}(\mathbf{x}_1) \cdots d\mathbb{P}_{\mathbf{X}}(\mathbf{x}_k),$$

for a fixed $k > 0$, a function $g : \mathcal{X}^k \to \mathbb{R}$ and $\mathbf{X}_1, \dots \mathbf{X}_k \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\mathbf{X}}$. Following Hoeffding [69], a natural estimator of $\theta(\mathbb{P}_{\mathbf{X}})$ is the U-statistic $\hat{\theta}(\mathbb{P}_{\mathbf{X}})$, defined by

$$\hat{\theta}(\mathbb{P}_{\mathbf{X}}) := \sum_{\sigma \in \mathfrak{I}_{k,n}} g\big(\mathbf{X}_{\sigma(1)}, \dots, \mathbf{X}_{\sigma(k)}\big),$$

where $\mathfrak{I}_{k,n}$ is the set of injective functions from $\{1, \dots, k\}$ to $\{1, \dots, n\}$. For an introduction to the theory of U-statistics, we refer to Koroljuk and Borovskich [84] and Serfling [123, Chapter 5].

In our framework, we assume that we do not only observe $\mathbf{X}$, but we observe in fact $(\mathbf{X}, \mathbf{Z})$ where $\mathbf{Z}$ is a p-dimensional covariate. We are now interested in functionals of the conditional law $\mathbb{P}_{\mathbf{X}|\mathbf{Z}}$. For each $\mathbf{z}_1, \dots, \mathbf{z}_k \in \mathcal{Z}$, where $\mathcal{Z}$ is a compact subset of $\mathbb{R}^p$, we can define such a functional $\theta_{\mathbf{z}_1,\dots\mathbf{z}_k}$ by

$$\begin{aligned}
\theta_{\mathbf{z}_1,\dots\mathbf{z}_k}(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\cdot}) &:= \theta(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_1}, \dots, \mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_k}) \\
&= \mathbb{E}_{\bigotimes_{i=1}^k \mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_i}}\big[g(\mathbf{X}_1, \dots, \mathbf{X}_k)\big] = \mathbb{E}\big[g(\mathbf{X}_1, \dots, \mathbf{X}_k)\big|\mathbf{Z}_i = \mathbf{z}_i, \forall i = 1, \dots, k\big] \\
&= \int g(\mathbf{x}_1, \dots, \mathbf{x}_k) d\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_1}(\mathbf{x}_1) \cdots d\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_k}(\mathbf{x}_k).
\end{aligned}$$

This can be seen as a generalization of $\theta(\mathbb{P}_{\mathbf{X}})$ to the conditional case. Indeed, when $\mathbf{X}$ and $\mathbf{Z}$ are independent, the new functional $\theta_{\mathbf{z}_1,\dots\mathbf{z}_k}(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\cdot})$ is equal to the unconditional functional $\theta(\mathbb{P}_{\mathbf{X}})$. For convenience, we will use the notation $\theta(\mathbf{z}_1, \dots, \mathbf{z}_k) := \theta_{\mathbf{z}_1,\dots\mathbf{z}_k}(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\cdot})$, treating the law of $(\mathbf{X}, \mathbf{Z})$ as fixed (but unknown). Stute [132] defined a kernel-based estimator $\hat{\theta}(\mathbf{z}_1, \dots, \mathbf{z}_k)$ of the conditional functional $\theta(\mathbf{z}_1, \dots, \mathbf{z}_k)$ by

$$\hat{\theta}(\mathbf{z}_1, \dots, \mathbf{z}_k) := \frac{\sum_{\sigma \in \mathfrak{I}_{k,n}} K_h\big(\mathbf{Z}_{\sigma(1)} - \mathbf{z}_1\big) \cdots K_h\big(\mathbf{Z}_{\sigma(k)} - \mathbf{z}_k\big)\, g\big(\mathbf{X}_{\sigma(1)}, \dots, \mathbf{X}_{\sigma(k)}\big)}{\sum_{\sigma \in \mathfrak{I}_{k,n}} K_h\big(\mathbf{Z}_{\sigma(1)} - \mathbf{z}_1\big) \cdots K_h\big(\mathbf{Z}_{\sigma(k)} - \mathbf{z}_k\big)}, \tag{8.1}$$

where $h > 0$ is the bandwidth, $K(\cdot)$ a kernel on $\mathbb{R}^p$, $K_h(\cdot) := h^{-p}K(\cdot/h)$, and $(\mathbf{X}_i, \mathbf{Z}_i) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\mathbf{X},\mathbf{Z}}$. Stute [132] proved the asymptotic normality of $\hat{\theta}(\mathbf{z}_1, \dots, \mathbf{z}_k)$ and its weak and strong consistency. Dony and Mason [45] derived its uniform in bandwidth consistency under VC-type conditions over a class of possible functions g.

Nevertheless, the estimator (8.1) has several weaknesses. First, interpretation of the whole hypersurface $(\mathbf{z}_1, \dots, \mathbf{z}_k) \mapsto \hat{\theta}(\mathbf{z}_1, \dots, \mathbf{z}_k)$ can be difficult. Indeed, the latter is a curve of dimension $1 + p \times k$, which is rather challenging to visualize even for small values of p and k. Second, for each new tuple $(\mathbf{z}_1, \dots, \mathbf{z}_k)$, computation of $\hat{\theta}(\mathbf{z}_1, \dots, \mathbf{z}_k)$ has a cost of $O(n^k)$. This means that if we want to estimate $\hat{\theta}(\mathbf{z}_1^{(i)}, \dots, \mathbf{z}_k^{(i)})$ for every $i = 1, \dots, N$, where $\big(\mathbf{z}_1^{(1)}, \dots, \mathbf{z}_k^{(1)}, \dots, \mathbf{z}_1^{(N)}, \dots, \mathbf{z}_k^{(N)}\big) \in \mathcal{Z}^{k \times N}$, then the total cost is $O(Nn^k)$. Third, it is well-known that kernel estimators are not very smooth, in the sense that they usually present many spurious local minima and maxima, that can be a problem in applications. Therefore, we may want to have estimators which are more monotone with respect to the conditioning variables $\mathbf{z}_1, \dots \mathbf{z}_k$, and have a simple functional form.

Another idea is to decompose the function $(\mathbf{z}_1, \dots, \mathbf{z}_k) \mapsto \theta(\mathbf{z}_1, \dots, \mathbf{z}_k)$ in a basis $(\psi_i)_{i \geq 0}$, generalizing the work of Derumigny and Fermanian [39]. This may not be always easy if the range of the function $\theta(\cdot, \cdots, \cdot)$ is a strict subset of $\mathbb{R}$. In that case, it is always possible to use a "link function" Λ, strictly increasing and continuously differentiable such that the range $\Lambda \circ \theta(\cdot, \cdots, \cdot)$ is exactly $\mathbb{R}$. Whatever the choice of Λ (including the identity function), we can decompose the latter function in any basis $(\psi_i)_{i \geq 0}$. If only a finite number $r > 0$ of elements of this basis are necessary to represent the whole function $\Lambda \circ \theta(\cdot, \cdots, \cdot)$ over $\mathcal{Z}^k$, then we have the following parametric model

$$\forall (\mathbf{z}_1, \dots, \mathbf{z}_k) \in \mathcal{Z}^k,\ \Lambda\big(\theta(\mathbf{z}_1, \dots, \mathbf{z}_k)\big) = \boldsymbol{\psi}(\mathbf{z}_1, \dots, \mathbf{z}_k)^T \beta^*, \tag{8.2}$$

where $\beta^* \in \mathbb{R}^r$ is the true parameter and $\boldsymbol{\psi}(\cdot) := \big(\psi_1(\cdot), \dots, \psi_r(\cdot)\big)^T \in \mathbb{R}^r$. In most applications, finding an appropriate basis $\boldsymbol{\psi}$ is not easy ; this will depend of the choice of the (conditional) functional θ. Therefore, the most simple solution consists in choosing a concatenation of several well-known basis such as polynomials, exponentials, sinuses and cosinuses, indicator functions, etc... They allow to take into account potential non-linearities and even discontinuities of the function $\Lambda \circ \theta(\cdot, \cdots, \cdot)$. The only condition is their linear independence, as seen in the following proposition (whose straightforward proof is omitted).

Proposition 8.1. *The parameter β^* is identifiable in Model (8.2) if and only if the functions $(\psi_1(\cdot), \dots, \psi_r(\cdot))$ are linearly independent $\mathbb{P}_{\mathbf{Z}}^{\otimes n}$-almost everywhere in the sense that, for all vectors $\mathbf{t} = (t_1, \dots, t_r) \in \mathbb{R}^r$, $\mathbb{P}_{\mathbf{Z}}^{\otimes n}\big(\boldsymbol{\psi}(\mathbf{Z}_1, \dots, \mathbf{Z}_n)^T\mathbf{t} = 0\big) = 1 \Longrightarrow \mathbf{t} = 0$.*

With such a choice of a wide and flexible class of functions, it is likely that not all these functions are relevant. This is what is know as sparsity, i.e. the number of non-zero coefficients of β^*, denoted by $|\mathcal{S}|$ = $|\beta^*|_0 \leq s$, for some $s \in \{1, \dots, r\}$, where $|\cdot|_0$ is the number of non-zero components of a vector of $\mathbb{R}^r$ and $\mathcal{S}$ is the set of non-zero components of β^*. Note that in this framework, r can be moderately large, for example 30 or 50 while the original dimension p is small, for example $p = 1$ or 2. This corresponds to the decomposition of a function, defined on a small-dimension domain, in a mildly large basis.

Remark 8.2. *At first sight, in Model (8.2), there seem to be no noise perturbing the variable of interest. In fact, it can be seen as a simple consequence of our formulation of the model. In the same way, the classical linear model $Y = \mathbf{X}^T\beta^* + \varepsilon$ can be rewritten as $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = \mathbf{x}^T\beta^*$ without any explicit noise. By definition, $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ is a deterministic function of a given $\mathbf{x}$. In our case, the corresponding fact is: $\Lambda\big(\theta(\mathbf{z}_1, \dots, \mathbf{z}_k)\big)$ is a deterministic function of the variables $(\mathbf{z}_1, \dots, \mathbf{z}_k)$. This means that we cannot write formally a model with noise, such as $\Lambda\big(\theta(\mathbf{z}_1, \dots, \mathbf{z}_k)\big) = \boldsymbol{\psi}(\mathbf{z}_1, \dots, \mathbf{z}_k)^T\beta^* + \varepsilon$ where ε is independent of the choice of $(\mathbf{z}_1, \dots, \mathbf{z}_k)$ since the left-hand side of the latter equality is a $(\mathbf{z}_1, \dots, \mathbf{z}_k)$-mesurable quantity, unless ε is constant almost surely.*

Contrary to more usual models, the explained variable $\Lambda\big(\theta(\mathbf{z}_1, \dots, \mathbf{z}_k)\big)$, is not observed in Model (8.2). Therefore, a direct estimation of the parameter β^* (for example, by the ordinary least squares, or by the Lasso) is unfeasible. In other words, even if the function $(\mathbf{z}_1, \dots, \mathbf{z}_k) \mapsto \Lambda\big(\theta(\mathbf{z}_1, \dots, \mathbf{z}_k)\big)$ is deterministic (by definition of conditional probabilities), finding the best β in Model (8.2) is far from being a numerical analysis problem since the function to be decomposed is unknown. Nevertheless, we will replace $\Lambda\big(\theta(\mathbf{z}_1, \dots, \mathbf{z}_k)\big)$ by the nonparametric estimate $\Lambda\big(\hat{\theta}(\mathbf{z}_1, \dots, \mathbf{z}_k)\big)$, and use it as an approximation of the explained variable.

More precisely, we fix a finite collection of points $\mathbf{z}'_1, \dots, \mathbf{z}'_{n'} \in \mathcal{Z}^{n'}$ and a collection $\mathfrak{I}_{k,n'}$ of injective functions $\sigma : \{1, \dots, k\} \to \{1, \dots, n'\}$. Note that we are not forced to include *all* the injective functions in $\mathfrak{I}_{k,n'}$, reducing its number of elements. This will allow us to decrease the computational cost of the procedure. For every $\sigma \in \mathfrak{I}_{k,n'}$, we estimate $\hat{\theta}(\mathbf{z}'_{\sigma(1)}, \dots, \mathbf{z}'_{\sigma(k)})$. Finally, the estimator $\hat{\beta}$ is defined as the minimizer of the following l_1-penalized criteria

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^r} \left[\frac{(n'-k)!}{n'!} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \left(\Lambda\Big(\hat{\theta}\big(\mathbf{z}'_{\sigma(1)}, \dots, \mathbf{z}'_{\sigma(k)}\big)\Big) - \boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)}, \dots, \mathbf{z}'_{\sigma(k)}\big)^T\beta \right)^2 + \lambda|\beta|_1 \right], \quad (8.3)$$

where λ is a positive tuning parameter (that may depend on n and n'), and $|\cdot|_q$ denotes the l_q norm, for $1 \leq q \leq \infty$. This procedure is summed up in the following Algorithm 15. Note that even if we study the general case with any $\lambda \geq 0$, corresponding properties of the unpenalized estimator can be derived by choosing the particular case $\lambda = 0$.

Algorithm 15: Two-step estimation of β

Input: A dataset $(X_{i,1}, X_{i,2}, \mathbf{Z}_i)$, $i = 1, \dots, n$

Input: A finite collection of points $\mathbf{z}'_1, \dots, \mathbf{z}'_{n'} \in \mathcal{Z}^{n'}$, selected for estimation

Input: A collection of N k-tuples for prediction $\big(\mathbf{z}_1^{(1)}, \dots, \mathbf{z}_k^{(1)}, \dots, \mathbf{z}_1^{(N)}, \dots, \mathbf{z}_k^{(N)}\big) \in \mathcal{Z}^{k \times N}$

for $\sigma \in \mathfrak{I}_{k,n'}$ **do**

  Compute the estimator $\hat{\theta}\big(\mathbf{z}'_{\sigma(1)}, \dots, \mathbf{z}'_{\sigma(k)}\big)$ using the sample $(\mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, \dots, n$;

end

Compute the minimizer $\hat{\beta}$ of (8.3) using the $\hat{\theta}\big(\mathbf{z}'_{\sigma(1)}, \dots, \mathbf{z}'_{\sigma(k)}\big)$, $j = 1, \dots, n'$, estimated in the above step ;

for $i \leftarrow 1$ **to** N **do**

  Compute the prediction $\tilde{\theta}(\mathbf{z}_1^{(i)}, \dots, \mathbf{z}_k^{(i)}) := \Lambda^{(-1)}\big(\boldsymbol{\psi}(\mathbf{z}_1^{(i)}, \dots, \mathbf{z}_k^{(i)})^T \hat{\beta}\big)$;

end

Output: An estimator $\hat{\beta}$ and N predictions $\tilde{\theta}(\mathbf{z}_1^{(i)}, \dots, \mathbf{z}_k^{(i)})$, $i = 1, \dots, N$.

Once an estimator $\hat{\beta}$ of β^* has been computed, the prediction of all the conditional functionals is reduced to the computation of $\Lambda^{(-1)}\big(\boldsymbol{\psi}(\mathbf{z}_1^{(i)}, \dots, \mathbf{z}_k^{(i)})^T \hat{\beta}\big) := \tilde{\theta}(\mathbf{z}_1^{(i)}, \dots, \mathbf{z}_k^{(i)})$, for every $i = 1, \dots, N$. The total computational cost of this new method is therefore $O(|\mathfrak{I}_{k,n'}| \cdot n'^k + |\mathfrak{I}_{k,n'}| \cdot r + Ns)$ operations. The first term corresponds to the cost of evaluating each non-parametric estimator (8.1). The second term corresponds to the minimization of the convex optimization program (8.3), and the last one is the prediction cost. Note that it can provide a huge improvement compared to the previously available estimator with a cost in $O(Nn^k)$ when $N \to \infty$, i.e. when we want to recover the full function $\theta(\cdot, \cdots, \cdot)$. Moreover, the speedup given by Algorithm 15 compared to the original conditional U-statistics (8.1) even increases with the sample size n, for moderate choices of n'.

A similar model, called *functional response* has already been studied ; see, e.g. Kowalski and Tu [85, Chapter 6.2]. They provide a method to estimate the parameter β^*, using generalized estimating equations. However, they only provides asymptotic results for their estimator, and their algorithm needs to solve a multi-dimensional equation which has no reason to be convex.

In Section 8.2, we provide both asymptotic properties and non-asymptotic bounds for the non-parametric estimator $\hat{\theta}$. Section 8.3 is devoted to the theoretical study of the two-step parametric estimator $\hat{\beta}$. Finally, we detail examples in Section 8.4. All proof have been postponed to the Appendix.

8.2 Theoretical properties of the nonparametric estimator $\hat{\theta}(\cdot)$

Remark that if g is symmetric, then $\hat{\theta}(\mathbf{z}_1, \dots, \mathbf{z}_k)$ is equal to

$$\hat{\theta}^{\uparrow}(\mathbf{z}_1, \dots, \mathbf{z}_k) := \frac{\sum_{\sigma \in \mathfrak{I}_{k,n}^{\uparrow}} K_h\big(\mathbf{Z}_{\sigma(1)} - \mathbf{z}_1\big) \cdots K_h\big(\mathbf{Z}_{\sigma(k)} - \mathbf{z}_k\big)\, g\big(\mathbf{X}_{\sigma(1)}, \dots, \mathbf{X}_{\sigma(k)}\big)}{\sum_{\sigma \in \mathfrak{I}_{k,n}^{\uparrow}} K_h\big(\mathbf{Z}_{\sigma(1)} - \mathbf{z}_1\big) \cdots K_h\big(\mathbf{Z}_{\sigma(k)} - \mathbf{z}_k\big)}, \tag{8.4}$$

where $\mathfrak{I}_{k,n}^{\uparrow}$ is the set of strictly increasing functions from $\{1, \dots, k\}$ to $\{1, \dots, n\}$. As a consequence, the cost of the computation of $\hat{\theta}^{\uparrow}$ is reduced by a factor of $k!$ compared to $\hat{\theta}$. The estimator $\hat{\theta}$ is well-defined if and only if $N_k(\mathbf{z}_1, \dots, \mathbf{z}_k) > 0$, where

$$N_k(\mathbf{z}_1, \dots, \mathbf{z}_k) := \frac{k!(n-k)!}{n!} \sum_{\sigma \in \mathfrak{I}_{k,n}^{\uparrow}} K_h\big(\mathbf{Z}_{\sigma(1)} - \mathbf{z}_1\big) \cdots K_h\big(\mathbf{Z}_{\sigma(k)} - \mathbf{z}_k\big). \tag{8.5}$$

8.2.1 Non-asymptotic bounds for N_k

To prove that our estimator $\hat{\theta}(\mathbf{z}_1, \dots, \mathbf{z}_k)$ exists with a probability that tends to 1, we will study the behavior of N_k. We will need the following assumptions to control the behavior of the kernel K and of the density of $\mathbf{Z}$.

Assumption 8.2.1. *The kernel $K(\cdot)$ is bounded, i.e. there exists a finite constant C_K such that $K(\cdot) \leq C_K$ and $\int K(\mathbf{u})d\mathbf{u} = 1$. The kernel is of order α for some $\alpha > 0$, i.e. for all $j = 1, \dots, \alpha - 1$ and all $1 \leq i_1, \dots, i_\alpha \leq p$, $\int K(\mathbf{u})\, u_{i_1} \dots u_{i_j}\, d\mathbf{u} = 0$.*

Assumption 8.2.2. *$f_{\mathbf{Z}}$ is α-times continuously differentiable on $\mathcal{Z}$ and there exists a finite constant $C_{K,\alpha}$ such that, for all $\mathbf{z}_1, \dots \mathbf{z}_k$,*

$$\int \Big|K\big(\mathbf{u}_1\big) \cdots K\big(\mathbf{u}_k\big)\Big| \sum_{m_1 + \cdots + m_k = \alpha} \binom{\alpha}{m_{1:k}} \cdot \prod_{i=1}^{k} \sum_{j_1, \dots, j_{m_i} = 1}^{p} \big|u_{i,j_1} \dots u_{i,j_{m_i}}\big| \sup_{t \in [0,1]} \left| \frac{\partial^{m_i} f_{\mathbf{Z}}}{\partial z_{j_1} \cdots \partial z_{j_{m_i}}}\big(\mathbf{z}_i + t\mathbf{u}_i\big) \right| d\mathbf{u}_1 \cdots d\mathbf{u}_k \leq C_{K,\alpha}$$

where $\binom{\alpha}{m_{1:k}} := \alpha! / \big(\prod_{i=1}^{k}(m_i!)\big)$ is the multinomial coefficient.

Assumption 8.2.3. *$f_{\mathbf{Z}}(\cdot) \leq f_{\mathbf{Z},max}$ for some finite constant $f_{\mathbf{Z},max}$.*

Lemma 8.3. *Under Assumptions 8.2.1, 8.2.2 and 8.2.3, we have for any $t > 0$,*

$$\mathbb{P}\bigg(\big|N_k(\mathbf{z}_1, \dots, \mathbf{z}_k) - \prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i)\big| \leq \frac{C_{K,\alpha}}{\alpha!} h^\alpha + t\bigg) \geq 1 - 2\exp\bigg(-\frac{[n/k]t^2}{h^{-kp}C_1 + h^{-kp}C_2 t}\bigg),$$

where $C_1 := 2f_{\mathbf{Z},max}^k ||K||_2^{2k}$, and $C_2 := (4/3)C_K^k$ and $||K||_2^2 := \int K^2$.

This Lemma is proved in Section 8.6.1. More can be said if the density $f_{\mathbf{Z}}$ is bounded below. Therefore, we will use the following assumption.

Assumption 8.2.4. *There exists a constant $f_{\mathbf{Z},min} > 0$ such that for every $\mathbf{z} \in \mathcal{Z}$, $f_{\mathbf{Z}}(\mathbf{z}) > f_{\mathbf{Z},min}$.*

If for some $\epsilon > 0$, we have $C_{K,\alpha}h^\alpha/\alpha! + t \leq f_{\mathbf{Z},min} - \epsilon$, then $\hat{f}(\mathbf{z}) \geq \epsilon > 0$ with probability larger than on the event whose probability is bound in Lemma 8.3. We should therefore choose the largest t possible, which yields the following corollary.

Corollary 8.4. *Under Assumptions 8.2.1-8.2.4, if $C_{K,\alpha}h^\alpha/\alpha! < f_{\mathbf{Z},min}$, then the random variable $N_k(\mathbf{z}_1, \dots, \mathbf{z}_k)$ is strictly positive with probability larger than*

$$1 - 2\exp\bigg(-\frac{[n/k]h^{kp}\big(f_{\mathbf{Z},min} - C_{K,\alpha}h^\alpha/\alpha!\big)^2}{C_1 + C_2\big(f_{\mathbf{Z},min} - C_{K,\alpha}h^\alpha/\alpha!\big)}\bigg),$$

guaranteeing the existence of the estimator $\hat{\theta}(\mathbf{z}_1, \dots, \mathbf{z}_k)$ on this event.

8.2.2 Non-asymptotic bounds in probability for $\hat{\theta}$

To establish bounds on $\hat{\theta}$ for every fixed n, we will need some assumptions on the joint law of $(\mathbf{X}, \mathbf{Z})$.

Assumption 8.2.5. *There exists a measure μ on $(\mathcal{X}, \mathcal{A})$ such that $\mathbb{P}_{\mathbf{X},\mathbf{Z}}$ is absolutely continuous with respect to $\mu \otimes Leb_p$, where Leb_p is the Lebesgue measure on $\mathbb{R}^p$.*

Assumption 8.2.6. *For every $\mathbf{x} \in \mathcal{X}$, $\mathbf{z} \mapsto f_{\mathbf{X},\mathbf{Z}}(\mathbf{x},\mathbf{z})$ is differentiable almost everywhere up to the order α. Moreover, there exists a finite constant $C_{g,f,\alpha} > 0$, such that, for every positive integers $m_1, \dots, m_k$ such that $\sum_{i=1}^k m_i = \alpha$, for every $0 \leq j_1, \dots, j_{m_i} \leq p$,*

$$\int \prod_{i=1}^k \left| \Big(g(\mathbf{x}_1, \dots, \mathbf{x}_k) - \mathbb{E}\big[g(\mathbf{X}_1, \dots, \mathbf{X}_k) \big| \mathbf{Z}_i = \mathbf{z}_i, \forall i = 1, \dots, k \big] \Big) \right.$$
$$\left. \cdot \left(\frac{\partial^{m_i} f_{\mathbf{X},\mathbf{Z}}}{\partial z_{j_1} \cdots \partial z_{j_{m_i}}}(\mathbf{x}_i, \mathbf{z}_i + \mathbf{u}_i) - \frac{\partial^{m_i} f_{\mathbf{X},\mathbf{Z}}}{\partial z_{j_1} \cdots \partial z_{j_{m_i}}}(\mathbf{x}_i, \mathbf{z}_i) \right) \right| d\mu(\mathbf{x}_1) \cdots d\mu(\mathbf{x}_k) \leq C_{g,f,\alpha} \prod_{i=1}^k \left| \mathbf{u}_i \right|_\infty,$$

for every choices of $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathcal{X}$ and $\mathbf{z}_1, \dots, \mathbf{z}_k \in \mathcal{Z}, \mathbf{u}_1, \dots, \mathbf{u}_k \in \mathbb{R}^p$ such that $\mathbf{z}_i + \mathbf{u}_i \in \mathcal{Z}$. There exists a constant $C'_{K,\alpha}$ such that

$$\sum_{m_1 + \cdots + m_k = \alpha} \binom{n}{m_{1:k}} \int \prod_{i=1}^k K(\mathbf{u}_i) \sum_{j_1, \dots, j_{m_i} = 1}^p u_{i,j_1} \dots u_{i,j_{m_i}} \prod_{i=1}^k \left| \mathbf{u}_i \right|_\infty d\mathbf{u}_1 \cdots d\mathbf{u}_k \leq C'_{K,\alpha}.$$

An easy situation is the case when g is bounded, i.e. when the following assumption hold.

Assumption 8.2.7. *There exists a constant C_g such that $||g||_\infty \leq C_g < +\infty$.*

When g is not bounded, a weaker result can still be proved under a "conditional Bernstein" assumption. This assumption will help us to control the tail behavior of g so that exponential concentration bounds are available.

Assumption 8.2.8 (conditional Bernstein assumption)**.** *There exists a positive function B_g such that, for all $l \geq 1$ and $\mathbf{z}_1, \dots, \mathbf{z}_k \in \mathbb{R}^{kp}$,*

$$\mathbb{E}\Big[\big| g(\mathbf{X}_1, \dots, \mathbf{X}_k) \big|^l \Big| \mathbf{Z}_1 = \mathbf{z}_1, \dots, \mathbf{Z}_k = \mathbf{z}_k \Big] \leq B_g(\mathbf{z}_1, \dots, \mathbf{z}_k)^l l!,$$

such that $B_g(\mathbf{Z}_1, \dots, \mathbf{Z}_k) \leq \tilde{B}_g$ almost surely, for some finite positive constant $\tilde{B}_g$.

As a shortcut notation, we will define also $B_{g,\mathbf{z}} := B_g(\mathbf{z}_1, \dots, \mathbf{z}_k)$. The following proposition is proved in Section 8.6.2.

Proposition 8.5 (Exponential bound for the estimator $\hat{\theta}(\mathbf{z}_1, \dots, \mathbf{z}_k)$, with fixed $\mathbf{z}_1, \dots \mathbf{z}_k \in \mathcal{Z}^k$)**.** *Assume either Assumption 8.2.7 or the weaker Assumption 8.2.8. Under Assumptions 8.2.1-8.2.6, for every $t, t' > 0$ such that $C_{K,\alpha} h^\alpha / \alpha! + t < f_{\mathbf{Z},min}/2$, we have*

$$\mathbb{P}\left(\left| \hat{\theta}(\mathbf{z}_1, \dots, \mathbf{z}_k) - \theta(\mathbf{z}_1, \dots, \mathbf{z}_k) \right| < \big(1 + C_3 h^\alpha + C_4 t\big) \times \big(C_5 h^{k+\alpha} + t'\big) \right)$$
$$\geq 1 - 2\exp\left(- \frac{[n/k] t^2 h^{kp}}{C_1 + C_2 t} \right) - 2\exp\left(- \frac{[n/k] t'^2 h^{kp}}{C_6 + C_7 t'} \right),$$

where $C_3 := 4 f_{\mathbf{Z},max}^k f_{\mathbf{Z},min}^{-2k} C_{K,\alpha}/\alpha!$, $C_4 := 4 f_{\mathbf{Z},max}^k f_{\mathbf{Z},min}^{-2k}$ and $C_5 := C_{g,f,\alpha} C'_{K,\alpha} f_{\mathbf{Z},min}^{-k}/\alpha!$.

If Assumption 8.2.7 is satisfied, the result holds with the following values: $C_6 := 2C_g^2 f_{\mathbf{Z},max}^k f_{\mathbf{Z},min}^{-2k} ||K||_2^{2k}$, $C_7 := (8/3) C_K^k C_g^k f_{\mathbf{Z},min}^{-k}$; in the case of Assumption 8.2.8, the result holds with the following alternative values: $\tilde{C}_6 := 128\big(B_{g,\mathbf{z}} + \tilde{B}_g\big)^2 C_K^{2k-1} f_{\mathbf{Z},min}^{-2k}$, $\tilde{C}_7 := 2\big(B_{g,\mathbf{z}} + \tilde{B}_g\big) C_K^k f_{\mathbf{Z},min}^{-k}$.

8.2.3 Asymptotic results for $\hat{\theta}$

The estimator $\hat{\theta}(\mathbf{z}_1, \dots, \mathbf{z}_k)$ has been first studied by Stute (1991) [132]. He proved the consistency and the asymptotic normality of $\hat{\theta}(\mathbf{z}_1, \dots, \mathbf{z}_k)$. We recall his results as they will be used in Section 8.3.

Assumption 8.2.9. *(i)* $h_n \to 0$ *and* $nh_n^p \to \infty$ *;*

(ii) $K(\mathbf{z}) \geq C_{K,1}\mathbb{1}_{\{|\mathbf{z}|_\infty \leq C_{K,1}\}}$ *for some* $C_{K,1}$, $C_{K,2} > 0$ *;*

(iii) there exists a decreasing function $H : \mathbb{R}_+ \to \mathbb{R}_+$*, and positive constants* c_1, c_2 *such that* $H(t) \underset{t\to\infty}{=} o(t^{-1})$ *and* $c_1 H(|\mathbf{z}|_\infty) \leq K(\mathbf{z}) \leq c_2 H(|\mathbf{z}|_\infty)$.

Proposition 8.6 (Consistency of $\hat{\theta}$, Theorem 2 in Stute [132])**.** *Under Assumption 8.2.9, for* $\mathbb{P}_{\mathbf{Z}}^{\otimes k}$*-almost all* $(\mathbf{z}_1, \ldots, \mathbf{z}_k)$, $\hat{\theta}(\mathbf{z}_1, \ldots, \mathbf{z}_k) \xrightarrow{\mathbb{P}} \theta(\mathbf{z}_1, \ldots, \mathbf{z}_k)$ *as* $n \to \infty$.

We introduce now a few more notation to state the asymptotic normality of $\hat{\theta}$. For $1 \leq j, l, m \leq k$ and $\mathbf{z}_1, \ldots, \mathbf{z}_{3k} \in \mathcal{Z}^{3k}$, define

$$\begin{aligned}
\theta_{j,l}(\mathbf{z}_1, \ldots, \mathbf{z}_k) &:= \mathbb{E}\big[g(\mathbf{X}_1, \ldots, \mathbf{X}_{j-1}, \mathbf{X}, \mathbf{X}_{j+1}, \ldots, \mathbf{X}_k)g(\mathbf{X}_{k+1}, \ldots, \mathbf{X}_{k+l-1}, \mathbf{X}, \mathbf{X}_{k+l+1}, \ldots, \mathbf{X}_{2k}) \\
&\quad \big|\mathbf{Z} = \mathbf{z}_j \ ; \ \mathbf{Z}_i = \mathbf{z}_i, \forall i = 1, \ldots, k, i \neq j \ ; \ \mathbf{Z}_{k+i} = \mathbf{z}_i, \forall i = 1, \ldots, k, i \neq l\big], \\
\tilde{\theta}_{j,l}(\mathbf{z}_1, \ldots, \mathbf{z}_{2k}) &:= \mathbb{E}\big[g(\mathbf{X}_1, \ldots, \mathbf{X}_{j-1}, \mathbf{X}, \mathbf{X}_{j+1}, \ldots, \mathbf{X}_k)g(\mathbf{X}_{k+1}, \ldots, \mathbf{X}_{k+l-1}, \mathbf{X}, \mathbf{X}_{k+l+1}, \ldots, \mathbf{X}_{2k}) \\
&\quad \big|\mathbf{Z} = \mathbf{z}_j \ ; \ \mathbf{Z}_i = \mathbf{z}_i, \forall i = 1, \ldots, 2k, i \notin \{j, k+l\}\big]. \qquad (8.6) \\
\theta_{j,l,m}(\mathbf{z}_1, \ldots, \mathbf{z}_{3k}) &:= \mathbb{E}\big[g(\mathbf{X}_1, \ldots, \mathbf{X}_{j-1}, \mathbf{X}, \mathbf{X}_{j+1}, \ldots, \mathbf{X}_k) \\
&\quad g(\mathbf{X}_{k+1}, \ldots, \mathbf{X}_{k+l-1}, \mathbf{X}, \mathbf{X}_{k+l+1}, \ldots, \mathbf{X}_{2k})g(\mathbf{X}_{2k+1}, \ldots, \mathbf{X}_{2k+m-1}, \mathbf{X}, \mathbf{X}_{2k+m+1}, \ldots, \mathbf{X}_{3k}) \\
&\quad \big|\mathbf{Z} = \mathbf{z}_j \ ; \ \mathbf{Z}_i = \mathbf{z}_i, \forall i = 1, \ldots, 3k, i \notin \{j, k+l, 2k+m\}\big].
\end{aligned}$$

Assumption 8.2.10. *(i)* $h_n \to 0$ *and* $nh_n^p \to \infty$ *;*

(ii) K *is symmetric at* 0 *and bounded with compact support ;*

(iii) $\theta_{j,l}$ *is continuous at* $(\mathbf{z}_1, \ldots, \mathbf{z}_k)$ *for all* $1 \leq j, l \leq k$ *;*

(iv) θ *is two times continuously differentiable in a neighborhood of* $(\mathbf{z}_1, \ldots, \mathbf{z}_k)$ *;*

(v) $\theta_{j,l,m}$ *is bounded in a neighborhood of the point* $(\mathbf{z}_1, \ldots, \mathbf{z}_k, \mathbf{z}_1, \ldots, \mathbf{z}_k, \mathbf{z}_1, \ldots, \mathbf{z}_k) \in \mathcal{Z}^{3k}$ *for all* $1 \leq j, l, m \leq k$ *;*

(vi) $f_{\mathbf{Z}}$ *is twice differentiable in neighborhoods of* $\mathbf{z}_i, 1 \leq i \leq k$.

Proposition 8.7 (Asymptotic normality of $\hat{\theta}$, Corollary 2.4 in Stute [132])**.** *Under Assumption 8.2.10,* $\sqrt{nh_n^p}\big(\hat{\theta}(\mathbf{z}_1, \ldots, \mathbf{z}_k) - \theta(\mathbf{z}_1, \ldots, \mathbf{z}_k)\big) \xrightarrow{D} \mathcal{N}(0, \rho^2)$,
where $\rho^2 := \sum_{j,l=1}^k \mathbb{1}_{\{\mathbf{z}_j = \mathbf{z}_l\}}\big(\theta_{j,l}(\mathbf{z}_1, \ldots, \mathbf{z}_k) - \theta^2(\mathbf{z}_1, \ldots, \mathbf{z}_k)\big)||K||_2^2 / f_{\mathbf{Z}}(\mathbf{z}_j)$.

Moreover, let N *be a positive integer, and* $\big(\mathbf{z}_1^{(1)}, \ldots, \mathbf{z}_k^{(1)}, \ldots, \mathbf{z}_1^{(N)}, \ldots, \mathbf{z}_k^{(N)}\big) \in \mathcal{Z}^{k\times N}$*. Then under similar regularity conditions,* $\sqrt{nh_n^p}\big(\hat{\theta}(\mathbf{z}_1^{(i)}, \ldots, \mathbf{z}_k^{(i)}) - \theta(\mathbf{z}_1^{(i)}, \ldots, \mathbf{z}_k^{(i)})\big)_{i=1,\ldots,N} \xrightarrow{D} \mathcal{N}(0, \mathbb{H})$*, where for* $1 \leq \tilde{j}, \tilde{l} \leq N$*,*

$$[\mathbb{H}]_{\tilde{j},\tilde{l}} := \sum_{j,l=1}^k \mathbb{1}_{\left\{\mathbf{z}_j^{(\tilde{j})} = \mathbf{z}_l^{(\tilde{l})}\right\}}\left(\tilde{\theta}_{j,l}\left(\mathbf{z}_1^{(\tilde{j})}, \ldots, \mathbf{z}_k^{(\tilde{j})}, \mathbf{z}_1^{(\tilde{l})}, \ldots, \mathbf{z}_k^{(\tilde{l})}\right) - \theta\left(\mathbf{z}_1^{(\tilde{j})}, \ldots, \mathbf{z}_k^{(\tilde{j})}\right)\theta\left(\mathbf{z}_1^{(\tilde{l})}, \ldots, \mathbf{z}_k^{(\tilde{l})}\right)\right)\frac{||K||_2^2}{f_{\mathbf{Z}}\left(\mathbf{z}_j^{(\tilde{j})}\right)}.$$

Note that the second part of Proposition 8.7 above is a consequence of the first one. Indeed, for every $(c_1, \ldots, c_N) \in \mathbb{R}^N$, we can define $\theta\big(\mathbf{z}_1^{(1)}, \ldots, \mathbf{z}_k^{(1)}, \ldots, \mathbf{z}_1^{(N)}, \ldots, \mathbf{z}_k^{(N)}\big) := \sum_{\tilde{i}=1}^N c_{\tilde{i}}\theta(\mathbf{z}_1^{(\tilde{i})}, \ldots, \mathbf{z}_k^{(\tilde{i})})$ and corresponding versions of g, $\hat{\theta}$ and ρ^2. Finally, the conclusion follows from the Cramér-Wold device.

8.3 Theoretical properties of the estimator $\hat{\beta}$

Let us define the matrix $\mathbb{Z}$ of dimension $|\mathfrak{I}_{k,n'}| \times r$ by $[\mathbb{Z}']_{i,j} := \psi_j\big(\mathbf{z}'_{\sigma_i(1)}, \ldots, \mathbf{z}'_{\sigma_i(k)}\big)$, where $1 \leq i \leq |\mathfrak{I}_{k,n'}|$, $1 \leq j \leq r$ and σ_i is the i-th element of $\mathfrak{I}_{k,n'}$. The chosen order of $\mathfrak{I}_{k,n'}$ is arbitrary and has no impact in practice. In the same way, we define the vector $\mathbf{Y}$ of dimension $|\mathfrak{I}_{k,n'}|$ defined by $Y_i := \Lambda\Big(\hat{\theta}\big(\mathbf{z}'_{\sigma_i(1)}, \ldots, \mathbf{z}'_{\sigma_i(k)}\big)\Big)$, such that the criterion (8.3) is in the standard Lasso form

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^r} \Big[||\mathbf{Y} - \mathbb{Z}'\beta||^2 + \lambda|\beta|_1\Big],$$

where for any vector $\mathbf{v}$ of size $|\mathfrak{I}_{k,n'}|$, its scaled norm is defined by $||\mathbf{v}|| := |\mathbf{v}|_2/\sqrt{|\mathfrak{I}_{k,n'}|}$. Following [39], we define $\xi_{i,n}$, for $1 \leq i \leq |\mathfrak{I}_{k,n'}|$, by

$$\begin{aligned}\xi_{i,n} = \xi_{\sigma_i,n} &:= \Lambda\Big(\hat{\theta}\big(\mathbf{z}'_{\sigma_i(1)}, \ldots, \mathbf{z}'_{\sigma_i(k)}\big)\Big) - \boldsymbol{\psi}\big(\mathbf{z}'_{\sigma_i(1)}, \ldots, \mathbf{z}'_{\sigma_i(k)}\big)^T\beta^* \\ &= \Lambda\Big(\hat{\theta}\big(\mathbf{z}'_{\sigma_i(1)}, \ldots, \mathbf{z}'_{\sigma_i(k)}\big)\Big) - \Lambda\Big(\theta\big(\mathbf{z}'_{\sigma_i(1)}, \ldots, \mathbf{z}'_{\sigma_i(k)}\big)\Big).\end{aligned}$$

8.3.1 Non-asymptotic bounds on $\hat{\theta}$

We will also use the *Restricted Eigenvalue* (RE) condition, introduced by Bickel, Ritov and Tsybakov [19]. For $c_0 > 0$ and $s \in \{1, \ldots, p\}$, it is defined as follows,

$RE(s, c_0)$ **condition :** *The design matrix $\mathbb{Z}'$ satisfies*

$$\kappa(s, c_0) := \min_{\substack{J_0 \subset \{1,\ldots,r\} \\ |J_0| \leq s}} \min_{\substack{\delta \neq 0 \\ |\delta_{J_0^C}|_1 \leq c_0|\delta_{J_0}|_1}} \frac{||\mathbb{Z}'\delta||}{|\delta|_2} > 0.$$

Note that this condition is very mild, and is satisfied with a high probability for a large class of random matrices: see Bellec et al. [12, Section 8.1] for references and a discussion. We will also need the following regularity assumption on the function $\Lambda(\cdot)$.

Assumption 8.3.1. *The function $\mathbf{z} \mapsto \boldsymbol{\psi}(\mathbf{z})$ are bounded on $\mathcal{Z}$ by a constant $C_{\boldsymbol{\psi}}$. Moreover, $\Lambda(\cdot)$ is continuously differentiable. Let $\mathcal{T}$ be the range of θ, from $\mathcal{Z}^k$ towards $\mathbb{R}$. On an open neighborhood of $\mathcal{T}$, the derivative of $\Lambda(\cdot)$ is bounded by a constant $C_{\Lambda'}$.*

The following theorem is proved in Section 8.6.3.

Theorem 8.8. *Assume either Assumption 8.2.7 or the weaker Assumption 8.2.8. Suppose that Assumptions 8.2.1-8.2.6 and 8.3.1 hold and that the design matrix $\mathbb{Z}'$ satisfies the $RE(s, 3)$ condition. Choose the tuning parameter as $\lambda = \gamma t$, with $\gamma \geq 4$ and $t > 0$, and assume that we choose h small enough such that*

$$h \leq \min\left(\Big(\frac{f_{\mathbf{Z},min}\alpha!}{4\,C_{K,\alpha}}\Big)^{1/\alpha},\ \Big(\frac{t}{2C_5C_8}\Big)^{1/(k+\alpha)}\right), \tag{8.7}$$

where $C_8 := C_{\boldsymbol{\psi}}C_{\Lambda'}\big(1 + C_4 f_{\mathbf{Z},min}/2\big)$. Then, we have

$$\begin{aligned}&\mathbb{P}\Big(||\mathbb{Z}'(\hat{\beta} - \beta^*)|| \leq \frac{4(\gamma+1)t\sqrt{s}}{\kappa(s,3)} \text{ and } |\hat{\beta} - \beta^*|_q \leq \frac{4^{2/q}(\gamma+1)ts^{1/q}}{\kappa^2(s,3)},\ \text{for every } 1 \leq q \leq 2\Big) \\ &\qquad \geq 1 - 2\sum_{\sigma \in \mathfrak{I}_{k,n'}}\left[\exp\Big(-\frac{[n/k]f^2_{\mathbf{Z},min}h^{kp}}{16C_1 + 4C_2 f_{\mathbf{Z},min}}\Big) + \exp\Big(-\frac{[n/k]t^2h^{kp}}{4C_8^2C_{6,\sigma} + 2C_8C_{7,\sigma}t}\Big)\right].\end{aligned} \tag{8.8}$$

If Assumption 8.2.7 is satisfied, the result holds with $C_{6,\sigma}$ and $C_{7,\sigma}$ constant, respectively to C_6 and C_7 defined in Proposition 8.5 ; in the case of Assumption 8.2.8, the result holds with the following alternative values: $C_{6,\sigma} := 128\big(B_g(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}) + \tilde{B}_g\big)^2 C_K^{2k} f_{\mathbf{Z},min}^{-2k}$ *and*
$C_{7,\sigma} := 2\big(B_g(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}) + \tilde{B}_g\big) C_K^{k} f_{\mathbf{Z},min}^{-k}$.

The latter theorem gives some bounds that hold in probability for the prediction error $||\mathbb{Z}'(\hat{\beta}-\beta^*)||_{n'}$ and for the estimation error $|\hat{\beta}-\beta^*|_q$ with $1 \le q \le 2$ under the specification (8.2). Note that the influence of n' and r is hidden through the Restricted Eigenvalue number $\kappa(s,3)$. The result depends on three parameters to be chosen: γ, t, h. The choice of γ is easy, as a larger γ can only deteriorate the bound. We therefore have an interest in choosing the smallest γ possible, i.e. $\gamma = 3$.

The choice of the parameters t and h is more difficult. Indeed, looking at Equation (8.8), we can see that a lower t gives a smaller bound, but at the same time the probability of this event (which probability is bounded in Equation (8.8)) is lower. This induces a trade-off between the probability of the event and the size of the bound, as we want the smallest possible bound with the highest possible probability. Moreover, we cannot choose t too small, because of the lower bound (8.7): t is limited by a value proportional to h^α. On the one hand, we could choose a small h, but on the other hand, the probability of the event in Equation (8.8) will decrease. Summing up this reasoning in other words, we get the following conclusion: **low values of h and t give a smaller bound, while high values of h and t give a high probability**. Therefore, there is some kind of compromise to do, depending of the kind of bound one is looking for.

8.3.2 Asymptotic properties of $\hat{\beta}$ when $n \to \infty$ and for fixed n'

In this part, n' is still assumed to be fixed and we state the consistency and the asymptotic normality of $\hat{\beta}$ as $n \to \infty$. As above, we adopt a fixed design: the $\mathbf{z}'_i$ are arbitrarily fixed or, equivalently, our reasoning are made conditionally on the second sample. In this section, we follow Derumigny and Fermanian [39] by giving similar results. Proofs are identical and therefore omitted.

For $n, n' > 0$, denote by $\hat{\beta}_{n,n'}$ the estimator (8.3) with $h = h_n$ and $\lambda = \lambda_{n,n'}$.

Lemma 8.9. *We have* $\hat{\beta}_{n,n'} = \arg\min_{\beta\in\mathbb{R}^{p'}} \mathbb{G}_{n,n'}(\beta)$*, where*

$$\begin{aligned}\mathbb{G}_{n,n'}(\beta) := {} & \frac{2(n'-k)!}{n'!} \sum_{\sigma\in\mathfrak{I}_{k,n'}} \xi_{\sigma,n} \boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)^T(\beta^*-\beta) \\ & + \frac{(n'-k)!}{n'!} \sum_{\sigma\in\mathfrak{I}_{k,n'}} \big\{\boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)^T(\beta^*-\beta)\big\}^2 + \lambda_{n,n'}|\beta|_1. \end{aligned} \tag{8.9}$$

Theorem 8.10 (Consistency of $\hat{\beta}$)**.** *Under Assumption 8.2.9, if n' is fixed and $\lambda = \lambda_{n,n'} \to \lambda_0$, then, given $\mathbf{z}'_1,\dots,\mathbf{z}'_{n'}$ and as n tends to the infinity, $\hat{\beta}_{n,n'} \xrightarrow{\mathbb{P}} \beta^{**} := \inf_\beta \mathbb{G}_{\infty,n'}(\beta)$, where*

$$\mathbb{G}_{\infty,n'}(\beta) := \frac{1}{n'} \sum_{\sigma\in\mathfrak{I}_{k,n'}} \Big(\boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)^T(\beta^*-\beta)\Big)^2 + \lambda_0|\beta|_1.$$

In particular, if $\lambda_0 = 0$ and $< \{\boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big) : \sigma \in \mathfrak{I}_{k,n'}\} > = \mathbb{R}^r$, then $\hat{\beta}_{n,n'} \xrightarrow{\mathbb{P}} \beta^*$.

Theorem 8.11 (Asymptotic law of the estimator)**.** *Under Assumption 8.2.10, and if* $\lambda_{n,n'}(nh_{n,n'}^p)^{1/2}$ *tends to* ℓ *when* $n \to \infty$*, we have* $(nh_{n,n'}^p)^{1/2}(\hat{\beta}_{n,n'} - \beta^*) \xrightarrow{D} \mathbf{u}^* := \arg\min_{\mathbf{u}\in\mathbb{R}^r} \mathbb{F}_{\infty,n'}(\mathbf{u})$*, given* $\mathbf{z}'_1, \ldots, \mathbf{z}'_{n'}$*, where*

$$\mathbb{F}_{\infty,n'}(\mathbf{u}) := \frac{2(n'-k)!}{n'!} \sum_{\sigma\in\mathfrak{I}_{k,n'}} \sum_{j=1}^{r} W_\sigma \psi_j\big(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\big) u_j + \frac{(n'-k)!}{n'!} \sum_{\sigma\in\mathfrak{I}_{k,n'}} \Big(\boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\big)^T \mathbf{u}\Big)^2$$
$$+ \ell \sum_{i=1}^{r} \big(|u_i| \mathbb{1}_{\{\beta_i^*=0\}} + u_i \operatorname{sign}(\beta_i^*) \mathbb{1}_{\{\beta_i^*\neq 0\}}\big),$$

with $\mathbf{W} = (W_\sigma)_{\sigma\in\mathfrak{I}_{k,n'}} \sim \mathcal{N}\left(0, \tilde{\mathbb{H}}\right)$ *where*

$$[\tilde{\mathbb{H}}]_{\sigma,\varsigma} := \sum_{j,l=1}^{k} \mathbb{1}_{\left\{\mathbf{z}'_{\sigma(j)} = \mathbf{z}'_{\varsigma(l)}\right\}} \frac{||K||_2^2}{f_{\mathbf{Z}}\left(\mathbf{z}'_{\sigma(j)}\right)} \Lambda'\left(\theta\left(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\right)\right) \Lambda'\left(\theta\left(\mathbf{z}'_{\varsigma(1)}, \ldots, \mathbf{z}'_{\varsigma(k)}\right)\right)$$
$$\cdot \left(\tilde{\theta}_{j,l}\left(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}, \mathbf{z}'_{\varsigma(1)}, \ldots, \mathbf{z}'_{\varsigma(k)}\right) - \theta\left(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\right) \theta\left(\mathbf{z}'_{\varsigma(1)}, \ldots, \mathbf{z}'_{\varsigma(k)}\right)\right),$$

and $\tilde{\theta}_{j,l}$ *is as defined in Equation (8.6).*

Moreover, $\limsup_{n\to\infty} \mathbb{P}\left(\mathcal{S}_n = \mathcal{S}\right) = c < 1$*, where* $\mathcal{S}_n := \{j : \hat{\beta}_j \neq 0\}$ *and* $\mathcal{S} := \{j : \beta_j \neq 0\}$.

A usual way of obtaining the oracle property is to modify our estimator in an "adaptive" way. Following Zou [146], consider a preliminary "rough" estimator of β^*, denoted by $\tilde{\beta}_n$, or more simply $\tilde{\beta}$. Moreover $\nu_n(\tilde{\beta}_n - \beta^*)$ is assumed to be asymptotically normal, for some deterministic sequence (ν_n) that tends to the infinity. Now, let us consider the same optimization program as in (8.3) but with a random tuning parameter given by $\lambda_{n,n'} := \tilde{\lambda}_{n,n'}/|\tilde{\beta}_n|^\delta$, for some constant $\delta > 0$ and some positive deterministic sequence $(\tilde{\lambda}_{n,n'})$. The corresponding adaptive estimator (solution of the modified Equation (8.3)) will be denoted by $\check{\beta}_{n,n'}$, or simply $\check{\beta}$. Hereafter, we still set $\mathcal{S}_n = \{j : \check{\beta}_j \neq 0\}$.

Theorem 8.12 (Asymptotic law of the adaptive estimator of β)**.** *Under Assumption 8.2.10, if the following convergence hold* $\tilde{\lambda}_{n,n'}(nh_{n,n'}^p)^{1/2} \to \ell \geq 0$ *and* $\tilde{\lambda}_{n,n'}(nh_{n,n'}^p)^{1/2}\nu_n^\delta \to \infty$ *when* $n \to \infty$*, we have*

$$(nh_{n,n'}^p)^{1/2}(\check{\beta}_{n,n'} - \beta^*)_{\mathcal{S}} \xrightarrow{D} \mathbf{u}^{**}_{\mathcal{S}} := \underset{\mathbf{u}_{\mathcal{S}}\in\mathbb{R}^s}{\arg\min}\, \check{\mathbb{F}}_{\infty,n'}(\mathbf{u}_{\mathcal{S}}),$$ *where*

$$\check{\mathbb{F}}_{\infty,n'}(\mathbf{u}_{\mathcal{S}}) := \frac{2(n'-k)!}{n'!} \sum_{\sigma\in\mathfrak{I}_{k,n'}} \sum_{j\in\mathcal{S}} W_\sigma \psi_j(\mathbf{z}'_i) u_j + \frac{(n'-k)!}{n'!} \sum_{\sigma\in\mathfrak{I}_{k,n'}} \Big(\sum_{j\in\mathcal{S}} \psi_j(\mathbf{z}'_i) u_j\Big)^2$$
$$+ \ell \sum_{i\in\mathcal{S}} \frac{u_i}{|\beta_i^*|^\delta} \operatorname{sign}(\beta_i^*),$$

with $\mathbf{W} = (W_\sigma)_{\sigma\in\mathfrak{I}_{k,n'}} \sim \mathcal{N}\big(0, \tilde{\mathbb{H}}\big)$.
Moreover, when $\ell = 0$*, the oracle property is fulfilled:* $\mathbb{P}\left(\mathcal{S}_n = \mathcal{S}\right) \to 1$ *as* $n \to \infty$.

8.3.3 Asymptotic properties of $\hat{\beta}$ jointly in (n, n')

Now, we consider the framework in which both n and n' are going to infinity, while the dimensions p and r stay fixed. We now provide a consistency result for $\hat{\beta}_{n,n'}$.

Theorem 8.13 (Consistency of $\hat{\beta}_{n,n'}$, jointly in (n, n'))**.** *Assume that Assumptions 8.2.1- 8.3.1 are satisfied. Assume that* $\sum_{\sigma\in\mathfrak{I}_{k,n'}} \boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\big)\boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\big)^T/n'$ *converges to a matrix* $M_{\psi,\mathbf{z}'}$*, as* $n' \to \infty$*. Assume that* $\lambda_{n,n'} \to \lambda_0$ *and* $n' \exp(-Anh^{kp}) \to 0$ *for every* $A > 0$*, when* $(n, n') \to \infty$*. Then* $\hat{\beta}_{n,n'} \xrightarrow{\mathbb{P}} \arg\min_{\beta\in\mathbb{R}^r} \mathbb{G}_{\infty,\infty}(\beta)$*, as* $(n, n') \to \infty$*, where* $\mathbb{G}_{\infty,\infty}(\beta) := (\beta^* - \beta)M_{\psi,\mathbf{z}'}(\beta^* - \beta)^T + \lambda_0|\beta|_1$*. Moreover, if* $\lambda_0 = 0$ *and* $M_{\psi,\mathbf{z}'}$ *is invertible, then* $\hat{\beta}_{n,n'}$ *is consistent and tends to the true value* β^*.

Note that, since the sequence $(\mathbf{z}'_i)$ is deterministic, we only assume the convergence of the sequence of deterministic matrices $\sum_{\sigma\in\mathfrak{I}_{k,n'}} \boldsymbol{\psi}(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)})\boldsymbol{\psi}(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)})^T/n'$ in $\mathbb{R}^{r^2}$. Moreover, if the "second subset" $(\mathbf{z}'_i)_{i=1,\dots,n'}$ were a random sample (drawn along the law $\mathbb{P}_{\mathbf{Z}}$), the latter convergence would be understood "in probability". And if $\mathbb{P}_{\mathbf{Z}}$ satisfies the identifiability condition (Proposition 8.1), then $M_{\psi,\mathbf{z}'}$ would be invertible and $\hat{\beta}_{n,n'} \to \beta^*$ in probability. Now, we want to go one step further and derive the asymptotic law of the estimator $\hat{\beta}_{n,n'}$. For this, we will need the following assumption.

Assumption 8.3.2. *(i) The support of the kernel $K(\cdot)$ is included into $[-1,1]^p$. Moreover, for all n, n' and every $(i,j) \in \{1,\dots,n'\}^2$, $i \neq j$, we have $|\mathbf{z}'_i - \mathbf{z}'_j|_\infty > 2h_{n,n'}$.*

(ii) (a) $n'(nh_{n,n'}^{p+4\alpha} + h_{n,n'}^{2\alpha} + h_{n,n'}^{p} + (nh_{n,n'}^{p})^{-1}) \to 0$, (b) $\lambda_{n,n'}(n'\,n\,h_{n,n'}^{p})^{1/2} \to 0$, (c) $n'\,n\,h_{n,n'}^{p} \to \infty$ and $n\,h_{n,n'}^{p+2\alpha-\epsilon}/\ln n' \to \infty$ for some $\epsilon \in [0, 2\alpha[$.

(iii) The distribution $\mathbb{P}_{\mathbf{z}',n'} := |\mathfrak{I}_{k,n'}|^{-1}\sum_{\sigma\in\mathfrak{I}_{k,n'}} \delta_{(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)})}$ weakly converges as $n' \to \infty$, to a distribution $\mathbb{P}_{\mathbf{z}',k,\infty}$ on $\mathbb{R}^{kp}$. There exists a distribution $\mathbb{P}_{\mathbf{z}',\infty}$ on $\mathbb{R}^{kp}$, with a density $f_{\mathbf{z}',\infty}$ with respect to the p-dimensional Lebesgue measure such that $\mathbb{P}_{\mathbf{z}',k,\infty} = \mathbb{P}_{\mathbf{z}',\infty}^{\otimes k}$.

(iv) The matrix $V_1 := \int \boldsymbol{\psi}(\mathbf{z}'_1,\dots\mathbf{z}'_k)\boldsymbol{\psi}(\mathbf{z}'_1,\dots\mathbf{z}'_k)^T f_{\mathbf{z}',\infty}(\mathbf{z}'_1)\cdots f_{\mathbf{z}',\infty}(\mathbf{z}'_k)d\mathbf{z}'_1\cdots d\mathbf{z}'_k$ is non-singular.

(v) $\Lambda(\cdot)$ is two times continuously differentiable. Let $\mathcal{T}$ be the range of θ, from $\mathcal{Z}^k$ towards $\mathbb{R}$. On an open neighborhood of $\mathcal{T}$, the second derivative of $\Lambda(\cdot)$ is bounded by a constant $C_{\Lambda''}$.

(vi) some integrals exists and are finite. We will especially need the folowing ones.

$$\tilde{V}_1 := \int \theta(\mathbf{z}'_1,\dots,\mathbf{z}'_k)\Lambda'\Big(\theta(\mathbf{z}'_1,\dots,\mathbf{z}'_k)\Big)\boldsymbol{\psi}(\mathbf{z}'_1,\dots,\mathbf{z}'_k) f_{\mathbf{z}',\infty}(\mathbf{z}_1)\cdots f_{\mathbf{z}',\infty}(\mathbf{z}_k)\,d\mathbf{z}_1\cdots d\mathbf{z}_k\,;$$

$$V_2 := \int \frac{||K||_2^2}{f_{\mathbf{Z}}(\mathbf{z}'_1)} g(\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_k)g(\mathbf{x}_1,\mathbf{y}_2,\dots,\mathbf{y}_k)\Lambda'^2\big(\theta(\mathbf{z}'_1,\dots,\mathbf{z}'_k)\big)\boldsymbol{\psi}(\mathbf{z}'_1,\dots,\mathbf{z}'_k)\boldsymbol{\psi}(\mathbf{z}'_1,\dots,\mathbf{z}'_k)^T$$
$$\times f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_1}(\mathbf{x}_1)\,d\mu(\mathbf{x}_1)d\mu(\mathbf{z}'_1)\prod_{i=2}^{k} f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_i}(\mathbf{y}_i)f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_i}(\mathbf{x}_i)f_{\mathbf{z}',\infty}(\mathbf{z}'_i)\,d\mu(\mathbf{x}_i)d\mu(\mathbf{y}_i)d\mathbf{z}'_i.$$

Theorem 8.14 (Asymptotic law of $\hat{\beta}_{n,n'}$, jointly in (n,n'))**.** *Under Assumptions 8.2.1-8.2.5 and under Assumption 8.3.2, we have*

$$(n \times n' \times h_{n,n'}^{p})^{1/2}(\hat{\beta}_{n,n'} - \beta^*) \xrightarrow{D} \mathcal{N}(0, \tilde{V}_{as}),$$

where $\tilde{V}_{as} := V_1^{-1}V_2V_1^{-1}$, V_1 is the matrix defined in Assumption 8.3.2(iv), and V_2 in Assumption 8.3.2(v).

This theorem is proved in Section 8.7.

8.4 Applications and examples

Following Example 4.4 in Stute [132], we consider the function $g(x_1,x_2) := \mathbb{1}\{x_1 \le x_2\}$, with $k=2$. In this case $\theta(\mathbf{z}_1,\mathbf{z}_2) = \mathbb{P}(X_1 \le X_2 | \mathbf{Z}_1 = \mathbf{z}_1, \mathbf{Z}_2 = \mathbf{z}_2)$. The parameter $\theta(\mathbf{z}_1,\mathbf{z}_2)$ quantifies the probability that the quantity of interest X be smaller if we knew that $\mathbf{Z} = \mathbf{z}_1$ than if we knew that $\mathbf{Z} = \mathbf{z}_2$.

To illustrate our methods, we choose a simple example, with the Epanechnikov kernel, defined by $K(x) := (3/4)(1-u^2)\mathbb{1}|u| \le 1$. It is a kernel of order $\alpha = 2$, with $\int K^2 = 3/5$. Assumption 8.2.1 is then satisfied with $C_K := 3/4$. Fix $p=1$, $\mathcal{Z} = [-1,1]$, $\mathcal{X} = \mathbb{R}$, $f_Z(z) = \phi(z)\mathbb{1}\{|z| \le 1\}/(1-2\Phi(-1))$, where

Φ and ϕ are respectively the cdf and the density of the standard Gaussian distribution and $X|Z=z \sim \mathcal{N}(z,1)$, for every $z \in \mathcal{Z}$. For any $z_1, z_2 \in \mathcal{Z}$,

$$\begin{aligned}
&\int \Big|K\big(\mathbf{u}_1\big)\cdots K\big(\mathbf{u}_k\big)\Big| \sum_{m_1+\cdots+m_k=\alpha} \binom{\alpha}{m_{1:k}} \\
&\qquad \cdot \prod_{i=1}^{k} \sum_{j_1,\ldots,j_{m_i}=1}^{p} \big|u_{i,j_1}\ldots u_{i,j_{m_i}}\big| \sup_{t\in[0,1]} \left|\frac{\partial^{m_i} f_{\mathbf{Z}}}{\partial z_{j_1}\cdots\partial z_{j_{m_i}}}\big(\mathbf{z}_i + t\mathbf{u}_i\big)\right| d\mathbf{u}_1\cdots d\mathbf{u}_k \\
&= \int K(u_1)K(u_2) \sum_{m_1+m_2=2} \frac{2}{m_1!m_2!} \cdot \prod_{i=1}^{k} \big|u_i^{m_i}\big| \sup_{t\in[0,1]} \left|\frac{\partial^{m_i} f_Z}{\partial z^{m_i}}\big(\mathbf{z}_i + t\mathbf{u}_i\big)\right| du_1 du_2 \\
&\leq \int K(u_1)K(u_2)\cdot \Bigg(\big|u_2^2\big| \sup_{z\in\mathcal{Z}} |f_Z(z)| \sup_{z\in\mathcal{Z}} \left|\frac{\partial^2 f_Z}{\partial z^2}(z)\right| \\
&+ \frac{1}{2}\big|u_1u_2\big| \sup_{z\in\mathcal{Z}} \left|\frac{\partial f_Z}{\partial z}(z)\right| \sup_{z\in\mathcal{Z}} \left|\frac{\partial f_Z}{\partial z}(z)\right| + \big|u_1^2\big| \sup_{z\in\mathcal{Z}} \left|\frac{\partial^2 f_Z}{\partial z^2}(z)\right| \sup_{t\in[0,1]} |f_Z(z)| \Bigg) du_1 du_2 \\
&\leq 2\times 0.59^2 \times \int K(u)u^2 du + \frac{1}{2}\times 0.36^2 \Big(\int |u|K(u)du\Big)^2 \leq 0.2.
\end{aligned}$$

Assumption 8.2.2 is then satisfied with $C_{K,\alpha} = 0.2$. Assumption 8.2.3 is easily satisfied with $f_{Z,\max} = 1/\big(\sqrt{2\pi}(1-2\Phi(-1))\big) \leq 0.59$. Therefore, we can apply Lemma 8.3. We compute the constants $C_1 := 2f_{\mathbf{Z},max}^k ||K||_2^{2k} = 2\times 0.59^2 \times (3/5)^2 \leq 0.26$ and $C_2 := (4/3)C_K^k = (4/3)\times(3/4)^2 = 3/4$. Therefore, for any $n \geq 0$, $h, t > 0$, $z_1, z_2 \in \mathcal{Z}$, we have

$$\mathbb{P}\Big(\big|N_2(z_1,z_2) - f_Z(z_1)f_Z(z_2)\big| \leq 0.1h^\alpha + t\Big) \geq 1 - 2\exp\Big(-\frac{[n/2]t^2}{0.26h^2 + 0.75h^2 t}\Big),$$

Assumption 8.2.4 is satisfied with $f_{Z,\min} = \phi(1)/(1-2\Phi(-1)) > 0.35$, so that we can apply Corollary 8.4. Therefore, the estimator $\hat{\theta}(z_1,z_2)$ exists with probability greater than

$$1 - 2\exp\Big(-\frac{(n-1)h^2\big(0.35-0.1h^2\big)^2}{0.52 + 1.5\times\big(0.35-0.1h^2\big)}\Big),$$

Note that this probability is greater than 0.99 as soon as $n \geq 3\big(0.52 + 1.5\times(0.35-0.1h^2)\big)/\big(h^2(0.35-0.1h^2)^2\big)$. For example, with $h = 0.2$, it means that the estimator $\hat{\theta}(z_1,z_2)$ exists with a probability greater than 99% as soon as n is greater than 651.

We list below other possible examples of applications. Conditional moments constitute also a natural class of U-statistics. They include the conditional variance ($p_{\mathbf{X}} = 1$, $k = 2$, $g(X_1,X_2) = X_1^2 - X_1\cdot X_2$) and the conditional covariance ($p_{\mathbf{X}} = 2$, $k = 2$, $g(\mathbf{X}_1,\mathbf{X}_2) := X_{1,1}\cdot X_{2,1} - X_{1,1}\cdot X_{2,2}$). The conditional variance gives information about the volatility of X given the variable $\mathbf{Z}$. Conditional covariances can be used to describe how the dependence moves as a function of the conditioning variables $\mathbf{Z}$. Higher-order conditional moments (skewness, kurtosis, and so on) can also be estimated by higher-order conditional U-statistics, and they described respectively how the asymmetry and the behavior of the tails of X change as function of Z.

Gini's mean difference, an indicator of dispersion, can also be used in this framework. Formally, it is defined as the U-statistic with $p_{\mathbf{X}} = 1$, $k = 2$ and $g(X_1,X_2) := |X_1 - X_2|$. Its conditional version describes how two variables are far away, on average, given their conditioning variables $\mathbf{Z}$. for example, X could be the income of an individual, $\mathbf{Z}$ could be the position of its home, and $\theta(\mathbf{z}_1,\mathbf{z}_2)$ represent the average inequality between the income of two persons, one at point $\mathbf{z}_1$ and the other at point $\mathbf{z}_2$.

Other conditional dependence measures can also be written as conditional U-statistics, see e.g. Example 1.1.7 of Koroljuk and Borovskisch [84]. They show how a U-statistic of order $k = 5$ can be used

to estimated the dependence parameter

$$\theta = \iint \big(F_{1,2}(x,y) - F_{1,2}(x,\infty)F_{1,2}(\infty,y)\big)dF_{1,2}(x,y).$$

In our framework, we could consider a conditional version, given by

$$\theta(\mathbf{z}_1,\mathbf{z}_2) = \iint \big(F_{1,2|\mathbf{Z}=\mathbf{z}}(x,y) - F_{1,2|\mathbf{Z}=\mathbf{z}}(x,\infty)F_{1,2|\mathbf{Z}=\mathbf{z}}(\infty,y)\big)dF_{1,2|\mathbf{Z}=\mathbf{z}}(x,y),$$

where $\mathbf{X}$ is of dimension $p_{\mathbf{X}} = 2$. Finally, the conditional Kendall's tau is also included in our framework, with $p_{\mathbf{X}} = 2$, $k = 2$, $g(\mathbf{X}_1,\mathbf{X}_2) := 4 \cdot \mathbb{1}_{\{\mathbf{X}_1 < \mathbf{X}_2\}} - 1$.

8.5 Notations

In the proofs, we will use the following shortcut notations: $\mathbf{x}_{1:k}$ denotes the k-tuple $(\mathbf{x}_1,\dots,\mathbf{x}_k)$. Similarly, for a function σ, $\sigma(1:k)$ denotes the tuple $(\sigma(1),\dots,\sigma(k))$, and $\mathbf{X}_{\sigma(1:k)}$ is the k-tuple $(\mathbf{X}_{\sigma(1)},\dots,\mathbf{X}_{\sigma(k)})$. For any variable Y and any collection of given points $(\mathbf{z}_1,\dots\mathbf{z}_k)$, the conditional expectation $\mathbb{E}[Y|\mathbf{Z}_{1:k} = \mathbf{z}_{1:k}]$ denotes $\mathbb{E}[Y|\mathbf{Z}_1 = \mathbf{z}_1,\dots,\mathbf{Z}_k = \mathbf{z}_k]$. We denote by $\int \phi(\mathbf{z}_{1:k})d\mathbf{z}_{1:k}$ the integral $\int \phi(\mathbf{z}_1,\dots,\mathbf{z}_k)d\mathbf{z}_1\cdots d\mathbf{z}_k$ for any integrable function $\phi : \mathbb{R}^{k\times p} \to \mathbb{R}$, and by $\int g(\mathbf{x}_{1:k})d\mu^{\otimes k}(\mathbf{x}_{1:k})$ the integral $\int g(\mathbf{z}_1,\dots,\mathbf{z}_k)d\mu(\mathbf{x}_1)\cdots d\mu(\mathbf{x}_k)$ for any μ-integrable function $g : \mathcal{X}^k \to \mathbb{R}$.

8.6 Finite distance proofs for $\hat{\theta}$ and $\hat{\beta}$

For convenience, we recall Berk's (1970) inequality (see Theorem A in Serfling [123, p.201]). Note that, if $m = 1$, this reduces to Bernstein's inequality.

Lemma 8.15. *Let $k > 0$, $n \geq k$, $\mathbf{X}_1,\dots,\mathbf{X}_n$ i.i.d. random vectors with values in a measurable space $\mathcal{X}$ and $g : \mathcal{X}^k \to [a,b]$ be a real bounded function. Set $\theta := \mathbb{E}[g(\mathbf{X}_{1:k})]$ and $\sigma^2 := Var[g(\mathbf{X}_{1:k})]$. Then, for any $t > 0$,*

$$\mathbb{P}\left(\binom{n}{k}^{-1}\sum_{\sigma\in\mathfrak{I}^{\uparrow}_{k,n}} g\left(\mathbf{X}_{\sigma(1:k)}\right) - \theta \geq t\right) \leq \exp\left(-\frac{[n/k]t^2}{2\sigma^2 + (2/3)(b-\theta)t}\right),$$

where $\mathfrak{I}_{k,n}$ is the set of bijective functions from $\{1,\dots,k\}$ to $\{1,\dots,n\}$ and $\mathfrak{I}^{\uparrow}_{k,n}$ is the subset of $\mathfrak{I}_{k,n}$ made of increasing functions.

Note that g does not need to be symmetric for this bound to hold. Indeed, if g is not symmetric, we can nonetheless apply this lemma to the symmetrized version $\tilde{g}$ defined as $\tilde{g}(\mathbf{x}_{1:k}) := (k!)^{-1}\sum_{\sigma\in\mathfrak{I}_{k,k}} g(\mathbf{x}_{\sigma(1:k)})$, and we get the result.

8.6.1 Proof of Lemma 8.3

We decompose the quantity to bound into a stochastic part and a bias as follows:

$$N_k(\mathbf{z}_{1:k}) - \prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i) = \big(N_k(\mathbf{z}_{1:k}) - \mathbb{E}[N_k(\mathbf{z}_{1:k})]\big) + \big(\mathbb{E}[N_k(\mathbf{z}_{1:k})] - \prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i)\big).$$

We first bound the bias.

$$\begin{aligned}\left|\mathbb{E}\big[N_k(\mathbf{z}_{1:k})\big] - \prod_{i=1}^k f_{\mathbf{Z}}(\mathbf{z}_i)\right| &= \left|\mathbb{E}\Big[\binom{n}{k}^{-1} \sum_{\sigma\in\mathfrak{I}_{k,n}} \prod_{i=1}^k K_h\big(\mathbf{Z}_{\sigma(i)} - \mathbf{z}_i\big)\Big] - \prod_{i=1}^k f_{\mathbf{Z}}(\mathbf{z}_i)\right| \\ &= \left|\int \Big(\prod_{i=1}^k f_{\mathbf{Z}}(\mathbf{z}_i + h\mathbf{u}_i) - \prod_{i=1}^k f_{\mathbf{Z}}(\mathbf{z}_i)\Big) \prod_{i=1}^k K(\mathbf{u}_i) d\mathbf{u}_i\right| \\ &= \left|\int \Big(\phi_{\mathbf{z},\mathbf{u}}(1) - \phi_{\mathbf{z},\mathbf{u}}(0)\Big) \prod_{i=1}^k K(\mathbf{u}_i) d\mathbf{u}_i\right|,\end{aligned}$$

where $\phi_{\mathbf{z},\mathbf{u}}(t) := \prod_{j=1}^k f_{\mathbf{Z}}\big(\mathbf{z}_i + th\mathbf{u}_j\big)$ for $t \in [-1,1]$. Note that this function has at least the same regularity as $f_{\mathbf{Z}}$, so it is α-differentiable, and by a Taylor-Lagrange expansion, we get

$$\left|\mathbb{E}[N_k(\mathbf{z}_{1:k})] - \prod_{i=1}^k f_{\mathbf{Z}}(\mathbf{z}_i)\right| = \left|\int_{\mathbb{R}^{kp}} \left(\sum_{i=1}^{\alpha-1} \frac{1}{i!}\phi^{(i)}_{\mathbf{z},\mathbf{u}}(0) + \frac{1}{\alpha!}\phi^{(\alpha)}_{\mathbf{z},\mathbf{u}}(t_{\mathbf{z},\mathbf{u}})\right) \prod_{i=1}^k K(\mathbf{u}_i) d\mathbf{u}_i\right|.$$

For $l > 0$, we have

$$\begin{aligned}\phi^{(l)}_{\mathbf{z},\mathbf{u}}(0) &= \sum_{m_1+\cdots+m_k=l} \binom{\alpha}{m_{1:k}} \prod_{i=1}^k \frac{\partial^{m_i}\Big(f_{\mathbf{Z}}\big(\mathbf{z}_i + ht\mathbf{u}_i\big)\Big)}{\partial t^{m_i}}(0) \\ &= \sum_{m_1+\cdots+m_k=l} \binom{\alpha}{m_{1:k}} \prod_{i=1}^k \sum_{j_1,\ldots,j_{m_i}=1}^p h^{m_i} u_{i,j_1}\ldots u_{i,j_{m_i}} \frac{\partial^{m_i} f_{\mathbf{Z}}}{\partial z_{j_1}\cdots\partial z_{j_{m_i}}}\big(\mathbf{z}_i + t_{\mathbf{z},\mathbf{u}} h\mathbf{u}_i\big),\end{aligned}$$

where $\binom{\alpha}{m_{1:k}} := \alpha!/\big(\prod_{i=1}^k (m_i!)\big)$ is the multinomial coefficient. Using Assumption 8.2.1, for every $i = 1,\ldots,\alpha-1$, we get $\int K(\mathbf{u}_1)\cdots K(\mathbf{u}_k)\phi^{(i)}_{\mathbf{z},\mathbf{u}}(0) d\mathbf{u}_1\cdots d\mathbf{u}_k = 0$. Therefore, only the last term remains and we have

$$\left|\mathbb{E}[N_k(\mathbf{z}_{1:k})] - \prod_{i=1}^k f_{\mathbf{Z}}(\mathbf{z}_i)\right| = \left|\int \left(\frac{1}{\alpha!}\phi^{(\alpha)}_{\mathbf{z},\mathbf{u}}(t_{\mathbf{z},\mathbf{u}})\right) \prod_{i=1}^k K(\mathbf{u}_i) d\mathbf{u}_i\right| \le \frac{C_{K,\alpha}}{\alpha!} h^\alpha,$$

using Assumption 8.2.2.

Second, we bound the stochastic part. We have

$$N_k(\mathbf{z}_{1:k}) - \mathbb{E}[N_k(\mathbf{z}_{1:k})] = \frac{k!(n-k)!}{n!} \sum_{\sigma\in\mathfrak{I}^{\uparrow}_{k,n}} \prod_{i=1}^k K_h\big(\mathbf{Z}_{\sigma(i)} - \mathbf{z}_i\big) - \prod_{i=1}^k \mathbb{E}[K_h\big(\mathbf{Z}_i - \mathbf{z}_i\big)].$$

Then, we can apply Lemma 8.15 to the function g defined by $g(\tilde{\mathbf{z}}_1,\ldots,\tilde{\mathbf{z}}_k) := \prod_{i=1}^k K_h\big(\tilde{\mathbf{z}}_i - \mathbf{z}_i\big)$. Here, we have $b = -a = h^{-kp} C_K^k$, and

$$Var\big[g(\mathbf{Z}_1,\ldots,\mathbf{Z}_k)^2\big] \le \mathbb{E}\big[g(\mathbf{Z}_1,\ldots,\mathbf{Z}_k)^2\big] = \prod_{i=1}^k \mathbb{E}\big[K_h\big(\mathbf{Z}_i - \mathbf{z}_i\big)^2\big] \le h^{-kp} f^k_{\mathbf{Z},max} ||K||_2^{2k}.$$

Finally, we get

$$\mathbb{P}\left(\binom{n}{k}^{-1} N_k(\mathbf{z}_{1:k}) - \mathbb{E}[N_k(\mathbf{z}_{1:k})] \ge t\right) \le \exp\left(-\frac{[n/k]t^2}{2h^{-kp} f^k_{\mathbf{Z},max}||K||_2^{2k} + (4/3)h^{-kp} C_K^k t}\right),$$

□

8.6.2 Proof of Proposition 8.5

We have the following decomposition

$$|\hat{\theta}(\mathbf{z}_{1:k}) - \theta(\mathbf{z}_{1:k})|$$

$$= \Big| N_k(\mathbf{z}_{1:k})^{-1} \frac{(n-k)!}{n!} \sum_{\sigma \in \mathfrak{I}_{k,n}} \prod_{i=1}^{k} K_h(\mathbf{Z}_{\sigma(i)} - \mathbf{z}_i) \Big(g(\mathbf{X}_{\sigma(1:k)}) - \mathbb{E}\big[g(\mathbf{X}_{1:k}) \big| \mathbf{Z}_{1:k} = \mathbf{z}_{1:k} \big] \Big) \Big|$$

$$= \frac{\prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i)}{N_k(\mathbf{z}_1, \ldots, \mathbf{z}_k)} \cdot \Big| \frac{(n-k)!}{n!} \sum_{\sigma \in \mathfrak{I}_{k,n}} \prod_{i=1}^{k} \frac{K_h(\mathbf{Z}_{\sigma(i)} - \mathbf{z}_i)}{f_{\mathbf{Z}}(\mathbf{z}_i)} \Big(g(\mathbf{X}_{\sigma(1:k)}) - \mathbb{E}\big[g(\mathbf{X}_{1:k}) \big| \mathbf{Z}_{1:k} = \mathbf{z}_{1:k} \big] \Big) \Big|$$

$$=: \frac{\prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i)}{N_k(\mathbf{z}_1, \ldots, \mathbf{z}_k)} \cdot \Big| \sum_{\sigma \in \mathfrak{I}_{k,n}} S_\sigma \Big|.$$

The conclusion will follow from the next three lemmas, where we will bound separately $\prod_{i=1}^{k} f_{\mathbf{Z}}/N_k$, the bias term $\big| \sum_{\sigma \in \mathfrak{I}_{k,n}} \mathbb{E}[S_\sigma] \big|$ and the stochastic component $\big| \sum_{\sigma \in \mathfrak{I}_{k,n}} (S_\sigma - \mathbb{E}[S_\sigma]) \big|$.

Lemma 8.16 (Bound for $\prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i)/N_k$)**.** *Under Assumptions 8.2.1, 8.2.2, 8.2.3, and 8.2.4 and if for some $t > 0$, $C_{K,\alpha} h^\alpha / \alpha! + t < f_{\mathbf{Z},min}^k / 2$, we have*

$$\mathbb{P}\left(\left| \frac{1}{N_k(\mathbf{z}_{1:k})} - \frac{1}{\prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i)} \right| \leq \frac{4}{f_{\mathbf{Z},min}^{2k}} \left(\frac{C_{K,\alpha} h^\alpha}{\alpha!} + t \right) \right)$$
$$\geq 1 - 2 \exp\left(- \frac{[n/k] t^2}{2 h^{-kp} f_{\mathbf{Z},max}^k ||K||_2^{2k} + (4/3) h^{-kp} C_K^k t} \right),$$

and on the same event, $N_k(\mathbf{z}_{1:k})$ is strictly positive and

$$\frac{\prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i)}{N_k(\mathbf{z}_{1:k})} \leq 1 + \frac{4 f_{\mathbf{Z},max}^k}{f_{\mathbf{Z},min}^{2k}} \left(\frac{C_{K,\alpha} h^\alpha}{\alpha!} + t \right).$$

Proof : Using the mean value inequality for the function $x \mapsto 1/x$, we get

$$\left| \frac{1}{N_k(\mathbf{z}_{1:k})} - \frac{1}{\prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i)} \right| \leq \frac{1}{N_*^2} \big| N_k(\mathbf{z}_{1:k}) - \prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i) \big|,$$

where N_* lies between $N_k(\mathbf{z}_{1:k})$ and $\prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i)$. By Lemma 8.3, we get

$$\mathbb{P}\left(\big| N_k(\mathbf{z}_{1:k}) - \prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i) \big| \leq \frac{C_{K,\alpha}}{\alpha!} h^\alpha + t \right) \geq 1 - 2 \exp\left(- \frac{[n/k] t^2}{2 h^{-kp} f_{\mathbf{Z},max}^k ||K||_2^{2k} + (4/3) h^{-kp} C_K^k t} \right).$$

On this event, $\big| N_k(\mathbf{z}_{1:k}) - \prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i) \big| \leq (1/2) \prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i)$ by assumption, so that $f_{\mathbf{Z},min}^k / 2 \leq N_k(\mathbf{z}_{1:k})$. We have also $f_{\mathbf{Z},min}^k / 2 \leq \prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i)$. Thus, we have $f_{\mathbf{Z},min}^k / 2 \leq N_*$. Combining the previous inequalities, we finally get

$$\left| \frac{1}{N_k(\mathbf{z}_{1:k})} - \frac{1}{\prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i)} \right| \leq \frac{1}{N_*^2} \big| N_k(\mathbf{z}_{1:k}) - \prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i) \big| \leq \frac{4}{f_{\mathbf{Z},min}^{2k}} \left(\frac{C_{K,\alpha} h^\alpha}{\alpha!} + t \right).$$

□

Now, we provide a bound on the bias.

Lemma 8.17. *Under Assumptions 8.2.1 and 8.2.6, we have $\big| \mathbb{E}[S_\sigma] \big| \leq C_{g,f,\alpha} C_{K,\alpha} h^{k\alpha} / (f_{\mathbf{Z},min}^k \alpha!)$.*

Proof : We remark that

$$0 = \int \Big(g(\mathbf{x}_{1:k}) - \mathbb{E}\big[g(\mathbf{X}_{1:k}) \big| \mathbf{Z}_{1:k} = \mathbf{z}_{1:k} \big] \Big) f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_1}(\mathbf{x}_1) \cdots f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}_k}(\mathbf{x}_k) d\mu^{\otimes k}(\mathbf{x}_{1:k})$$
$$= \int \Big(g(\mathbf{x}_{1:k}) - \mathbb{E}\big[g(\mathbf{X}_{1:k}) \big| \mathbf{Z}_{1:k} = \mathbf{z}_{1:k} \big] \Big) \frac{f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1, \mathbf{z}_1) \cdots f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_k, \mathbf{z}_k)}{\prod_{i=1}^{k} f_{\mathbf{Z}}(\mathbf{z}_i)} d\mu^{\otimes k}(\mathbf{x}_{1:k}). \tag{8.10}$$

We have

$$\mathbb{E}[S_\sigma] = \mathbb{E}\Big[\frac{K_h(\mathbf{Z}_{\sigma(1)}-\mathbf{z}_1)\cdots K_h(\mathbf{Z}_{\sigma(k)}-\mathbf{z}_k)}{\prod_{i=1}^k f_{\mathbf{Z}}(\mathbf{z}_i)}\Big(g(\mathbf{X}_{\sigma(1)},\ldots,\mathbf{X}_{\sigma(k)}) - \mathbb{E}\big[g(\mathbf{X}_{1:k})\big|\mathbf{Z}_{1:k}=\mathbf{z}_{1:k}\big]\Big)\Big]$$

$$= \int \Big(g(\mathbf{x}_{1:k}) - \mathbb{E}\big[g(\mathbf{X}_{1:k})\big|\mathbf{Z}_{1:k}=\mathbf{z}_{1:k}\big]\Big)\prod_{i=1}^k \frac{K(\mathbf{u}_i)}{f_{\mathbf{Z}}(\mathbf{z}_i)} f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i,\mathbf{z}_i+h\mathbf{u}_i)\,d\mu(\mathbf{x}_i)d\mathbf{u}_i$$

$$= \int \Big(g(\mathbf{x}_{1:k}) - \mathbb{E}\big[g(\mathbf{X}_{1:k})\big|\mathbf{Z}_{1:k}=\mathbf{z}_{1:k}\big]\Big)\Big(\prod_{i=1}^k f_{\mathbf{X},\mathbf{Z}}\Big(\mathbf{x}_i,\mathbf{z}_i+h\mathbf{u}_i\Big) - \prod_{i=1}^k f_{\mathbf{X},\mathbf{Z}}\Big(\mathbf{x}_i,\mathbf{z}_i\Big)\Big)\prod_{i=1}^k \frac{K(\mathbf{u}_i)}{f_{\mathbf{Z}}(\mathbf{z}_i)}\,d\mu(\mathbf{x}_i)d\mathbf{u}_i.$$

We apply now the Taylor-Lagrange formula to the function

$$\phi_{\mathbf{x}_{1:k},\mathbf{u}_{1:k}}(t) := \prod_{i=1}^k f_{\mathbf{X},\mathbf{Z}}\Big(\mathbf{x}_i,\mathbf{z}_i+h\mathbf{u}_i\Big),$$

and get

$$\mathbb{E}[S_\sigma] = \int \Big(g(\mathbf{x}_{1:k}) - \mathbb{E}\big[g(\mathbf{X}_{1:k})\big|\mathbf{Z}_{1:k}=\mathbf{z}_{1:k}\big]\Big)\Big(\phi_{\mathbf{x}_{1:k},\mathbf{u}_{1:k}}(t)(1)-\phi_{\mathbf{x}_{1:k},\mathbf{u}_{1:k}}(t)(0)\Big)\prod_{i=1}^k \frac{K(\mathbf{u}_i)}{f_{\mathbf{Z}}(\mathbf{z}_i)}\,d\mu(\mathbf{x}_i)d\mathbf{u}_i$$

$$= \int \Big(g(\mathbf{x}_{1:k}) - \mathbb{E}\big[g(\mathbf{X}_{1:k})\big|\mathbf{Z}_{1:k}=\mathbf{z}_{1:k}\big]\Big)$$
$$\cdot\Big(\sum_{j=1}^{\alpha-1}\frac{1}{j!}\phi_{\mathbf{x}_{1:k},\mathbf{u}_{1:k}}(t)^{(j)}(0) + \frac{1}{\alpha!}\phi_{\mathbf{x}_{1:k},\mathbf{u}_{1:k}}(t)^{(\alpha)}(t_{\mathbf{x},\mathbf{u}})\Big)\prod_{i=1}^k \frac{K(\mathbf{u}_i)}{f_{\mathbf{Z}}(\mathbf{z}_i)}\,d\mu(\mathbf{x}_i)d\mathbf{u}_i$$

$$= \int \Big(g(\mathbf{x}_{1:k}) - \mathbb{E}\big[g(\mathbf{X}_{1:k})\big|\mathbf{Z}_{1:k}=\mathbf{z}_{1:k}\big]\Big)$$
$$\cdot\Big(\frac{1}{\alpha!}\phi_{\mathbf{x}_{1:k},\mathbf{u}_{1:k}}(t)^{(\alpha)}(t_{\mathbf{x},\mathbf{u}})\Big)\prod_{i=1}^k \frac{K(\mathbf{u}_i)}{f_{\mathbf{Z}}(\mathbf{z}_i)}\,d\mu(\mathbf{x}_i)d\mathbf{u}_i$$

$$= \int \Big(g(\mathbf{x}_{1:k}) - \mathbb{E}\big[g(\mathbf{X}_{1:k})\big|\mathbf{Z}_{1:k}=\mathbf{z}_{1:k}\big]\Big)$$
$$\cdot\frac{1}{\alpha!}\Big(\phi_{\mathbf{x}_{1:k},\mathbf{u}_{1:k}}(t)^{(\alpha)}(t_{\mathbf{x},\mathbf{u}}) - \phi_{\mathbf{x}_{1:k},\mathbf{u}_{1:k}}(t)^{(\alpha)}(0)\Big)\prod_{i=1}^k \frac{K(\mathbf{u}_i)}{f_{\mathbf{Z}}(\mathbf{z}_i)}\,d\mu(\mathbf{x}_i)d\mathbf{u}_i.$$

For every real t, we have

$$\phi^{(\alpha)}(t) = \sum_{m_1+\cdots+m_k=\alpha}\binom{n}{m_{1:k}}\prod_{i=1}^k \frac{\partial^{m_i}\Big(f_{\mathbf{X},\mathbf{Z}}\big(\mathbf{x}_i,\mathbf{z}_i+ht\mathbf{u}_i\big)\Big)}{\partial t^{m_i}}$$

$$= \sum_{m_1+\cdots+m_k=\alpha}\binom{n}{m_{1:k}}\prod_{i=1}^k\sum_{j_1,\ldots,j_{m_i}=1}^p h^{m_i}u_{i,j_1}\ldots u_{i,j_{m_i}}\frac{\partial^{m_i}f_{\mathbf{X},\mathbf{Z}}}{\partial z_{j_1}\cdots\partial z_{j_{m_i}}}\big(\mathbf{x}_i,\mathbf{z}_i+ht\mathbf{u}_i\big)$$

$$= h^\alpha\sum_{m_1+\cdots+m_k=\alpha}\binom{n}{m_{1:k}}\prod_{i=1}^k\sum_{j_1,\ldots,j_{m_i}=1}^p u_{i,j_1}\ldots u_{i,j_{m_i}}\frac{\partial^{m_i}f_{\mathbf{X},\mathbf{Z}}}{\partial z_{j_1}\cdots\partial z_{j_{m_i}}}\big(\mathbf{x}_i,\mathbf{z}_i+ht\mathbf{u}_i\big). \tag{8.11}$$

Therefore, we get

$$\mathbb{E}[S_\sigma] = \sum_{m_1+\cdots+m_k=\alpha}\binom{n}{m_{1:k}}\int\prod_{i=1}^k\frac{K(\mathbf{u}_i)}{\prod_{i=1}^k f_{\mathbf{Z}}(\mathbf{z}_i)}\sum_{j_1,\ldots,j_{m_i}=1}^p u_{i,j_1}\ldots u_{i,j_{m_i}}$$
$$\cdot\Big(g(\mathbf{x}_{1:k}) - \mathbb{E}\big[g(\mathbf{X}_{1:k})\big|\mathbf{Z}_{1:k}=\mathbf{z}_{1:k}\big]\Big)$$
$$\cdot\left(\frac{\partial^{m_i}f_{\mathbf{X},\mathbf{Z}}}{\partial z_{j_1}\cdots\partial z_{j_{m_i}}}\big(\mathbf{x}_i,\mathbf{z}_i+ht\mathbf{u}_i\big) - \frac{\partial^{m_i}f_{\mathbf{X},\mathbf{Z}}}{\partial z_{j_1}\cdots\partial z_{j_{m_i}}}\big(\mathbf{x}_i,\mathbf{z}_i\big)\right)d\mu(\mathbf{x}_1)d\mathbf{u}_1\cdots d\mu(\mathbf{x}_k)d\mathbf{u}_k,$$

and, using Assumption 8.2.6, this yields

$$\big|\mathbb{E}[S_\sigma]\big| \le \frac{C_{g,f,\alpha}C_{K,\alpha}h^{\alpha+k}}{f_{\mathbf{Z},min}^k\alpha!}.$$

□

Now we bound the stochastic component. We have the following equality

$$\Big|\sum_{\sigma\in\mathfrak{I}_{k,n}}\big(S_\sigma-\mathbb{E}[S_\sigma]\big)\Big| = \Big|\frac{(n-k)!}{n!}\sum_{\sigma\in\mathfrak{I}_{k,n}} g\big((\mathbf{X}_{\sigma(1)},\mathbf{Z}_{\sigma(1)}),\dots,(\mathbf{X}_{\sigma(k)},\mathbf{Z}_{\sigma(k)})\big)\Big|$$

with the function $\tilde{g}$ defined by

$$\begin{aligned}
&\tilde{g}\big((\mathbf{X}_1,\mathbf{Z}_1),\dots,(\mathbf{X}_k,\mathbf{Z}_k)\big)\\
&=\frac{K_h\big(\mathbf{Z}_1-\mathbf{z}_1\big)\cdots K_h\big(\mathbf{Z}_k-\mathbf{z}_k\big)}{\prod_{i=1}^k f_{\mathbf{Z}}(\mathbf{z}_i)}\Big(g(\mathbf{X}_{1:k})-\mathbb{E}\big[g(\mathbf{X}_{1:k})\big|\mathbf{Z}_{1:k}=\mathbf{z}_{1:k}\big]\Big)\\
&-\mathbb{E}\left[\frac{K_h\big(\mathbf{Z}_1-\mathbf{z}_1\big)\cdots K_h\big(\mathbf{Z}_k-\mathbf{z}_k\big)}{\prod_{i=1}^k f_{\mathbf{Z}}(\mathbf{z}_i)}\Big(g(\mathbf{X}_{1:k})-\mathbb{E}\big[g(\mathbf{X}_{1:k})\big|\mathbf{Z}_{1:k}=\mathbf{z}_{1:k}\big]\Big)\right]
\end{aligned}$$

By construction, $\mathbb{E}\Big[\tilde{g}\big((\mathbf{X}_1,\mathbf{Z}_1),\dots,(\mathbf{X}_k,\mathbf{Z}_k)\big)\Big]=0$. If $\tilde{g}$ is bounded, we can derive an immediate bound for this stochastic component. Indeed, we would have $||\tilde{g}||_\infty \le 4C_K^k h^{-kp} C_g^k / f_{\mathbf{Z},min}^k$. Moreover, we have

$$\begin{aligned}
Var\Big[\tilde{g}\big((\mathbf{X}_1,\mathbf{Z}_1),\dots,(\mathbf{X}_k,\mathbf{Z}_k)\big)\Big] &\le \mathbb{E}\left[\frac{K_h^2\big(\mathbf{Z}_1-\mathbf{z}_1\big)\cdots K_h^2\big(\mathbf{Z}_k-\mathbf{z}_k\big)}{\prod_{i=1}^k f_{\mathbf{Z}}^2(\mathbf{z}_i)} g^2(\mathbf{X}_1,\dots,\mathbf{X}_k)\right]\\
&\le C_g^2 f_{\mathbf{Z},max}^k f_{\mathbf{Z},min}^{-2k} h^{-kp}||K||_2^{2k}.
\end{aligned}$$

Therefore, we can apply Lemma 8.15, and we get

$$\mathbb{P}\Big(\Big|\sum_{\sigma\in\mathfrak{I}_{k,n}}\big(S_\sigma-\mathbb{E}[S_\sigma]\big)\Big|>t\Big)\le 2\exp\left(-\frac{[n/k]t^2}{2C_g^2 f_{\mathbf{Z},max}^k f_{\mathbf{Z},min}^{-2k} h^{-kp}||K||_2^{2k}+(8/3)C_K^k h^{-kp} C_g^k f_{\mathbf{Z},min}^{-k} t}\right).$$

In the following Lemma 8.18, our goal will be to bound the stochastic component using only Assumption 8.2.8 on the conditional moments of g.

Lemma 8.18. *Under Assumptions 8.2.1, 8.2.4 and 8.2.8, for every $t>0$, we have*

$$\mathbb{P}\left(\sum_{\sigma\in\mathfrak{I}_{k,n}} S_\sigma-\mathbb{E}[S_\sigma]>t\right)\le\exp\left(-\frac{t^2 f_{\mathbf{Z},min}^{2k} h^{kp}[n/k]}{128\big(B_{g,\mathbf{z}}+\tilde{B}_g\big)^2 C_K^{2k-1}+2t\big(B_{g,\mathbf{z}}+\tilde{B}_g\big)C_K^k f_{\mathbf{Z},min}^k}\right).$$

Proof: Using the same decomposition for U-statistics as in Hoeffding [70], we obtain

$$\sum_{\sigma\in\mathfrak{I}_{k,n}} S_\sigma-\mathbb{E}[S_\sigma]=\frac{1}{n!}\sum_{\sigma\in\mathfrak{I}_{n,n}}\frac{1}{[n/k]}\sum_{i=1}^{[n/k]} V_{n,i,\sigma},$$

where

$$V_{n,i,\sigma}:=\tilde{g}\Big(\big(\mathbf{X}_{\sigma(1+(i-1)k)},\mathbf{Z}_{\sigma(2+(i-1)k)}\big),\dots,\big(\mathbf{X}_{\sigma(ik)},\mathbf{Z}_{\sigma(jk)}\big)\Big).$$

For any $\lambda>0$, we have

$$\begin{aligned}
\mathbb{P}\left(\sum_{\sigma\in\mathfrak{I}_{k,n}} S_\sigma-\mathbb{E}[S_\sigma]>t\right) &\le e^{-\lambda t}\mathbb{E}\left[\exp\left(\lambda\sum_{\sigma\in\mathfrak{I}_{k,n}} S_\sigma-\mathbb{E}[S_\sigma]\right)\right]\\
&\le e^{-\lambda t}\mathbb{E}\left[\exp\left(\lambda\frac{1}{n!}\sum_{\sigma\in\mathfrak{I}_{n,n}}\frac{1}{[n/k]}\sum_{i=1}^{[n/k]} V_{n,i,\sigma}\right)\right]
\end{aligned}$$

$$\leq e^{-\lambda t}\frac{1}{n!}\sum_{\sigma\in\mathfrak{I}_{n,n}}\mathbb{E}\left[\exp\left(\lambda\frac{1}{[n/k]}\sum_{i=1}^{[n/k]}V_{n,i,\sigma}\right)\right]$$
$$\leq e^{-\lambda t}\frac{1}{n!}\sum_{\sigma\in\mathfrak{I}_{n,n}}\prod_{i=1}^{[n/k]}\mathbb{E}\left[\exp\left(\lambda\frac{1}{[n/k]}V_{n,i,\sigma}\right)\right]$$
$$\leq e^{-\lambda t}\left(\sup_{\sigma\in\mathfrak{I}_{n,n},\, i=1,\dots,[n/k]}\mathbb{E}\left[\exp\left(\lambda[n/k]^{-1}V_{n,i,\sigma}\right)\right]\right)^{[n/k]}. \tag{8.12}$$

Let $l\geq 2$. Using the inequality $(a+b+c+d)^l\leq 4^l(a^l+b^l+c^l+d^l)$, we get

$$\mathbb{E}\big[|V_{n,i,\sigma}|^l\big]=\mathbb{E}\big[|V_{n,1,\sigma}|^l\big]\leq 4^l\,\mathbb{E}\Big[|g(\mathbf{X}_{\sigma(1)},\dots,\mathbf{X}_{\sigma(k)})|^l\prod_{i=1}^k\frac{|K_h|^l\big(\mathbf{Z}_{\sigma(i)}-\mathbf{z}_i\big)}{f^l_{\mathbf{Z}}(\mathbf{z}_i)}\Big]$$
$$+4^l\,\mathbb{E}\Big[\big|\mathbb{E}\big[g(\mathbf{X}_{1:k})\big|\mathbf{Z}_{1:k}=\mathbf{z}_{1:k}\big]\big|^l\prod_{i=1}^k\frac{|K_h|^l\big(\mathbf{Z}_{\sigma(i)}-\mathbf{z}_i\big)}{f^l_{\mathbf{Z}}(\mathbf{z}_i)}\Big]$$
$$+4^l\left|\mathbb{E}\Big[g(\mathbf{X}_{\sigma(1)},\dots,\mathbf{X}_{\sigma(k)})\prod_{i=1}^k\frac{K_h\big(\mathbf{Z}_{\sigma(i)}-\mathbf{z}_i\big)}{f^l_{\mathbf{Z}}(\mathbf{z}_i)}\Big]\right|^l$$
$$+4^l\left|\mathbb{E}\Big[\big|\mathbb{E}\big[g(\mathbf{X}_{1:k})\big|\mathbf{Z}_{1:k}=\mathbf{z}_{1:k}\big]\big|\prod_{i=1}^k\frac{K_h\big(\mathbf{Z}_{\sigma(i)}-\mathbf{z}_i\big)}{f^l_{\mathbf{Z}}(\mathbf{z}_i)}\Big]\right|^l$$

Using Jensen's inequality for the function $x\mapsto|x|^p$ with the second, third and fourth terms, and the law of iterated expectations for the first and the third terms, we get

$$\mathbb{E}\big[|V_{n,i,\sigma}|^l\big]\leq 4^l\cdot 2\,\mathbb{E}\Big[\mathbb{E}\big[|g(\mathbf{X}_{\sigma(1)},\dots,\mathbf{X}_{\sigma(k)})|^l\big|\mathbf{Z}_{\sigma(1)},\dots,\mathbf{Z}_{\sigma(k)}\big]\prod_{i=1}^k\frac{|K_h|^l\big(\mathbf{Z}_{\sigma(i)}-\mathbf{z}_i\big)}{f^l_{\mathbf{Z}}(\mathbf{z}_i)}\Big]$$
$$+4^l\cdot 2\,\mathbb{E}\Big[\mathbb{E}\big[\big|g(\mathbf{X}_{1:k})\big|^l\big|\mathbf{Z}_i=\mathbf{z}_i,\forall i=1,\dots,k\big]\prod_{i=1}^k\frac{|K_h|^l\big(\mathbf{Z}_{\sigma(i)}-\mathbf{z}_i\big)}{f^l_{\mathbf{Z}}(\mathbf{z}_i)}\Big]$$
$$\leq 4^l\cdot 2\,\mathbb{E}\Big[\Big(B^l_g(\mathbf{Z}_1,\dots,\mathbf{Z}_k)+B^l_g(\mathbf{z}_1,\dots,\mathbf{z}_k)\Big)^l l!\prod_{i=1}^k\frac{|K_h|^l\big(\mathbf{Z}_{\sigma(i)}-\mathbf{z}_i\big)}{f^l_{\mathbf{Z}}(\mathbf{z}_i)}\Big]$$
$$\leq 4^l\cdot 2\,\big(\tilde{B}^l_g+B^l_g(\mathbf{z}_1,\dots,\mathbf{z}_k)\big)l!(h^{-kp}C_K^k f^{-k}_{\mathbf{Z},min})^{l-1}\,f^{-k}_{\mathbf{Z},min}$$
$$\leq 2\Big(4\big(\tilde{B}_g+B_{g,\mathbf{z}}\big)h^{-kp}C_K^k f^{-k}_{\mathbf{Z},min}\Big)^l l!\,h^{kp}C_K^{-1},$$

where $B_{g,\mathbf{z}}:=B_g(\mathbf{z}_1,\dots,\mathbf{z}_k)$. Remarking that $\mathbb{E}[V_{n,i,\sigma}]=0$ by construction of $\tilde{g}$, we obtain

$$\mathbb{E}\left[\exp\left(\lambda[n/k]^{-1}V_{n,i,\sigma}\right)\right]=1+\sum_{l=2}^\infty\frac{\mathbb{E}\big[(\lambda[n/k]^{-1}V_{n,i,\sigma})^l\big]}{l!}$$
$$\leq 1+2C_K^{-1}h^{kp}\sum_{l=2}^\infty(4\lambda[n/k]^{-1}\big(B_{g,\mathbf{z}}+\tilde{B}_g\big)h^{-kp}C_K^k f^{-k}_{\mathbf{Z},min})^l$$
$$\leq 1+2C_K^{-1}h^{kp}\cdot\frac{\Big(4\lambda[n/k]^{-1}\big(B_{g,\mathbf{z}}+\tilde{B}_g\big)h^{-kp}C_K^k f^{-k}_{\mathbf{Z},min}\Big)^2}{1-4\lambda[n/k]^{-1}\big(B_{g,\mathbf{z}}+\tilde{B}_g\big)h^{-kp}C_K^k f^{-k}_{\mathbf{Z},min}}$$
$$\leq\exp\left(\frac{32\lambda^2[n/k]^{-2}\big(B_{g,\mathbf{z}}+\tilde{B}_g\big)^2h^{-kp}C_K^{2k-1}f^{-2k}_{\mathbf{Z},min}}{1-4\lambda[n/k]^{-1}\big(B_{g,\mathbf{z}}+\tilde{B}_g\big)h^{-kp}C_K^k f^{-k}_{\mathbf{Z},min}}\right),$$

where the last statement follows from the inequality $1+x\leq\exp(x)$. Combining the latter bound with Equation (8.12), we get

$$\mathbb{P}\left(\sum_{\sigma\in\mathfrak{I}_{k,n}}S_\sigma-\mathbb{E}[S_\sigma]>t\right)\leq\exp\left(-\lambda t+\frac{32\lambda^2\big(B_{g,\mathbf{z}}+\tilde{B}_g\big)^2C_K^{2k-1}}{f^{2k}_{\mathbf{Z},min}h^{kp}[n/k]-4\lambda\big(B_{g,\mathbf{z}}+\tilde{B}_g\big)C_K^k f^k_{\mathbf{Z},min}}\right). \tag{8.13}$$

Remarking that the right-hand side term inside the exponential is of the form $-\lambda t + \frac{a\lambda^2}{b-c\lambda}$, we choose the value

$$\lambda_* = \frac{tb}{2a+tc} = \frac{t f_{\mathbf{Z},min}^{2k} h^{kp}[n/k]}{64\big(B_{g,\mathbf{z}} + \tilde{B}_g\big)^2 C_K^{2k-1} + t\big(B_{g,\mathbf{z}} + \tilde{B}_g\big) C_K^k f_{\mathbf{Z},min}^k} \tag{8.14}$$

such that $-\lambda_* t + \frac{a\lambda_*^2}{b-c\lambda_*} = -\frac{t^2 b}{4a+2ct} = -\frac{t}{2}\lambda_*$. Therefore, the right-hand side term of Equation (8.13) can be simplified, and combining this with Equation (8.14), we obtain

$$\mathbb{P}\left(\sum_{\sigma\in\mathfrak{I}_{k,n}} S_\sigma - \mathbb{E}[S_\sigma] > t\right) \leq \exp\left(-\frac{t^2 f_{\mathbf{Z},min}^{2k} h^{kp}[n/k]}{128\big(B_{g,\mathbf{z}} + \tilde{B}_g\big)^2 C_K^{2k-1} + 2t\big(B_{g,\mathbf{z}} + \tilde{B}_g\big) C_K^k f_{\mathbf{Z},min}^k}\right).$$

□

8.6.3 Proof of Theorem 8.8

By Proposition 8.5, for every $t_1, t_2 > 0$ such that $C_{K,\alpha} h^\alpha/\alpha! + t < f_{\mathbf{Z},min}/2$, we have

$$\mathbb{P}\Bigg(|\hat{\theta}(\mathbf{z}_1,\ldots,\mathbf{z}_k) - \theta(\mathbf{z}_1,\ldots,\mathbf{z}_k)| < \big(1 + C_3 h^\alpha + C_4 t_1\big) \times \big(C_5 h^{k+\alpha} + t_2\big)\Bigg)$$
$$\geq 1 - 2\exp\left(-\frac{[n/k] t_1^2 h^{kp}}{C_1 + C_2 t_1}\right) - 2\exp\left(-\frac{[n/k] t_2^2 h^{kp}}{C_6 + C_7 t_2}\right),$$

We apply this proposition to every k-tuple $\big(\mathbf{z}'_{\sigma(1)},\ldots,\mathbf{z}'_{\sigma(k)}\big)$ where $\sigma \in \mathfrak{I}_{k,n'}$. Combining it with Assumption 8.3.1, we get

$$\mathbb{P}\Bigg(\sup_i |\xi_{i,n}| < C_{\Lambda'}\big(1 + C_3 h^\alpha + C_4 t_1\big) \times \big(C_5 h^{k+\alpha} + t_2\big)\Bigg)$$
$$\geq 1 - 2\sum_{i=1}^{|\mathfrak{I}_{k,n'}|}\left[\exp\left(-\frac{[n/k] t_1^2 h^{kp}}{C_1 + C_2 t_1}\right) + \exp\left(-\frac{[n/k] t_2^2 h^{kp}}{C_6 + C_7 t_2}\right)\right],$$

Choosing $t_1 := f_{\mathbf{Z},min}/4$ and using the bound (8.7) on h, we get

$$\mathbb{P}\Bigg(\sup_i |\xi_{i,n}| < C_{\Lambda'}\big(1 + C_3 \frac{f_{\mathbf{Z},min}\alpha!}{4C_{K,\alpha}} + C_4 \frac{f_{\mathbf{Z},min}}{4}\big) \times \big(C_5 h^{k+\alpha} + t_2\big)\Bigg)$$
$$\geq 1 - 2\sum_{i=1}^{|\mathfrak{I}_{k,n'}|}\left[\exp\left(-\frac{[n/k] f_{\mathbf{Z},min}^2 h^{kp}}{16C_1 + 4C_2 f_{\mathbf{Z},min}}\right) + \exp\left(-\frac{[n/k] t_2^2 h^{kp}}{C_6 + C_7 t_2}\right)\right].$$

Choosing $t_2 = t/(2C_8) = t/\Big(2C_{\boldsymbol{\psi}} C_{\Lambda'}\big(1 + C_3 \frac{f_{\mathbf{Z},min}\alpha!}{4C_{K,\alpha}} + C_4 \frac{f_{\mathbf{Z},min}}{4}\big)\Big)$, and using the bound (8.7) on h^α, we get

$$\mathbb{P}\left(\sup_i |\xi_{i,n}| < t/C_{\boldsymbol{\psi}}\right) \geq 1 - 2\sum_{i=1}^{|\mathfrak{I}_{k,n'}|}\left[\exp\left(-\frac{[n/k] f_{\mathbf{Z},min}^2 h^{kp}}{16C_1 + 4C_2 f_{\mathbf{Z},min}}\right) + \exp\left(-\frac{[n/k] t^2 h^{kp}}{4C_8^2 C_6 + 2C_8 C_7 t}\right)\right].$$

On the same event, we have $\max_{j=1,\ldots,p'}\left|\frac{1}{n'}\sum_{i=1}^{n'} Z'_{i,j}\xi_{i,n}\right| \leq t$, by Assumption 8.3.1. The conclusion results from the following lemma.

Lemma 8.19 (From [39, Lemma 25]). *Assume that* $\max_{j=1,\ldots,p'}\left|\frac{1}{n'}\sum_{i=1}^{n'} Z'_{i,j}\xi_{i,n}\right| \leq t$, *for some* $t > 0$, *that the assumption* $RE(s,3)$ *is satisfied, and that the tuning parameter is given by* $\lambda = \gamma t$, *with* $\gamma \geq 4$. *Then,* $||\mathbb{Z}'(\hat{\beta} - \beta^*)|| \leq \frac{4(\gamma+1)t\sqrt{s}}{\kappa(s,3)}$ *and* $|\hat{\beta} - \beta^*|_q \leq \frac{4^{2/q}(\gamma+1)t s^{1/q}}{\kappa^2(s,3)}$, *for every* $1 \leq q \leq 2$.

□

8.7 Proof of Theorem 8.14

Define $\tilde{r}_{n,n'} := (n \times n' \times h_{n,n'}^p)^{1/2}$, $\mathbf{u} := \tilde{r}_{n,n'}(\beta - \beta^*)$ and $\hat{\mathbf{u}}_{n,n'} := \tilde{r}_{n,n'}(\hat{\beta}_{n,n'} - \beta^*)$, so that $\hat{\beta}_{n,n'} = \beta^* + \hat{\mathbf{u}}_{n,n'}/\tilde{r}_{n,n'}$. We define for every $\mathbf{u} \in \mathbb{R}^{p'}$,

$$\mathbb{F}_{n,n'}(\mathbf{u}) := \frac{-2\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \xi_{\sigma,n} \boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\big)^T \mathbf{u} + \frac{1}{|\mathfrak{I}_{k,n'}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \big\{\boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\big)^T \mathbf{u}\big\}^2 + \lambda_{n,n'} \tilde{r}_{n,n'}^2 \Big(\big|\beta^* + \frac{\mathbf{u}}{\tilde{r}_{n,n'}}\big|_1 - |\beta^*|_1 \Big), \tag{8.15}$$

and we obtain $\hat{\mathbf{u}}_{n,n'} = \arg\min_{\mathbf{u} \in \mathbb{R}^{p'}} \mathbb{F}_{n,n'}(\mathbf{u})$ applying Lemma 8.9.

Lemma 8.20. *Under the same assumptions as in Theorem 8.14,*

$$T_1 := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \xi_{\sigma,n} \boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\big) \xrightarrow{D} \mathcal{N}(0, V_2).$$

This lemma is proved in Section 8.7.1. It will help to control the first term of Equation (8.15), which is simply $-2T_1^T \mathbf{u}$.

Concerning the second term of Equation (8.15), using Assumption 8.3.2(iii), we have for every $\mathbf{u} \in \mathbb{R}^{p'}$

$$\frac{1}{|\mathfrak{I}_{k,n'}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \big\{\boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\big)^T \mathbf{u}\big\}^2 \to \int \big(\boldsymbol{\psi}(\mathbf{z}'_1, \ldots, \mathbf{z}'_k)^T \mathbf{u}\big)^2 f_{\mathbf{z}',\infty}(\mathbf{z}'_1) \cdots f_{\mathbf{z}',\infty}(\mathbf{z}'_k)\, d\mathbf{z}'_1 \cdots d\mathbf{z}'_k. \tag{8.16}$$

This has to be read as a convergence of a sequence of real numbers indexed by $\mathbf{u}$, because the design points $\mathbf{z}'_i$ are deterministic. We also have, for any $\mathbf{u} \in \mathbb{R}^{p'}$ and when n is large enough,

$$\big|\beta^* + \frac{\mathbf{u}}{\tilde{r}_{n,n'}}\big|_1 - |\beta^*|_1 = \sum_{i=1}^{p'} \Big(\frac{|u_i|}{\tilde{r}_{n,n'}} \mathbb{1}_{\{\beta_i^* = 0\}} + \frac{u_i}{\tilde{r}_{n,n'}} \operatorname{sign}(\beta_i^*) \mathbb{1}_{\{\beta_i^* \neq 0\}} \Big).$$

Therefore, by Assumption 8.3.2(ii)(b), for every $\mathbf{u} \in \mathbb{R}^{p'}$,

$$\lambda_{n,n'} \tilde{r}_{n,n'}^2 \Big(\big|\beta^* + \frac{\mathbf{u}}{\tilde{r}_{n,n'}}\big|_1 - |\beta^*|_1 \Big) \to 0, \tag{8.17}$$

when (n, n') tends to the infinity. Combining Lemma 8.20 and Equations (8.15-8.17), and defining the function $\mathbb{F}_{\infty,\infty}$ by

$$\mathbb{F}_{\infty,\infty}(\mathbf{u}) := 2\tilde{\mathbf{W}}^T \mathbf{u} + \int \big(\boldsymbol{\psi}(\mathbf{z}'_1, \ldots, \mathbf{z}'_k)^T \mathbf{u}\big)^2 f_{\mathbf{z}',\infty}(\mathbf{z}'_1) \cdots f_{\mathbf{z}',\infty}(\mathbf{z}'_k)\, d\mathbf{z}'_1 \cdots d\mathbf{z}'_k,$$

where $\mathbf{u} \in \mathbb{R}^r$ and $\tilde{W} \sim \mathcal{N}(0, V_2)$, we obtain that every finite-dimensional margin of $\mathbb{F}_{n,n'}$ weakly converges to the corresponding margin of $\mathbb{F}_{\infty,\infty}$. Now, applying the convexity lemma, we get

$$\hat{\mathbf{u}}_{n,n'} \xrightarrow{D} \mathbf{u}_{\infty,\infty}, \text{ where } \mathbf{u}_{\infty,\infty} := \arg\min_{\mathbf{u} \in \mathbb{R}^r} \mathbb{F}_{\infty,\infty}(\mathbf{u}).$$

Since $\mathbb{F}_{\infty,\infty}(\mathbf{u})$ is a continuously differentiable convex function, apply the first-order condition $\nabla \mathbb{F}_{\infty,\infty}(\mathbf{u}) = 0$, which yields

$$2\tilde{\mathbf{W}} + 2 \int \boldsymbol{\psi}(\mathbf{z}'_1, \ldots, \mathbf{z}'_k) \boldsymbol{\psi}(\mathbf{z}'_1, \ldots, \mathbf{z}'_k)^T \mathbf{u}_{\infty,\infty} f_{\mathbf{z}',\infty}(\mathbf{z}'_1) \cdots f_{\mathbf{z}',\infty}(\mathbf{z}'_k)\, d\mathbf{z}'_1 \cdots d\mathbf{z}'_k = 0.$$

As a consequence $\mathbf{u}_{\infty,\infty} = -V_1^{-1} \tilde{\mathbf{W}} \sim \mathcal{N}(0, \tilde{V}_{as})$, using Assumption 8.3.2(iv). We finally obtain $\tilde{r}_{n,n'}\big(\hat{\beta}_{n,n'} - \beta^*\big) \xrightarrow{D} \mathcal{N}\big(0, \tilde{V}_{as}\big)$, as claimed. □

8.7.1 Proof of Lemma 8.20

Using a Taylor expansion yields

$$\begin{aligned}T_1 &:= \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}|}\sum_{\sigma\in\mathfrak{I}_{k,n'}}\xi_{\sigma,n}\boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)\\ &= \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}|}\sum_{\sigma\in\mathfrak{I}_{k,n'}}\Big(\Lambda\Big(\hat{\theta}\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)\Big)-\Lambda\Big(\theta\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)\Big)\Big)\boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)\\ &= T_2+T_3,\end{aligned}$$

where the main term is

$$T_2 := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}|}\sum_{\sigma\in\mathfrak{I}_{k,n'}}\Lambda'\Big(\theta\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)\Big)\Big(\hat{\theta}\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)-\theta\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)\Big)\boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big),$$

and the remainder is

$$T_3 := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}|}\sum_{\sigma\in\mathfrak{I}_{k,n'}}\alpha_{3,\sigma}\cdot\Big(\hat{\theta}\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)-\theta\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)\Big)^2\boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big),$$

with $\forall\sigma\in\mathfrak{I}_{k,n'}$, $|\alpha_{3,\sigma}|\le C_{\Lambda''}/2$, by Assumption 8.3.2(v).

Let us define $\overline{\boldsymbol{\psi}}_\sigma := \Lambda'\Big(\theta\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)\Big)\boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)$, for every $\sigma\in\mathfrak{I}_{k,n'}$. Using the definition (8.1), we rewrite $T_2 := T_4+T_5$ where

$$T_4 := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}|\cdot|\mathfrak{I}_{k,n}|}\sum_{\sigma\in\mathfrak{I}_{k,n'}}\sum_{\varsigma\in\mathfrak{I}_{k,n}}\frac{\prod_{i=1}^k K_h\big(\mathbf{Z}_{\varsigma(i)}-\mathbf{z}'_{\sigma(i)}\big)}{\prod_{i=1}^k f_{\mathbf{Z}}(\mathbf{z}'_{\sigma(i)})}\Big(g\big(\mathbf{X}_{\varsigma(1)},\dots,\mathbf{X}_{\varsigma(k)}\big)-\theta\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)\Big)\overline{\boldsymbol{\psi}}_\sigma,$$

$$\begin{aligned}T_5 := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}|\cdot|\mathfrak{I}_{k,n}|}&\sum_{\sigma\in\mathfrak{I}_{k,n'}}\sum_{\varsigma\in\mathfrak{I}_{k,n}}\prod_{i=1}^k K_h\big(\mathbf{Z}_{\varsigma(i)}-\mathbf{z}'_{\sigma(i)}\big)\Big(g\big(\mathbf{X}_{\varsigma(1)},\dots,\mathbf{X}_{\varsigma(k)}\big)-\theta\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)\Big)\\ &\times\Big(\frac{1}{N_k(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)})}-\frac{1}{\prod_{i=1}^k f_{\mathbf{Z}}(\mathbf{z}'_{\sigma(i)})}\Big)\overline{\boldsymbol{\psi}}_\sigma.\end{aligned}$$

To lighten the notations, we will define $K_{\sigma,\varsigma} := \prod_{i=1}^k K_h\big(\mathbf{Z}_{\varsigma(i)}-\mathbf{z}'_{\sigma(i)}\big)$, $g_\varsigma := g\big(\mathbf{X}_{\varsigma(1)},\dots,\mathbf{X}_{\varsigma(k)}\big)$, $\theta_\sigma := \theta\big(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)}\big)$, $f_{\mathbf{Z}',\sigma} := \prod_{i=1}^k f_{\mathbf{Z}}(\mathbf{z}'_{\sigma(i)})$, and $N_\sigma := N_k(\mathbf{z}'_{\sigma(1)},\dots,\mathbf{z}'_{\sigma(k)})$, for every $\sigma\in\mathfrak{I}_{k,n'}$ and $\varsigma\in\mathfrak{I}_{k,n}$, so that

$$T_4 := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}|\cdot|\mathfrak{I}_{k,n}|}\sum_{\sigma\in\mathfrak{I}_{k,n'}}\sum_{\varsigma\in\mathfrak{I}_{k,n}}\frac{K_{\sigma,\varsigma}}{f_{\mathbf{Z}',\sigma}}(g_\varsigma-\theta_\sigma)\overline{\boldsymbol{\psi}}_\sigma, \tag{8.18}$$

$$T_5 := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}|\cdot|\mathfrak{I}_{k,n}|}\sum_{\sigma\in\mathfrak{I}_{k,n'}}\sum_{\varsigma\in\mathfrak{I}_{k,n}}K_{\sigma,\varsigma}(g_\varsigma-\theta_\sigma)\Big(\frac{1}{N_\sigma}-\frac{1}{f_{\mathbf{Z}',\sigma}}\Big)\overline{\boldsymbol{\psi}}_\sigma. \tag{8.19}$$

Using α-order limited expansions, we get

$$\begin{aligned}\mathbb{E}[T_4] &= \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}|}\sum_{\sigma\in\mathfrak{I}_{k,n'}}\int\frac{\prod_{i=1}^k K_h\big(\mathbf{z}_i-\mathbf{z}'_{\sigma(i)}\big)}{f_{\mathbf{Z}',\sigma}}\Big(g\big(\mathbf{x}_{1:k}\big)-\theta_\sigma\Big)\prod_{i=1}^k f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i,\mathbf{z}_i)d\mu^{\otimes k}(\mathbf{x}_{1:k})d\mathbf{z}_{1:k} \qquad (8.20)\\ &= \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}|}\sum_{\sigma\in\mathfrak{I}_{k,n'}}\int\frac{\prod_{i=1}^k K\big(\mathbf{t}_i\big)}{f_{\mathbf{Z}',\sigma}}\Big(g\big(\mathbf{x}_{1:k}\big)-\theta_\sigma\Big)\prod_{i=1}^k f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i,\mathbf{z}'_{\sigma(i)}+h\mathbf{t}_i)d\mu^{\otimes k}(\mathbf{x}_{1:k})d\mathbf{t}_{1:k}\\ &= \frac{\tilde{r}_{n,n'}h^{k\alpha}}{|\mathfrak{I}_{k,n'}|}\sum_{\sigma\in\mathfrak{I}_{k,n'}}\int\frac{\prod_{i=1}^k K\big(\mathbf{t}_i\big)}{f_{\mathbf{Z}',\sigma}}\Big(g\big(\mathbf{x}_{1:k}\big)-\theta_\sigma\Big)\prod_{i=1}^k d_{\mathbf{Z}}^{(\alpha)}f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i,\mathbf{z}^*_{\sigma(i)})d\mu^{\otimes k}(\mathbf{x}_{1:k})d\mathbf{t}_{1:k}\\ &= O\Big(\tilde{r}_{n,n'}h^{k\alpha}\Big) = O\Big((n\times n'\times h_{n,n'}^{p+2k\alpha})^{1/2}\Big) = o(1),\end{aligned}$$

where above, $\mathbf{z}_i^*$ denote some vectors in $\mathbb{R}^p$ such that $|\mathbf{z}'_i - \mathbf{z}_i^*|_\infty \leq 1$, depending on $\mathbf{z}'_i$ and $\mathbf{x}_i$.

We can therefore use the centered version of T_4, defined as

$$T_4 - \mathbb{E}[T_4] = \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{\varsigma \in \mathfrak{I}_{k,n}} g_{\sigma,\varsigma},$$

$$g_{\sigma,\varsigma} := \frac{\overline{\psi}_\sigma}{f_{\mathbf{Z}',\sigma}} \Big(K_{\sigma,\varsigma}\big(g_\varsigma - \theta_\sigma\big) - \mathbb{E}\big[K_{\sigma,\varsigma}\big(g_\varsigma - \theta_\sigma\big)\big] \Big).$$

Computation of the limit of the variance matrix $Var[T_4]$.

We have $Var[T_4] = \mathbb{E}[T_4 T_4^T] + o(1)$.

$$Var[T_4] = \frac{\tilde{r}^2_{n,n'}}{|\mathfrak{I}_{k,n'}|^2 \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\sigma,\overline{\sigma} \in \mathfrak{I}_{k,n'}} \sum_{\varsigma,\overline{\varsigma} \in \mathfrak{I}_{k,n}} \mathbb{E}[g_{\sigma,\varsigma} g^T_{\overline{\sigma},\overline{\varsigma}}] + o(1).$$

By independence, $\mathbb{E}[g_{\sigma,\varsigma} g^T_{\overline{\sigma},\overline{\varsigma}}] = 0$ as soon as $\varsigma \cap \overline{\varsigma} = \emptyset$, where we identify a permutation ς and its image $\varsigma(\{1,\dots,k\})$. Therefore, we get

$$\begin{aligned} Var[T_4] &\simeq \frac{nn' h^p_{n,n'}}{|\mathfrak{I}_{k,n'}|^2 \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\sigma,\overline{\sigma} \in \mathfrak{I}_{k,n'}} \sum_{\substack{\varsigma,\overline{\varsigma} \in \mathfrak{I}_{k,n} \\ \varsigma \cap \overline{\varsigma} \neq \emptyset}} \mathbb{E}\big[g_{\sigma,\varsigma}\, g^T_{\overline{\sigma},\overline{\varsigma}}\big] \\ &= \frac{nn' h^p_{n,n'}}{|\mathfrak{I}_{k,n'}|^2 \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\sigma,\overline{\sigma} \in \mathfrak{I}_{k,n'}} \sum_{\substack{\varsigma,\overline{\varsigma} \in \mathfrak{I}_{k,n} \\ \varsigma \cap \overline{\varsigma} \neq \emptyset}} g_{\sigma,\varsigma,\overline{\sigma},\overline{\varsigma}} - \tilde{g}_\sigma \tilde{g}^T_{\overline{\sigma}}, \end{aligned}$$

where $\tilde{g}_\sigma := \overline{\psi}_\sigma \mathbb{E}\big[K_{\sigma,\varsigma}\big(g_\varsigma - \theta_\sigma\big)\big] / f_{\mathbf{Z}',\sigma}$ and

$$g_{\sigma,\varsigma,\overline{\sigma},\overline{\varsigma}} := \frac{\overline{\psi}_\sigma \overline{\psi}^T_{\overline{\sigma}}}{f_{\mathbf{Z}',\sigma} f_{\mathbf{Z}',\overline{\sigma}}} \mathbb{E}\Big[K_{\sigma,\varsigma} K_{\overline{\sigma},\overline{\varsigma}} \big(g_\varsigma - \theta_\sigma\big)\big(g_{\overline{\varsigma}} - \theta_{\overline{\sigma}}\big) \Big].$$

Assume now that $\varsigma \cap \overline{\varsigma}$ is of cardinality 1, i.e. there exists only one couple $(j,\overline{j}) \in \{1,\dots,k\}^2$ such that $\varsigma(j) = \overline{\varsigma}(\overline{j})$. Then,

$$\begin{aligned} g_{\sigma,\varsigma,\overline{\sigma},\overline{\varsigma}} = \frac{\overline{\psi}_\sigma \overline{\psi}^T_{\overline{\sigma}}}{f_{\mathbf{Z}',\sigma} f_{\mathbf{Z}',\overline{\sigma}}} \int & \big(g(\mathbf{X}_{1:k}) - \theta_\sigma\big)\big(g(\mathbf{x}_{k+1},\dots,\mathbf{x}_{k+\overline{j}-1},\mathbf{x}_j,\mathbf{x}_{k+\overline{j}+1},\dots,\mathbf{x}_{2k}) - \theta_{\overline{\sigma}}\big) \\ & \cdot \prod_{i=1}^k K_h\big(\mathbf{z}_i - \mathbf{z}'_{\sigma(i)}\big) f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i,\mathbf{z}_i) d\mu(\mathbf{x}_i) d\mathbf{z}_i \cdot K_h\big(\mathbf{z}_j - \mathbf{z}'_{\overline{\sigma}(\overline{j})}\big) \\ & \cdot \prod_{\overline{i}=1,\,\overline{i}\neq\overline{j}}^k K_h\big(\mathbf{z}_{k+i} - \mathbf{z}'_{\overline{\sigma}(\overline{i})}\big) f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_{k+i},\mathbf{z}_{k+i}) d\mu(\mathbf{x}_{k+i}) d\mathbf{z}_{k+i} \\ = \frac{\overline{\psi}_\sigma \overline{\psi}^T_{\overline{\sigma}}}{f_{\mathbf{Z}}(\mathbf{z}_j)} \int & \big(g(\mathbf{X}_{1:k}) - \theta_\sigma\big)\big(g(\mathbf{x}_{k+1},\dots,\mathbf{x}_{k+\overline{j}-1},\mathbf{x}_j,\mathbf{x}_{k+\overline{j}+1},\dots,\mathbf{x}_{2k}) - \theta_{\overline{\sigma}}\big) \\ & \cdot \prod_{i=1}^k K(\mathbf{t}_i) \frac{f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i,\mathbf{z}'_{\sigma(i)} + h\mathbf{t}_i)}{f_{\mathbf{Z}}(\mathbf{z}'_{\sigma(i)})} d\mu(\mathbf{x}_i) d\mathbf{t}_i \cdot h^{-p} K\left(\mathbf{t}_i + \frac{\mathbf{z}'_{\sigma(j)} - \mathbf{z}'_{\overline{\sigma}(\overline{j})}}{h}\right) \\ & \cdot \prod_{\overline{i}=1,\,\overline{i}\neq\overline{j}}^k K(\mathbf{t}_{k+i}) \frac{f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_{k+i},\mathbf{z}'_{\overline{\sigma}(\overline{i})} + h\mathbf{t}_{k+i})}{f_{\mathbf{Z}}(\mathbf{z}_{k+i})} d\mu(\mathbf{x}_{k+i}) d\mathbf{t}_{k+i} \\ \simeq \frac{\overline{\psi}_\sigma \overline{\psi}^T_{\overline{\sigma}}}{f_{\mathbf{Z}}(\mathbf{z}_j)} \int & \big(g(\mathbf{X}_{1:k}) - \theta_\sigma\big)\big(g(\mathbf{x}_{k+1},\dots,\mathbf{x}_{k+\overline{j}-1},\mathbf{x}_j,\mathbf{x}_{k+\overline{j}+1},\dots,\mathbf{x}_{2k}) - \theta_{\overline{\sigma}}\big) \\ & \cdot \prod_{i=1}^k K(\mathbf{t}_i) \frac{f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i,\mathbf{z}'_{\sigma(i)})}{f_{\mathbf{Z}}(\mathbf{z}_i)} d\mu(\mathbf{x}_i) d\mathbf{t}_i \cdot h^{-p} K\left(\mathbf{t}_i + \frac{\mathbf{z}'_{\sigma(j)} - \mathbf{z}'_{\overline{\sigma}(\overline{j})}}{h}\right) \\ & \cdot \prod_{\overline{i}=1,\,\overline{i}\neq\overline{j}}^k K(\mathbf{t}_{k+i}) \frac{f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_{k+i},\mathbf{z}'_{\overline{\sigma}(\overline{i})})}{f_{\mathbf{Z}}(\mathbf{z}'_{\overline{\sigma}(\overline{i})})} d\mu(\mathbf{x}_{k+i}) d\mathbf{t}_{k+i}. \end{aligned}$$

By assumption, this is zero unless $\sigma(j) = \overline{\sigma}(\overline{j})$. In this case, it can be simplified, giving

$$g_{\sigma,\varsigma,\overline{\sigma},\overline{\varsigma}} \simeq \frac{\overline{\boldsymbol{\psi}}_{\sigma}\overline{\boldsymbol{\psi}}_{\overline{\sigma}}^{T}}{f_{\mathbf{Z}}(\mathbf{z}_j)h^p} \int K^2 \int \big(g(\mathbf{x}_{1:k}) - \theta_{\sigma}\big)\big(g(\mathbf{x}_{k:2k,\overline{j}\to j}) - \theta_{\overline{\sigma}}\big) \cdot \prod_{i=1}^{k} f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{\sigma(i)}}(\mathbf{x}_k)d\mu(\mathbf{x}_i) \prod_{\overline{i}=1,\,\overline{i}\neq\overline{j}}^{k} f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{\overline{\sigma}(\overline{i})}}(\mathbf{x}_{k+i})d\mu(\mathbf{x}_{k+i}) \;=:\; h^{-p}g_{\sigma,\overline{\sigma},j,\overline{j}},$$

where $\mathbf{x}_{k:2k,\overline{j}\to j} := (\mathbf{x}_{k+1}, \ldots, \mathbf{x}_{k+\overline{j}-1}, \mathbf{x}_j, \mathbf{x}_{k+\overline{j}+1}, \ldots, \mathbf{x}_{2k})$.

Note that, if $\varsigma \cap \overline{\varsigma}$ is of cardinality strictly greater than 1, some supplementary powers of h^{-p} arise thanks to the repeated kernels in ς and $\overline{\varsigma}$. As a consequence, they are of lower order and therefore negligible. Using α-order expansions as in Equation (8.20), we get $\sup_{\sigma} |\tilde{g}_{\sigma}| = O(h^{k\alpha})$. Thus,

$$\begin{aligned} Var[T_4] &\simeq O\big(nn'h_{n,n'}^{p+2k\alpha}\big) + \frac{nn'h_{n,n'}^{p}}{|\mathfrak{I}_{k,n'}|^2 \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\varsigma\in\mathfrak{I}_{k,n}} \sum_{j,\overline{j}=1}^{k} \sum_{\substack{\overline{\varsigma}\in\mathfrak{I}_{k,n} \\ \varsigma(j)=\overline{\varsigma}(\overline{j}),\,|\varsigma\cap\overline{\varsigma}|=1}} \sum_{\sigma,\overline{\sigma}\in\mathfrak{I}_{k,n'},\,\sigma(j)=\overline{\sigma}(\overline{j})} h^{-p}g_{\sigma,\overline{\sigma},j,\overline{j}} \\ &\simeq \frac{n'}{|\mathfrak{I}_{k,n'}|^2} \sum_{j,\overline{j}=1}^{k} \sum_{\sigma,\overline{\sigma}\in\mathfrak{I}_{k,n'},\,\sigma(j)=\overline{\sigma}(\overline{j})} g_{\sigma,\overline{\sigma},j,\overline{j}} \\ &\to \sum_{j,\overline{j}=1}^{k} g_{j,\overline{j},\infty} = V_2, \end{aligned}$$

where

$$\begin{aligned} g_{j,\overline{j},\infty} := \int \Lambda'\Big(\theta\big(\mathbf{z}'_{1:k}\big)\Big)\Lambda'\Big(\theta\big(\mathbf{z}'_{k:2k,\overline{j}\to j}\big)\Big)\boldsymbol{\psi}\big(\mathbf{z}'_{1:k}\big)\boldsymbol{\psi}^{T}\big(\mathbf{z}'_{k:2k,\overline{j}\to j}\big)\frac{\int K^2}{f_{\mathbf{Z}}(\mathbf{z}'_j)} \int \big(g(\mathbf{x}_{1:k}) - \theta(\mathbf{z}'_{1:k})\big) \\ \cdot \big(g(\mathbf{x}_{k:2k,\overline{j}\to j}) - \theta(\mathbf{z}'_{k:2k,\overline{j}\to j})\big) \prod_{i=1,i\neq k+\overline{j}}^{2k} f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_i}(\mathbf{x}_i)f_{\mathbf{Z}',\infty}(\mathbf{z}'_i)d\mu(\mathbf{x}_i)d\mathbf{z}'_i. \end{aligned}$$

In Section 8.7.2, we will prove that T_4 is asymptotically Gaussian ; therefore, its asymptotic variance will be given by V_2.

Now, decompose the term T_5, defined in Equation (8.19), using a Taylor expansion of the function $x \mapsto 1/(1+x)$ at 0.

$$\frac{1}{N_{\sigma}} - \frac{1}{f_{\mathbf{Z}',\sigma}} = \frac{1}{f_{\mathbf{Z}',\sigma}}\left(\frac{1}{1 + \frac{N_{\sigma} - f_{\mathbf{Z}',\sigma}}{f_{\mathbf{Z}',\sigma}}} - 1\right) = -\frac{N_{\sigma} - f_{\mathbf{Z}',\sigma}}{f_{\mathbf{Z}',\sigma}^2} + T_{7,\sigma},$$

where

$$T_{7,\sigma} := \frac{1}{f_{\mathbf{Z}',\sigma}}(1+\alpha_{7,\sigma})^{-3}\left(\frac{N_{\sigma} - f_{\mathbf{Z}',\sigma}}{f_{\mathbf{Z}',\sigma}}\right)^2, \text{ with } |\alpha_{7,\sigma}| \leq \left|\frac{N_{\sigma} - f_{\mathbf{Z}',\sigma}}{f_{\mathbf{Z}',\sigma}}\right|.$$

We have therefore the decomposition $T_5 = -T_6 + T_7$, where

$$T_6 := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|} \sum_{\sigma\in\mathfrak{I}_{k,n'}} \sum_{\varsigma\in\mathfrak{I}_{k,n}} K_{\sigma,\varsigma}\big(g_{\varsigma} - \theta_{\sigma}\big)\frac{N_{\sigma} - f_{\mathbf{Z}',\sigma}}{f_{\mathbf{Z}',\sigma}^2}\overline{\boldsymbol{\psi}}_{\sigma}, \tag{8.21}$$

$$T_7 := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|} \sum_{\sigma\in\mathfrak{I}_{k,n'}} \sum_{\varsigma\in\mathfrak{I}_{k,n}} K_{\sigma,\varsigma}\big(g_{\varsigma} - \theta_{\sigma}\big)T_{7,\sigma}\overline{\boldsymbol{\psi}}_{\sigma}. \tag{8.22}$$

Summing up all the previous equation, we get

$$T_1 = (T_4 - \mathbb{E}[T_4]) - T_6 + T_7 + T_3 + o(1).$$

Afterwards, we will prove that all the remainders terms T_6, T_7 and T_3 are negligible, i.e. they tend to zero in probability. These results are respectively proved in Subsections 8.7.3, 8.7.4 and 8.7.5. Combining all these elements with the asymptotic normality of T_4 (proved in Subsection 8.7.2), we get $T_1 \xrightarrow{D} \mathcal{N}(0, V_2)$, as claimed. □

8.7.2 Proof of the asymptotic normality of T_4

Using the Hájek projection of T_4, we define

$$
\begin{aligned}
&T_4 - \mathbb{E}[T_4] = T_{4,1} + T_{4,2}, \text{ where} \\
&T_{4,1} := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{\varsigma \in \mathfrak{I}_{k,n}} \sum_{i=1}^{k} \mathbb{E}[g_{\sigma,\varsigma} | \varsigma(i)], \\
&T_{4,2} := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{\varsigma \in \mathfrak{I}_{k,n}} \left(g_{\sigma,\varsigma} - \sum_{i=1,\dots,k} \mathbb{E}[g_{\sigma,\varsigma} | \varsigma(i)] \right),
\end{aligned}
$$

denoting by $|i$ the conditioning with respect to $(\mathbf{X}_i, \mathbf{Z}_i)$, for $i \in \{1, \dots, n\}$. We will show that $T_{4,1}$ is asymptotically normal, and that $T_{4,2} = o(1)$.

Using the fact that the $(\mathbf{X}_i, \mathbf{Z}_i)_i$ are i.i.d., and denoting by Id the injective function $i \mapsto i$, we have

$$
\begin{aligned}
T_{4,1} &= \frac{k\tilde{r}_{n,n'}}{n|\mathfrak{I}_{k,n'}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{i=1}^{n} \mathbb{E}\left[\frac{\overline{\psi}_\sigma}{f_{\mathbf{Z}',\sigma}} K_{\sigma,Id}\big(g_{Id} - \theta_\sigma\big) - \overline{g}_\sigma \middle| i \right] \\
&\simeq \frac{k\tilde{r}_{n,n'}}{n|\mathfrak{I}_{k,n'}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{i=1}^{n} \mathbb{E}\left[\frac{\overline{\psi}_\sigma}{f_{\mathbf{Z}',\sigma}} K_{\sigma,Id}\big(g_{Id} - \theta_\sigma\big) \middle| i \right] =: \sum_{i=1}^{n} \alpha_{4,i,n},
\end{aligned}
$$

because $\sup_\sigma |\overline{g}_\sigma| = O(h^{k\alpha})$, as proved in the previous section, hence negligible. The $\alpha_{4,i,n}$, for $1 \leq i \leq n$, form a triangular array of i.i.d. variables. To prove the asymptotic normality of $T_{4,1}$, it remains to check Lyapunov's condition, i.e. we will show that $\sum_{i=1}^{n} \mathbb{E}\big[|\alpha_{4,i,n}|_\infty^3\big] \to 0$. We have

$$
\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}\big[|\alpha_{4,i,n}|_\infty^3\big] = n\,\mathbb{E}\big[|\alpha_{4,1,n}|_\infty^3\big] \\
&= \frac{k^3 n \tilde{r}_{n,n'}^3}{n^3 |\mathfrak{I}_{k,n'}|^3} \sum_{\sigma,\nu,\vartheta \in \mathfrak{I}_{k,n'}} \frac{\overline{\psi}_\sigma \otimes \overline{\psi}_\nu \otimes \overline{\psi}_\vartheta}{f_{\mathbf{Z}',\sigma} f_{\mathbf{Z}',\nu} f_{\mathbf{Z}',\vartheta}} \mathbb{E}\left[\mathbb{E}\Big[K_{\sigma,Id}\big(g_{Id} - \theta_\sigma\big)\Big|1\Big] \mathbb{E}\Big[K_{\nu,Id}\big(g_{Id} - \theta_\nu)\big)\Big|1\Big] \mathbb{E}\Big[K_{\vartheta,Id}\big(g_{Id} - \theta_\vartheta)\big)\Big|1\Big] \right] \\
&= \frac{k^3 \tilde{r}_{n,n'}^3}{n^2 |\mathfrak{I}_{k,n'}|^3} \sum_{\sigma,\nu,\vartheta \in \mathfrak{I}_{k,n'}} \frac{\overline{\psi}_\sigma \otimes \overline{\psi}_\nu \otimes \overline{\psi}_\vartheta}{f_{\mathbf{Z}}(\mathbf{z}'_{\nu(1)}) f_{\mathbf{Z}}(\mathbf{z}'_{\vartheta(1)})} \int K_h\big(\mathbf{z}_1 - \mathbf{z}'_{\sigma(1)}\big) K_h\big(\mathbf{z}_1 - \mathbf{z}'_{\nu(1)}\big) K_h\big(\mathbf{z}_1 - \mathbf{z}'_{\vartheta(1)}\big) \\
&\cdot \prod_{i=2}^{k} K_h\big(\mathbf{z}_i - \mathbf{z}'_{\sigma(i)}\big) K_h\big(\mathbf{z}_{k+i} - \mathbf{z}'_{\nu(i)}\big) K_h\big(\mathbf{z}_{2k+i} - \mathbf{z}'_{\vartheta(i)}\big) \\
&\cdot \big(g(\mathbf{x}_{1:k}) - \theta_\sigma\big)\big(g(\mathbf{x}_1, \mathbf{x}_{(k+2):(2k)}) - \theta_\nu\big)\big(g(\mathbf{x}_1, \mathbf{x}_{(2k+2):(3k)}) - \theta_\vartheta\big) \\
&\cdot \prod_{i=1}^{k} \frac{f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i, \mathbf{z}_i)}{f_{\mathbf{Z}}(\mathbf{z}'_{\sigma(i)})} d\mu(\mathbf{x}_i) d\mathbf{z}_i \prod_{i=2}^{k} \frac{f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_{k+i}, \mathbf{z}_{k+i})}{f_{\mathbf{Z}}(\mathbf{z}'_{\nu(i)})} d\mu(\mathbf{x}_{k+i}) d\mathbf{z}_{k+i} \prod_{i=2}^{k} \frac{f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_{2k+i}, \mathbf{z}_{2k+i})}{f_{\mathbf{Z}}(\mathbf{z}'_{\vartheta(i)})} d\mu(\mathbf{x}_{2k+i}) d\mathbf{z}_{2k+i} \\
&\simeq \frac{k^3 \tilde{r}_{n,n'}^3}{n^2 |\mathfrak{I}_{k,n'}|^3} \sum_{\sigma,\nu,\vartheta \in \mathfrak{I}_{k,n'}} \frac{\overline{\psi}_\sigma \otimes \overline{\psi}_\nu \otimes \overline{\psi}_\vartheta}{f_{\mathbf{Z}}(\mathbf{z}'_{\nu(1)}) f_{\mathbf{Z}}(\mathbf{z}'_{\vartheta(1)})} \int h^{-2p} K(\mathbf{t}_1) K\left(\mathbf{t}_1 + \frac{\mathbf{z}'_{\sigma(1)} - \mathbf{z}'_{\nu(1)}}{h}\right) K\left(\mathbf{t}_1 + \frac{\mathbf{z}'_{\sigma(1)} - \mathbf{z}'_{\vartheta(1)}}{h}\right) \\
&\cdot \prod_{i=2}^{k} K_h\big(\mathbf{t}_i\big) K_h\big(\mathbf{t}_{k+i}\big) K_h\big(\mathbf{t}_{2k+i}\big) \big(g(\mathbf{x}_{1:k}) - \theta_\sigma\big)\big(g(\mathbf{x}_1, \mathbf{x}_{(k+2):(2k)}) - \theta_\nu\big)\big(g(\mathbf{x}_1, \mathbf{x}_{(2k+2):(3k)}) - \theta_\vartheta\big) \\
&\cdot \prod_{i=1}^{k} f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{\sigma(i)}}(\mathbf{x}_i) d\mu(\mathbf{x}_i) d\mathbf{z}_i \prod_{i=2}^{k} f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{\nu(i)}}(\mathbf{x}_{k+i}) d\mu(\mathbf{x}_{k+i}) d\mathbf{z}_{k+i} \prod_{i=2}^{k} f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{\vartheta(i)}}(\mathbf{x}_{2k+i}, \mathbf{t}_{2k+i}) d\mu(\mathbf{x}_{2k+i}) d\mathbf{t}_{2k+i},
\end{aligned}
$$

where in the last equivalent, we use a change of variable from the $\mathbf{z}_i$ to the $\mathbf{t}_i$, and then the continuity of the density $f_{\mathbf{X},\mathbf{Z}}$ with respect to $\mathbf{z}$, because $h = o(1)$.

Because of our assumptions, the terms of the sum for which $\sigma(1) \neq 1$ or $\nu(1) \neq 1$ are zero. Therefore, we get

$$\sum_{i=1}^{n} \mathbb{E}\big[|\alpha_{4,i,n}|_\infty^3\big] = \frac{\tilde{r}_{n,n'}^3 h^{-2p}}{n^2 |\mathfrak{I}_{k,n'}|^3} \sum_{\sigma,\nu,\vartheta \in \mathfrak{I}_{k,n'}, \sigma(1)=\nu(1)=1} O(1) = O\left(\frac{(nn'h^p)^{3/2}}{n^2 n'^2 h^{2p}}\right) = O\left(\frac{1}{(nn'h^p)^{1/2}}\right) = o(1).$$

We prove now that $T_{4,2} = o(1)$. Note first that, by construction, $\mathbb{E}[T_{4,2}] = 0$. Computing its variance, we get

$$\begin{aligned}
\mathbb{E}\big[T_{4,2}T_{4,2}^T\big] &= \mathbb{E}\left[\frac{\tilde{r}_{n,n'}^2}{|\mathfrak{I}_{k,n'}|^2 \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\sigma,\overline{\sigma} \in \mathfrak{I}_{k,n'}} \sum_{\varsigma,\overline{\varsigma} \in \mathfrak{I}_{k,n}} \left(g_{\sigma,\varsigma} - \sum_{i=1,\dots,k} \mathbb{E}\big[g_{\sigma,\varsigma}\big|\varsigma(i)\big]\right)\left(g_{\overline{\sigma},\overline{\varsigma}} - \sum_{\bar{i}=1,\dots,k} \mathbb{E}\big[g_{\overline{\sigma},\overline{\varsigma}}\big|\overline{\varsigma}(\bar{i})\big]\right)^T\right] \\
&=: \frac{\tilde{r}_{n,n'}^2}{|\mathfrak{I}_{k,n'}|^2 \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\sigma,\overline{\sigma} \in \mathfrak{I}_{k,n'}} \sum_{\varsigma,\overline{\varsigma} \in \mathfrak{I}_{k,n}} \mathbb{E}\Big[\tilde{g}_{\sigma,\overline{\sigma},\varsigma,\overline{\varsigma}}\Big]. \qquad (8.23)
\end{aligned}$$

Because of $\mathbb{E}[g_{\sigma,\varsigma}] = 0$ and by independence, the terms in the latter sum for which $\varsigma \cap \overline{\varsigma} = \emptyset$ are zero. Otherwise, there exists $j_1, j_2 \in \{1,\dots,k\}$ such that $\varsigma(j_1) = \overline{\varsigma}(j_2)$. If $\varsigma \cap \overline{\varsigma}$ is of cardinal 1, meaning that there is no other identities between elements of ς and $\overline{\varsigma}$, then we will show that the corresponding term is zero as well. We place ourselves in this case, assuming that $|\varsigma \cap \overline{\varsigma}| = 1$, and we get

$$\begin{aligned}
\mathbb{E}\Big[\tilde{g}_{\sigma,\overline{\sigma},\varsigma,\overline{\varsigma}}\Big] &= \mathbb{E}\left[\left(g_{\sigma,\varsigma} - \sum_{i=1,\dots,k} \mathbb{E}\big[g_{\sigma,\varsigma}\big|\varsigma(i)\big]\right)\left(g_{\overline{\sigma},\overline{\varsigma}}^T - \sum_{\bar{i}=1,\dots,k} \mathbb{E}\big[g_{\overline{\sigma},\overline{\varsigma}}^T\big|\overline{\varsigma}(\bar{i})\big]\right)\right] \\
&= \mathbb{E}\left[\left(g_{\sigma,\varsigma} - \mathbb{E}\big[g_{\sigma,\varsigma}\big|\varsigma(j_1)\big]\right)\left(g_{\overline{\sigma},\overline{\varsigma}}^T - \mathbb{E}\big[g_{\overline{\sigma},\overline{\varsigma}}^T\big|\overline{\varsigma}(j_2)\big]\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(g_{\sigma,\varsigma} - \mathbb{E}\big[g_{\sigma,\varsigma}\big|\varsigma(j_1)\big]\right)\left(g_{\overline{\sigma},\overline{\varsigma}}^T - \mathbb{E}\big[g_{\overline{\sigma},\overline{\varsigma}}^T\big|\varsigma(j_1)\big]\right)\Big|\varsigma(j_1)\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\Big[g_{\sigma,\varsigma} g_{\overline{\sigma},\overline{\varsigma}}^T\Big|\varsigma(j_1)\Big]\right] - \mathbb{E}\left[\mathbb{E}\big[g_{\sigma,\varsigma}\big|\varsigma(j_1)\big]\mathbb{E}\big[g_{\overline{\sigma},\overline{\varsigma}}^T\big|\varsigma(j_1)\big]\right] = 0.
\end{aligned}$$

Therefore, non-zero terms in Equation (8.23) correspond to the case where there exists $j_3 \neq j_1, j_4 \neq j_1$ such that $\varsigma(j_3) = \overline{\varsigma}(j_4)$. It is equivalent to $|\varsigma \cap \overline{\varsigma}| \geq 2$. We will ignore higher-order terms, i.e. the ones for which $|\varsigma \cap \overline{\varsigma}| > 2$, as they yield higher powers of h^p and are therefore negligible. Finally, Equation (8.23) becomes

$$\mathbb{E}\big[T_{4,2}T_{4,2}^T\big] \simeq \frac{\tilde{r}_{n,n'}^2}{|\mathfrak{I}_{k,n'}|^2 \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\sigma,\overline{\sigma} \in \mathfrak{I}_{k,n'}} \sum_{\substack{\varsigma,\overline{\varsigma} \in \mathfrak{I}_{k,n} \\ |\varsigma \cap \overline{\varsigma}| = 2}} \left(\mathbb{E}\Big[g_{\sigma,\varsigma} g_{\overline{\sigma},\overline{\varsigma}}^T\Big] - 2k\mathbb{E}\left[\mathbb{E}\big[g_{\sigma,\varsigma}\big|\varsigma(i)\big]\mathbb{E}\big[g_{\overline{\sigma},\overline{\varsigma}}^T\big|\overline{\varsigma}(\bar{i})\big]\right]\right).$$

As before, using change of variables and limited expansions, we can prove that

$$\frac{\tilde{r}_{n,n'}^2}{|\mathfrak{I}_{k,n'}|^2 \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\sigma,\overline{\sigma} \in \mathfrak{I}_{k,n'}} \sum_{\substack{\varsigma,\overline{\varsigma} \in \mathfrak{I}_{k,n} \\ |\varsigma \cap \overline{\varsigma}| = 2}} \mathbb{E}\Big[g_{\sigma,\varsigma} g_{\overline{\sigma},\overline{\varsigma}}^T\Big] = o(1),$$

and similarly for the other term.

8.7.3 Convergence of T_6 to 0

Using Equation (8.21), we have $T_6 = T_{6,1} + T_{6,2}$, where

$$T_{6,1} := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{\varsigma \in \mathfrak{I}_{k,n}} K_{\sigma,\varsigma}\big(g_\varsigma - \theta_\sigma\big) \frac{N_\sigma - \mathbb{E}[N_\sigma]}{f^2_{\mathbf{Z}',\sigma}} \overline{\boldsymbol{\psi}}_\sigma, \tag{8.24}$$

$$T_{6,2} := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{\varsigma \in \mathfrak{I}_{k,n}} K_{\sigma,\varsigma}\big(g_\varsigma - \theta_\sigma\big) \frac{\mathbb{E}[N_\sigma] - f_{\mathbf{Z}',\sigma}}{f^2_{\mathbf{Z}',\sigma}} \overline{\boldsymbol{\psi}}_\sigma. \tag{8.25}$$

We first prove that $T_{6,1} = o(1)$. Using Equation (8.5), we have

$$\begin{aligned}
T_{6,1} &= \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{\varsigma \in \mathfrak{I}_{k,n}} \frac{1}{f^2_{\mathbf{Z}',\sigma}} K_{\sigma,\varsigma}\big(g_\varsigma - \theta_\sigma\big)\big(N_k(\mathbf{z}'_{\sigma(1:k)}) - \mathbb{E}[N_k(\mathbf{z}'_{\sigma(1:k)})]\big)\overline{\boldsymbol{\psi}}_\sigma \\
&= \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{\varsigma \in \mathfrak{I}_{k,n}} \frac{1}{f^2_{\mathbf{Z}',\sigma}} K_{\sigma,\varsigma}\big(g_\varsigma - \theta_\sigma\big) \sum_{\nu \in \mathfrak{I}_{k,n}} \Big(\prod_{i=1}^k K_h\big(\mathbf{Z}_{\nu(i)} - \mathbf{z}'_{\sigma(i)}\big) - \mathbb{E}\Big[\prod_{i=1}^k K_h\big(\mathbf{Z}_{\nu(i)} - \mathbf{z}'_{\sigma(i)}\big)\Big]\Big)\overline{\boldsymbol{\psi}}_\sigma \\
&= \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{\varsigma,\nu \in \mathfrak{I}_{k,n}} \frac{1}{f^2_{\mathbf{Z}',\sigma}} K_{\sigma,\varsigma}\big(g_\varsigma - \theta_\sigma\big)\Big(K_{\sigma,\nu} - \mathbb{E}\big[K_{\sigma,\nu}\big]\Big)\overline{\boldsymbol{\psi}}_\sigma.
\end{aligned}$$

The terms for which $|\varsigma \cap \nu| \geq 1$ induce some powers of $(nh^p)^{-1}$, and are therefore negligible. We remove them to obtain an equivalent random vector $\overline{T}_{6,1}$, which is centered. Therefore it is sufficient to show that its second moment tends to 0.

$$\mathbb{E}\big[\overline{T}_{6,1}\overline{T}_{6,1}^T\big] = \frac{\tilde{r}^2_{n,n'}}{|\mathfrak{I}_{k,n'}|^2 \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\sigma,\overline{\sigma} \in \mathfrak{I}_{k,n'}} \sum_{\substack{\varsigma,\nu \in \mathfrak{I}_{k,n} \\ \varsigma \cap \nu = \emptyset}} \sum_{\substack{\overline{\varsigma},\overline{\nu} \in \mathfrak{I}_{k,n} \\ \overline{\varsigma} \cap \overline{\nu} = \emptyset}} \frac{\overline{\boldsymbol{\psi}}_\sigma}{f^2_{\mathbf{Z}',\sigma}} \frac{\overline{\boldsymbol{\psi}}_{\overline{\sigma}}^T}{f^2_{\mathbf{Z}',\overline{\sigma}}} g_{\sigma,\overline{\sigma},\varsigma,\overline{\varsigma},\nu,\overline{\nu}},$$

$$g_{\sigma,\overline{\sigma},\varsigma,\overline{\varsigma},\nu,\overline{\nu}} := \mathbb{E}\Big[K_{\sigma,\varsigma}\big(g_\varsigma - \theta_\sigma\big)\Big(K_{\sigma,\nu} - \mathbb{E}\big[K_{\sigma,\nu}\big]\Big)K_{\overline{\sigma},\overline{\varsigma}}\big(g_{\overline{\varsigma}} - \theta_{\overline{\sigma}}\big)\Big(K_{\overline{\sigma},\overline{\nu}} - \mathbb{E}\big[K_{\overline{\sigma},\overline{\nu}}\big]\Big)\Big].$$

The term $g_{\sigma,\overline{\sigma},\varsigma,\overline{\varsigma},\nu,\overline{\nu}}$ is 0 in two cases : if $\nu \cap (\varsigma \cup \overline{\varsigma} \cup \overline{\nu})$ or if $\overline{\nu} \cap (\varsigma \cup \overline{\varsigma} \cup \nu)$. This condition can be written as

$$\emptyset = \big[\nu \cap (\overline{\varsigma} \cup \overline{\nu})\big] \cup \big[\overline{\nu} \cap (\varsigma \cup \nu)\big] = (\nu \cup \overline{\nu}) \cap (\overline{\varsigma} \cup \overline{\nu}) \cap (\varsigma \cup \nu).$$

We deduce that non-zero terms arise only when there exists $j_1, j_2 \in \{1, \dots, k\}$ such that: $\nu(j_1) = \overline{\nu}(j_2)$ or $\nu(j_1) = \overline{\varsigma}(j_2)$ or $\overline{\nu}(j_1) = \varsigma(j_2)$. Therefore, we can write $\mathbb{E}\big[\overline{T}_{6,1}\overline{T}_{6,1}^T\big] = T_{6,1,1} + T_{6,1,2} + T_{6,1,3}$, where

$$T_{6,1,1} = \frac{\tilde{r}^2_{n,n'}}{|\mathfrak{I}_{k,n'}|^2 \cdot |\mathfrak{I}_{k,n}|^2} \sum_{j_1,j_2=1}^k \sum_{\sigma,\overline{\sigma} \in \mathfrak{I}_{k,n'}} \sum_{\substack{\varsigma,\nu \in \mathfrak{I}_{k,n} \\ \varsigma \cap \nu = \emptyset}} \sum_{\substack{\overline{\varsigma},\overline{\nu} \in \mathfrak{I}_{k,n} \\ \overline{\varsigma} \cap \overline{\nu} = \emptyset, \overline{\nu}(j_2) = \nu(j_1)}} \frac{\overline{\boldsymbol{\psi}}_\sigma}{f^2_{\mathbf{Z}',\sigma}} \frac{\overline{\boldsymbol{\psi}}_{\overline{\sigma}}^T}{f^2_{\mathbf{Z}',\overline{\sigma}}} g_{\sigma,\overline{\sigma},\varsigma,\overline{\varsigma},\nu,\overline{\nu}},$$

$$T_{6,1,2} = \frac{\tilde{r}^2_{n,n'}}{|\mathfrak{I}_{k,n'}|^2 \cdot |\mathfrak{I}_{k,n}|^2} \sum_{j_1,j_2=1}^k \sum_{\sigma,\overline{\sigma} \in \mathfrak{I}_{k,n'}} \sum_{\substack{\varsigma,\nu \in \mathfrak{I}_{k,n} \\ \varsigma \cap \nu = \emptyset}} \sum_{\substack{\overline{\varsigma},\overline{\nu} \in \mathfrak{I}_{k,n} \\ \overline{\varsigma} \cap \overline{\nu} = \emptyset, \overline{\varsigma}(j_2) = \nu(j_1)}} \frac{\overline{\boldsymbol{\psi}}_\sigma}{f^2_{\mathbf{Z}',\sigma}} \frac{\overline{\boldsymbol{\psi}}_{\overline{\sigma}}^T}{f^2_{\mathbf{Z}',\overline{\sigma}}} g_{\sigma,\overline{\sigma},\varsigma,\overline{\varsigma},\nu,\overline{\nu}},$$

$$T_{6,1,3} = \frac{\tilde{r}^2_{n,n'}}{|\mathfrak{I}_{k,n'}|^2 \cdot |\mathfrak{I}_{k,n}|^2} \sum_{j_1,j_2=1}^k \sum_{\sigma,\overline{\sigma} \in \mathfrak{I}_{k,n'}} \sum_{\substack{\varsigma,\nu \in \mathfrak{I}_{k,n} \\ \varsigma \cap \nu = \emptyset}} \sum_{\substack{\overline{\varsigma},\overline{\nu} \in \mathfrak{I}_{k,n} \\ \overline{\varsigma} \cap \overline{\nu} = \emptyset, \overline{\nu}(j_1) = \varsigma(j_2)}} \frac{\overline{\boldsymbol{\psi}}_\sigma}{f^2_{\mathbf{Z}',\sigma}} \frac{\overline{\boldsymbol{\psi}}_{\overline{\sigma}}^T}{f^2_{\mathbf{Z}',\overline{\sigma}}} g_{\sigma,\overline{\sigma},\varsigma,\overline{\varsigma},\nu,\overline{\nu}},$$

We will prove that $T_{6,1,1} = o(1)$. The two other terms can be treated in a similar way. Because of our assumptions, the terms for which $\overline{\sigma}(j_1) \neq \sigma(j_2)$ are zero. This divides the number of possible terms by n'. By using limited expansions as in Equation (8.20), we get that $g_{\sigma,\overline{\sigma},\varsigma,\overline{\varsigma},\nu,\overline{\nu}} = O(h^{k\alpha - p})$. Therefore, we have $T_{6,1,1} = O\big(\frac{nn'h^p}{nn'} h^{k\alpha - p}\big) = O(h^{k\alpha}) = o(1)$.

Concerning $T_{6,2}$, its variance matrix is given by

$$Var\big[T_{6,2}\big] = \frac{\tilde{r}_{n,n'}^2}{|\mathfrak{I}_{k,n'}|^2 \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\sigma,\overline{\sigma}\in\mathfrak{I}_{k,n'}} \sum_{\varsigma,\overline{\varsigma}\in\mathfrak{I}_{k,n}} \frac{\mathbb{E}[N_\sigma] - f_{\mathbf{Z}',\sigma}}{f_{\mathbf{Z}',\sigma}^2} \frac{\mathbb{E}[N_{\overline{\sigma}}] - f_{\mathbf{Z}',\overline{\sigma}}}{f_{\mathbf{Z}',\overline{\sigma}}^2} \overline{\boldsymbol{\psi}}_\sigma \overline{\boldsymbol{\psi}}_{\overline{\sigma}} \overline{g}_{\sigma,\overline{\sigma},\varsigma,\overline{\varsigma}},$$

$$\overline{g}_{\sigma,\overline{\sigma},\varsigma,\overline{\varsigma}} := \mathbb{E}\Big[K_{\sigma,\varsigma} K_{\overline{\sigma},\overline{\varsigma}} \big(g_\varsigma - \theta_\sigma\big)\big(g_{\overline{\varsigma}} - \theta_{\overline{\sigma}}\big)\Big] - \mathbb{E}\Big[K_{\sigma,\varsigma}\big(g_\varsigma - \theta_\sigma\big)\Big]\mathbb{E}\Big[K_{\overline{\sigma},\overline{\varsigma}}\big(g_{\overline{\varsigma}} - \theta_{\overline{\sigma}}\big)\Big].$$

Note that $\overline{g}_{\sigma,\overline{\sigma},\varsigma,\overline{\varsigma}} = 0$ when $\varsigma \cap \overline{\varsigma} = \emptyset$. This divides the number of terms in the sum above by n, and imposes that $\sigma \cap \overline{\varsigma} \neq 0$, which divides the number of terms in the sum above by another n'. Finally, limited expansions gives a bound of $h^{k\alpha-p}$. Summing up all these elements, we obtain $Var\big[T_{6,2}\big] = O(\frac{\tilde{r}_{n,n'}^2}{nn'} h^{k\alpha-p}) = O(h^{k\alpha}) = o(1)$. Similarly, we get $\mathbb{E}\big[T_{6,2}\big] = o(1)$ by a Taylor expansion.

8.7.4 Convergence of T_7 to 0

We recall Equation (8.22):

$$T_7 = \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|} \sum_{\sigma\in\mathfrak{I}_{k,n'}} \sum_{\varsigma\in\mathfrak{I}_{k,n}} K_{\sigma,\varsigma}\big(g_\varsigma - \theta_\sigma\big) T_{7,\sigma} \overline{\boldsymbol{\psi}}_\sigma,$$

$$T_{7,\sigma} := \frac{1}{f_{\mathbf{Z}',\sigma}}(1+\alpha_{7,\sigma})^{-3} \left(\frac{N_\sigma - f_{\mathbf{Z}',\sigma}}{f_{\mathbf{Z}',\sigma}}\right)^2, \text{ with } |\alpha_{7,\sigma}| \leq \left|\frac{N_\sigma - f_{\mathbf{Z}',\sigma}}{f_{\mathbf{Z}',\sigma}}\right|.$$

By Lemma 8.3 applied to $\mathbf{z}_1 = \mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}_{n'} = \mathbf{z}'_{\sigma(n')}$, for $\sigma \in \mathfrak{I}_{k,n'}$, we get

$$\mathbb{P}\bigg(\sup_{\sigma\in\mathfrak{I}_{k,n'}} \big|N_\sigma - f_{\mathbf{Z}',\sigma}\big| \leq \frac{C_{K,\alpha}}{\alpha} h^\alpha + t\bigg) \geq 1 - 2\exp\bigg(-\frac{[n/k]t^2}{h^{-kp}C_1 + h^{-kp}C_2 t}\bigg),$$

for any $t > 0$. Therefore, $\sup_{\sigma\in\mathfrak{I}_{k,n'}} |T_{7,\sigma}| = O_{\mathbb{P}}(h^{2\alpha})$ by choosing $t = h^{\alpha/k}$. Then,

$$|T_7| \leq \sup_{\sigma\in\mathfrak{I}_{k,n'}} |T_{7,\sigma}| \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|} \sum_{\sigma\in\mathfrak{I}_{k,n'}} \sum_{\varsigma\in\mathfrak{I}_{k,n}} |K_{\sigma,\varsigma}| \cdot \big|g_\varsigma - \theta_\sigma\big| \cdot |\overline{\boldsymbol{\psi}}_\sigma|.$$

The expectation of the double sum is $O(h^\alpha)$, by α-order limited expansions. By Markov's inequality, we deduce

$$T_7 = O_{\mathbb{P}}\Big(\tilde{r}_{n,n'} \sup_{\sigma\in\mathfrak{I}_{k,n'}} |T_{7,\sigma}| h^\alpha)\Big) = O_{\mathbb{P}}(\tilde{r}_{n,n'} h^{3\alpha}) = O_{\mathbb{P}}\Big((nn' h^{p+3\alpha})^{1/2}\Big),$$

therefore $T_7 = o_{\mathbb{P}}(1)$.

8.7.5 Convergence of T_3 to 0

We have

$$T_3 := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}|} \sum_{\sigma\in\mathfrak{I}_{k,n'}} \alpha_{3,\sigma} \cdot \Big(\hat{\theta}\big(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\big) - \theta\big(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\big)\Big)^2 \boldsymbol{\psi}\big(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\big),$$

with $\forall \sigma \in \mathfrak{I}_{k,n'}$, $|\alpha_{3,\sigma}| \leq C_{\Lambda''}/2$. Therefore

$$\begin{aligned} T_3 &\lesssim \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}|} \sum_{\sigma\in\mathfrak{I}_{k,n'}} \Big(\hat{\theta}\big(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\big) - \theta\big(\mathbf{z}'_{\sigma(1)}, \ldots, \mathbf{z}'_{\sigma(k)}\big)\Big)^2 \\ &\lesssim \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}|} \bigg(\frac{1}{|\mathfrak{I}_{k,n}|} \sum_{\varsigma\in\mathfrak{I}_{k,n}} \frac{K_{\sigma,\varsigma}}{f_{\mathbf{Z}',\sigma}}\big(g_\varsigma - \theta_\sigma\big) + K_{\sigma,\varsigma}\big(g_\varsigma - \theta_\sigma\big)\Big(\frac{1}{N_\sigma} - \frac{1}{f_{\mathbf{Z}',\sigma}}\Big)\bigg)^2 = T_8 + T_9 + T_{10}, \end{aligned}$$

where

$$T_8 := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{\varsigma,\overline{\varsigma} \in \mathfrak{I}_{k,n}} \frac{K_{\sigma,\varsigma} K_{\sigma,\overline{\varsigma}}}{f^2_{\mathbf{Z}',\sigma}} \big(g_\varsigma - \theta_\sigma\big)\big(g_{\overline{\varsigma}} - \theta_\sigma\big),$$

$$T_9 := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{\varsigma,\overline{\varsigma} \in \mathfrak{I}_{k,n}} \frac{K_{\sigma,\varsigma} K_{\sigma,\overline{\varsigma}}}{f_{\mathbf{Z}',\sigma}} \big(g_\varsigma - \theta_\sigma\big)\big(g_{\overline{\varsigma}} - \theta_\sigma\big)\Big(\frac{1}{N_\sigma} - \frac{1}{f_{\mathbf{Z}',\sigma}}\Big),$$

$$T_{10} := \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{\varsigma,\overline{\varsigma} \in \mathfrak{I}_{k,n}} K_{\sigma,\varsigma} K_{\sigma,\overline{\varsigma}} \big(g_\varsigma - \theta_\sigma\big)\big(g_{\overline{\varsigma}} - \theta_\sigma\big)\Big(\frac{1}{N_\sigma} - \frac{1}{f_{\mathbf{Z}',\sigma}}\Big)^2.$$

We show that $T_8 = o(1)$. The two other terms can be treated in a similar way.

$$\begin{aligned}
\mathbb{E}\big[|T_8|\big] &= \mathbb{E}\Bigg[\frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{\varsigma,\overline{\varsigma} \in \mathfrak{I}_{k,n}} \frac{|K_{\sigma,\varsigma} K_{\sigma,\overline{\varsigma}}|}{f^2_{\mathbf{Z}',\sigma}} \big|g_\varsigma - \theta_\sigma\big| \cdot \big|g_{\overline{\varsigma}} - \theta_\sigma\big|\Bigg] \\
&= \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{\varsigma,\overline{\varsigma} \in \mathfrak{I}_{k,n}} \int \frac{\prod_{i=1}^k \Big|K_h\big(\mathbf{z}_{\varsigma(i)} - \mathbf{z}'_{\sigma(i)}\big) K_h\big(\mathbf{z}_{\overline{\varsigma}(i)} - \mathbf{z}'_{\sigma(i)}\big)\Big|}{f^2_{\mathbf{Z}',\sigma}} \\
&\qquad \cdot \Big|g\big(\mathbf{x}_{\varsigma(1:k)}\big) - \theta_\sigma\Big|\Big|g\big(\mathbf{x}_{\overline{\varsigma}(1:k)}\big) - \theta_\sigma\Big| \prod_{i \,\in\, \varsigma(1:k) \cup \overline{\varsigma}(1:k)} f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i, \mathbf{z}_i) d\mu(\mathbf{x}_i) d\mathbf{z}_i.
\end{aligned}$$

Note that terms for which $\varsigma \neq \overline{\varsigma} \in \mathfrak{I}_{k,n'}$ are zero, because the $\mathbf{z}'_i$ are distinct and because of our Assumption 8.3.2(i). Therefore, we get

$$\begin{aligned}
\mathbb{E}\big[|T_8|\big] &= \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|^2} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \sum_{\varsigma \in \mathfrak{I}_{k,n}} \int \frac{\prod_{i=1}^k K_h\big(\mathbf{z}_{\varsigma(i)} - \mathbf{z}'_{\sigma(i)}\big)^2}{f^2_{\mathbf{Z}',\sigma}} \Big(g\big(\mathbf{x}_{\varsigma(1:k)}\big) - \theta_\sigma\Big)^2 \prod_{i \,\in\, \varsigma(1:k)} f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i, \mathbf{z}_i) d\mu(\mathbf{x}_i) d\mathbf{z}_i \\
&= \frac{\tilde{r}_{n,n'}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \int \frac{\prod_{i=1}^k K_h\big(\mathbf{z}_i - \mathbf{z}'_{\sigma(i)}\big)^2}{f^2_{\mathbf{Z}',\sigma}} \Big(g\big(\mathbf{x}_{\varsigma(1:k)}\big) - \theta_\sigma\Big)^2 \prod_{i=1}^k f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i, \mathbf{z}_i) d\mu(\mathbf{x}_i) d\mathbf{z}_i \\
&= \frac{\tilde{r}_{n,n'} h^{-kp}}{|\mathfrak{I}_{k,n'}| \cdot |\mathfrak{I}_{k,n}|} \sum_{\sigma \in \mathfrak{I}_{k,n'}} \int \frac{\prod_{i=1}^k K\big(\mathbf{t}_i\big)^2}{f^2_{\mathbf{Z}',\sigma}} \Big(g\big(\mathbf{x}_{\varsigma(1:k)}\big) - \theta_\sigma\Big)^2 \prod_{i=1}^k f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i, \mathbf{z}'_{\sigma(i)} + h\mathbf{t}_i) d\mu(\mathbf{x}_i) d\mathbf{z}_i \\
&= O\Big(\frac{\tilde{r}_{n,n'} h^{-kp}}{|\mathfrak{I}_{k,n}|}\Big) = O\Bigg(\Big(\frac{n \times n' \times h^{(1-k)p}}{|\mathfrak{I}_{k,n}|^2}\Big)^{1/2}\Bigg) = o(1). \,\square
\end{aligned}$$

Chapter 9

Confidence intervals for ratios of means: limitations of the delta method and honest confidence intervals

Abstract

In econometrics, many parameters of interest can be written as a ratio of two expectations. The main method to construct a confidence interval for such a parameter consists in an application of the delta method. Nevertheless, the delta method is an asymptotic procedure, and the obtained intervals may not be relevant if the sample size is small, or if we consider a sequence of models in which statisticians or econometricians study thinner and thinner effects as they have access to more and more data. We prove a generalization of the delta method for ratios of expectations in this "sequence of models" framework. Our paper complements it with a partial impossibility result: nonasymptotic confidence intervals can be built for ratios of expectations, but not at every level. Based on this, we propose an easy-to-compute rule-of-thumb index to appraise the reliability of asymptotic confidence intervals based on the delta method. Some simulations and applications to real data illustrate the practical usefulness of our rule of thumb and how our nonasymptotic confidence intervals compare to the asymptotic ones.

Keywords: Delta-method, confidence regions, uniformly valid inference, sequence of models.

Based on [42] : Derumigny, A., Girard, L., & Guyonvarch Y., On the construction of confidence intervals for ratios of expectations. *Arxiv preprint*, arXiv:1904.07111, 2019.

9.1 Introduction

In applied econometrics, the prevalent method for constructing confidence intervals (CIs) is asymptotic: the theoretical guarantees for most CIs used in practice hold only when the number of observations tends to infinity. For a large class of parameters, the construction of asymptotic CIs also relies on the delta method. In this paper, we focus on parameters that can be expressed as ratios of expectations for which the delta method is a standard procedure to conduct inference. The objective is twofold: provide tools to detect cases in which the delta method may behave poorly and develop inference procedures

that remain valid even in such cases.

Many popular estimands in economics take the form of ratios of expectations. A broad class of such examples are estimands that correspond to conditional expectations, since any conditional expectation with a discrete conditioning variable or a conditioning event can be written as a ratio of unconditional expectations. For instance, assume that we observe an i.i.d. sample of individuals indexed by $i \in \{1, \dots, n\}$ with W_i the wage of an individual and D_i an indicator equal to 1 whenever individual i belongs to some treatment group, say a training program; 0 otherwise. Suppose you are interested in the average wage of participants in the program. We have $\mathbb{E}[W \mid D = 1] = \mathbb{E}[WD] / \mathbb{E}[D]$ as D is binary, which is indeed a ratio of expectations.

As for any parameter, one major goal in practice is to construct confidence intervals that are reliable in finite samples. Nevertheless, CIs based on asymptotic justification are widely used in practice. In that respect, it is crucial to know how such asymptotic CIs perform when the sample size n is finite. For ratios of expectations, we document this issue on simulations (see details in Section 9.3.1). One of our findings is that the coverage of asymptotic CIs based on the delta method happens to be far below the nominal level, even for large sample sizes, when the expectation in the denominator is close to 0. For some scenarios, asymptotic CIs require above 100,000 observations to get reasonably close to their nominal level. Denominators close to 0 are not unusual in practice. Coming back to the treatment/wage example of the previous paragraph, a small denominator would correspond to a binary treatment with a low participation rate.

Besides, it is sometimes of interest to consider sequences of distributions indexed by the sample size as it can rationalize the practice of applied social researchers. The heuristic idea is that researchers can consider narrower effects as the data gets richer. This is similar to some frameworks that have been developed for weak instrumental variables. For instance, a researcher may look at the average value of a variable A of interest in a subgroup of the data. A subgroup could be defined as the intersection of, say, time, geographical area, gender, age, income brackets and so on. As the number of observations n grows, it is possible to consider subgroups g_n that become thinner and thinner (intersection of more and more variables, for examples). This practice could be modelled as estimating $\theta_n := \mathbb{E}[A \mid G_n = 1] = \mathbb{E}[AG_n] / \mathbb{P}(G_n = 1)$ where G_n is a binary variable that is equal to 1 if an individual belongs to the subgroup g_n and $\lim_{n \to +\infty} \mathbb{P}(G_n = 1) = 0$. In such a setting, it is unclear, even asymptotically, what the properties of CIs based on the delta method are. We show that usual CI can fail, and the limiting law of $\theta_n - \hat{\theta}_n$ may not be Gaussian any more, denoting by $\hat{\theta}_n$ the empirical ratio. In some cases, the difference $\theta_n - \hat{\theta}_n$ may actually have a Cauchy limit, for example. Complementing these asymptotic properties, we show on simulations that when the denominator in the ratio goes to zero, those CIs can behave very badly in finite samples too.

The guiding theme of this paper is therefore to develop valid inference procedures in those settings and a practical and easy-to-compute index assessing the reliability of asymptotic CIs for a given dataset with a finite sample size n.

This goal connects to a broad existing literature. In a nutshell, there exist old-established concentration results for expectations, namely upper bounds on the probability that an empirical mean departs from its expectation more than a given threshold, that enable to construct confidence intervals valid for any sample size and for large classes of probability distributions (see in particular [22]). To our knowledge, there is no such result for ratios of expectations. One of the contributions of this paper is to provide similar concentration results for ratios of expectations, which yield nonasymptotic confidence intervals

that are valid for classes of distributions that satisfy suitable moment bounds or support conditions. We consider distributions within a class characterized by a lower bound on the first moment for the denominator variable, and an upper bound on the second moment for both numerator and denominator variables[1].

In addition, our results highlight that there exists a critical confidence level, above which it is not possible to construct nonasymptotic CIs, uniformly valid on such classes, and that are almost surely bounded under every distribution of those classes. More precisely, we exhibit explicit upper and lower bounds on this critical confidence level: the former is a threshold above which we show it is impossible to construct such CIs; the latter is a threshold below which we show how to construct them.

These ideas closely relate to some impossibility results as regards the construction of confidence intervals. A large share of the research effort has concentrated on the problem of constructing confidence intervals for the expectation of a distribution. In an early contribution, [9] show that, when $\mathcal{P}$ is the set of all distributions on the real line with finite expectation, $\theta(P)$ is the expectation with respect to P and $\Theta = \mathbb{R}$, a confidence interval built from an i.i.d. sample of $n \in \mathbb{N}^*$ observations that has coverage $1-\alpha$ must contain any real number with probability at least $1-\alpha$ for every $P \in \mathcal{P}$. Broadly speaking, any confidence interval must have infinite length with positive probability for every $P \in \mathcal{P}$ to ensure a coverage of $1-\alpha > 0$.

Stronger results can be derived when one further restricts $\mathcal{P}$ or Θ. When $\mathcal{P}$ is taken to be the set of all distributions on the real line with variance uniformly bounded by a finite constant (henceforth called the *BC-case*), it is possible to show (using the Bienaymé-Chebyshev inequality) that for every $n \in \mathbb{N}^*$ and every $\alpha \in (0,1)$, there exists a confidence interval that is almost surely bounded under every $P \in \mathcal{P}$ and has coverage $1-\alpha$. In this case, the obtained CIs have the advantage that their length shrinks to 0 at the optimal rate $1/\sqrt{n}$. But on the downside, they are rarely of size $1-\alpha$, even asymptotically, except for some extreme distributions. This means that in practice, they are very conservative.

A strand of the literature has also investigated more complex problems in which $\theta(P)$ is not restricted to being an expectation. For very general parameters, [46] derives a generalization of [9]. An implication of the results in [46] is the existence of an impossibility theorem for ratios of expectations. Let P be a distribution on $\mathbb{R}^2$ with marginals P_X and P_Y. If $\theta(P) = \mathbb{E}_{P_X}[X]/\mathbb{E}_{P_Y}[Y]$, then for every $\alpha \in (0,1)$, it is impossible to build nontrivial CIs of coverage $1-\alpha$ when $\mathcal{P}$ is the set of all distributions on $\mathbb{R}^2$ with finite second moments and $\Theta = \{\theta = \mathbb{E}_{P_X}[X]/\mathbb{E}_{P_Y}[Y] : (\mathbb{E}_{P_X}[X], \mathbb{E}_{P_Y}[Y]) \in \mathbb{R} \times \mathbb{R}^*\}$. As will be explained below, this impossibility result breaks down as soon as $\mathcal{P}$ is chosen such that $|\mathbb{E}_{P_Y}[Y]|$ is bounded away from 0 uniformly over $\mathcal{P}$. Interestingly, the impossibility breaks down only partly in the sense that there remains an upper bound on confidence levels (that depends on n) above which it is impossible to build nontrivial CIs.

Other results for parameters built as differentiable transformations of one or more expectations are given in [118] and [113]: in particular, the latter gives conditions on $\mathcal{P}$ and $\theta : P \mapsto \theta(P)$ under which CIs constructed with the delta method are asymptotically of size $1-\alpha$ uniformly over $\mathcal{P}$.

An interesting consequence of our results is that, even if we assume a known positive lower bound on the expectation in the denominator, the limitation on the coverage of our nonasymptotic CIs remains. That point complements [46] and can be interpreted as a partial impossibility results: for a given sample size n, interesting CIs can be built but not at every confidence level. By contrast, provided the expectation

[1] We refer to this setting as the "Bienaymé-Chebyshev" (BC) case. In Appendix 9.9, we present similar results for distributions whose supports are bounded ("Hoeffding" case).

in the denominator is not null, the delta method gives CIs at every confidence level, but their coverage is only asymptotic.

That discrepancy may cast some doubts on the validity of asymptotic CIs. Hence, we suggest a rule-of-thumb index to assess the reliability of the delta method for ratios of expectations. The heuristic idea is simply, for a given sample, to compute the upper bound on the attainable level of our CIs[2]: for a level higher than that bound, asymptotic CIs based on the delta method might not reach the nominal level and could therefore be suspect. We illustrate the empirical usefulness of that rule on various simulations.

The rest of the paper is organized as follows. Section 9.2 details our framework and assumptions. In Section 9.3, we illustrate the weaknesses of the CIs based on the delta method with a denominator "close to 0" on simulations and detail the asymptotic behavior of the delta-method in such a framework. Section 9.4 is devoted to the construction of nonasymptotic confidence intervals and presents a lower bound on the critical confidence level. In Section 9.5, we derive an upper bound on the critical confidence level and a lower bound on the length of nonasymptotic CIs. This section also includes the description of a practical index to gauge the soundness of asymptotic CIs from the delta method. In Section 9.6, some simulations and applications to a real dataset are presented to illustrate our methods. Section 9.7 concludes. The proofs of all results are postponed to Appendix 9.8. Additional results under an alternative set of assumptions (Hoeffding case) are also detailed in Appendix 9.9. Finally, Appendix 9.10 shows supplementary simulations.

9.2 Our framework

Throughout the paper, for any random variable U and n i.i.d. replications $(U_{1,n}, \ldots, U_{n,n})$, we denote by $\overline{U}_n$ the empirical mean of U, that is $n^{-1}\sum_{i=1}^{n} U_{i,n}$. We present the two main assumptions on the data generating process that we maintain throughout the article.

Assumption 9.2.1. *For every $n \in \mathbb{N}^*$, we observe a sample $(X_{i,n}, Y_{i,n})_{i=1,\ldots,n} \overset{i.i.d.}{\sim} P_{X,Y,n}$, where $P_{X,Y,n}$ is a given distribution on $\mathbb{R}^2$. The real random variables $X_{1,n}$ and $Y_{1,n}$ are such that $\mathbb{E}\left[X_{1,n}^2\right] + \mathbb{E}\left[Y_{1,n}^2\right] < +\infty$, $\mathbb{P}\left(X_{1,n} = 0\right) < 1$ and $\mathbb{E}\left[Y_{1,n}\right] \geq l_{Y,n}$, where $l_{Y,n} > 0$ is known.*

Note that in practice, the value of $l_{Y,n}$ may not be available for the statistician. This is the reason why, in Section 9.5.2, we propose practical methods that do not need the knowledge of $l_{Y,n}$. It is worth noting that n indexes both the distribution $P_{X,Y,n}$ of the observations in this model, and the number of observations n. This encompasses the standard i.i.d. setup if the distribution of the observations does not change with n: for every $n \in \mathbb{N}^*$, $P_{X,Y,n} = P_{X,Y}$ for some given distribution $P_{X,Y}$.

The assumption $\mathbb{P}\left(X_{1,n} = 0\right) < 1$ simply rules out the possibility of having a numerator almost surely equal to 0, which is a very mild restriction in practice. The last part of the assumption is more interesting. As we assume the existence of a finite expectation, we can consider $\mathbb{E}\left[Y_{1,n}\right] \geq 0$ without loss of generality[3]. In order to have properly defined ratios of interest, we need to assume away a null denominator, namely suppose that for every $n \in \mathbb{N}^*$, $\mathbb{E}\left[Y_{1,n}\right] > 0$. The assumption $\mathbb{E}\left[Y_{1,n}\right] \geq l_{Y,n} > 0$ for every $n \in \mathbb{N}^*$ is stronger but, necessary to derive nonasymptotic CIs with maintained coverage and that are not trivial. Otherwise, if $l_{Y,n} = 0$, the impossibility theorem of [46] applies and prevents from constructing nontrivial CIs no matter the confidence level. Our assumption allows $l_{Y,n}$ to decrease to 0 though, which enables us to handle cases where the delta method may fail.

[2] Equivalently, for a stated nominal level for the confidence interval, we can compute the minimum required sample size.

[3] Otherwise, we simply replace $Y_{i,n}$ by its opposite $-Y_{i,n}$.

In a way, given the results of [46], Assumption 9.2.1 can be seen as close to the minimal hypothesis that allows for the possibility of nontrivial confidence intervals for a ratio of expectations in a nonasymptotic framework.

Assumption 9.2.2. *There exist finite constants $u_{X,n}, u_{Y,n}$ such that the second moments of $X_{1,n}$ and $Y_{1,n}$ are bounded, i.e. $\mathbb{E}\left[X_{1,n}^2\right] \leq u_{X,n}$, and $\mathbb{E}\left[Y_{1,n}^2\right] \leq u_{Y,n}$.*

This framework, where Assumptions 9.2.1 and 9.2.2 hold, will be denoted as the *BC case* since it is possible under this assumption to construct CIs using the Bienaymé-Chebyshev inequality. In Appendix 9.9, we present an adapted version of all our results under the assumption that $X_{1,n}$ and $Y_{1,n}$ have a bounded support instead of Assumption 9.2.2. This corresponds to what we will call the *"Hoeffding case"* because under such assumptions, we can use the Hoeffding inequality to build nonasymptotic confidence intervals with exponential speed of convergence.

Our objective is to make inference on the ratio of expectations $\mathbb{E}\left[X_{1,n}\right]/\mathbb{E}\left[Y_{1,n}\right]$ by constructing confidence intervals. To sum up, Assumptions 9.2.1 and 9.2.2 define a set $\mathcal{P}$ of distributions for some known constants $l_{Y,n}$, $u_{X,n}$ and $u_{Y,n}$. For a distribution $P_{X,Y,n}$ in $\mathcal{P}$, the parameter of interest is $\theta(P_{X,Y,n}) = \mathbb{E}\left[X_{1,n}\right]/\mathbb{E}\left[Y_{1,n}\right]$ with values in $\Theta = \mathbb{R}$. We will study confidence intervals C_n that are functions of the n bivariate observations $(X_{1,n}, Y_{1,n}), \ldots, (X_{n,n}, Y_{n,n})$.

In practice, it is possible that $\overline{Y}_n = 0$ for a given sample, and it may even happen with a strictly positive probability for some non-continuous distributions of Y. This means that $\hat{\theta}_n := \overline{X}_n/\overline{Y}_n$ does not exist for such samples. In such a case, it is difficult to construct a meaningful confidence interval. Different conventions are possible:

- We could choose to define $C_n = \mathbb{R}$ in this case. This means that the true parameter θ belongs to C_n, by definition. We believe that such a case would artificially improve the coverage of C_n, as it means that, the higher $\mathbb{P}(\overline{Y}_n = 0)$, the better our interval would be in term of coverage.
- We could choose $C_n = \emptyset$. This means that the hypothesis $\theta = \theta_0$ would be rejected for each $\theta_0 \in \mathbb{R}$, using the duality between tests and confidence intervals. This also something that we would like to avoid, as it may not seem reasonable to reject every choice of θ_0 for the only reason that it cannot be estimated with the current sample.
- Other choices are possible, for example $C_n = \{0\}$, but they do not seem reasonable either, as there is no reason to select only 0 in our confidence interval, especially if $\overline{X}_n \neq 0$.

For these reasons, we choose to let C_n undefined for such samples, in the same way as the ratios $x/0$ are undefined for any real x. In practice, it means that nothing can be said: for any value $x \in \mathbb{R}$, it is undefined whether x belongs or not to C_n (in the same way as it is undefined whether $1/0$ is smaller or higher than 2). Of course, when meeting such a case, the applied statistician could change the sample, for example by collecting more data. One could also consider sub-samples (possibly several and combine them in some way) of data for which the empirical mean of the denominator differs from 0. Nevertheless, the construction of satisfactory estimators in this case lies beyond the scope of this paper.

9.3 Limitations of the delta method

In general, the coverage of asymptotic CIs with nominal level $1-\alpha$ is uncontrolled for finite samples: for a sample of size n, the coverage of asymptotic CIs may be well below $1-\alpha$. Intuitively, this phenomenon

should be driven by "problematic" distributions in $\mathcal{P}$ in the following sense: when P is close to the boundary of $\mathcal{P}$, the probability $c(n,P) := \mathbb{P}_{P^{\otimes n}}\left(C_n \ni \theta(P)\right)$ may be much smaller than $1-\alpha$.

Let us focus on our case of interest where $\theta(P)$ is a ratio of expectations. For this problem, it is possible to build valid asymptotic CIs based on the delta method. We recall this fact in Section 9.3.1 and illustrate on simulations that when the expectation in the numerator is close to zero, $c(n,P)$ can be well below the nominal level of the CIs and it may require a very large number of observations to make $c(n,P)$ reasonably close to the nominal level. In Section 9.3.2, we investigate a more serious issue: in the sequence-of-model framework presented in Section 9.2, we let the expectation in the denominator not only be close to zero, but converge to zero as n increases. We show on simulations that depending on the speed at which the denominator goes to zero, $c(n,P)$ can either converge to the nominal level (more or less fast) or even not converge at all to the nominal level. This sheds light on a partial failure of the delta method when the denominator goes to zero that we derive formally in Section 9.3.3.

9.3.1 Asymptotic approximation takes time to hold

We first recall how to derive asymptotic confidence intervals based on the delta method in the case where, for every $n \in \mathbb{N}^*$, $P_{X,Y,n}$ is identical, hence denoted $P_{X,Y}$. In this case, we simply denote $\mathbb{E}[X] = \mathbb{E}\left[X_{1,n}\right]$ and $\mathbb{E}[Y] = \mathbb{E}\left[Y_{1,n}\right]$. Under Assumption 9.2.1, combining the multivariate central limit theorem and the delta method yields

$$\sqrt{n}\left(\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}[X]}{\mathbb{E}[Y]}\right) \xrightarrow[n\to+\infty]{d} \mathcal{N}\left(0, \nabla\phi' \Sigma \nabla\phi\right), \tag{9.1}$$

where $\nabla\phi'\Sigma\nabla\phi = \mathbb{V}\left[X\right]/\mathbb{E}[Y]^2 + \mathbb{E}[X]^2\mathbb{V}[Y]/\mathbb{E}[Y]^4 - 2\mathbb{C}ov\left[X,Y\right]\mathbb{E}[X]/\mathbb{E}[Y]^3$ and ϕ is the function $(x,y) \mapsto x/y$. As with expectations, we use the shortcuts $\mathbb{V}\left[X\right] = \mathbb{V}\left[X_{1,n}\right]$ and $\mathbb{C}ov\left[X,Y\right] = \mathbb{C}ov\left[X_{1,n}, Y_{1,n}\right]$. Based on (9.1), we can construct an asymptotic CI for the parameter of interest $\mathbb{E}[X]/\mathbb{E}[Y]$ using a consistent estimate of the asymptotic variance $\nabla\phi'\Sigma\nabla\phi$ and Slutsky's lemma.

To assess the quality of this CI, we compute $c(n,P)$ using simulations for different sample sizes n and distributions P and compare it to the nominal level. By definition, $c(n,P)$ forms an upper bound of the coverage rate. We therefore denote $c(n,P)$ the *maximal coverage*, in the sense that the true coverage of our CIs may be lower, but it could not be higher than $c_{n,P}$. We focus on CIs with a 95% nominal level. More specifically, for different sample sizes n and values of $\mathbb{E}[Y]$, we draw 5000 independent samples of size n $(X_i, Y_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(1,1) \otimes \mathcal{N}(\mathbb{E}[Y], 1)$. We compute the *maximal coverage* for CIs based on the delta method for every pair $(n, \mathbb{E}[Y])$ using the 5000 replications. $\mathbb{E}[Y]$ ranges from 0.01 (the denominator is close to zero) to 0.75 (the denominator is far from zero). Figure 9.1 sums up the results. For every n, it turns out that the closer $\mathbb{E}[Y]$ to 0, the smaller the *maximal coverage* of the delta method. When $\mathbb{E}[Y] = 0.01$, the *maximal coverage* gets close to the nominal level for n around 300,000.

9.3.2 Asymptotic results may not hold for sequences of models

Unlike the result displayed in (9.1), it is unclear how the quantity $\sqrt{n}\left(\overline{X}_n/\overline{Y}_n - \mathbb{E}[X]/\mathbb{E}[Y]\right)$ behaves asymptotically when we consider sequences of statistical models with the expectation in the denominator tending to 0 as n increases. For a given specification, Figure 9.2 shows the *maximal coverage* of asymptotic CIs based on the delta method when $\mathbb{E}\left[Y_{1,n}\right] = Cn^{-b}$ where C is set to 0.025 and b varies. For a speed $b \geq 1/2$, $1/2$ being the usual speed of the CLT, the *maximal coverage* of asymptotic CIs based on (9.1) is incorrect in the sense that it is far lower than the nominal level $1-\alpha$ and it does not

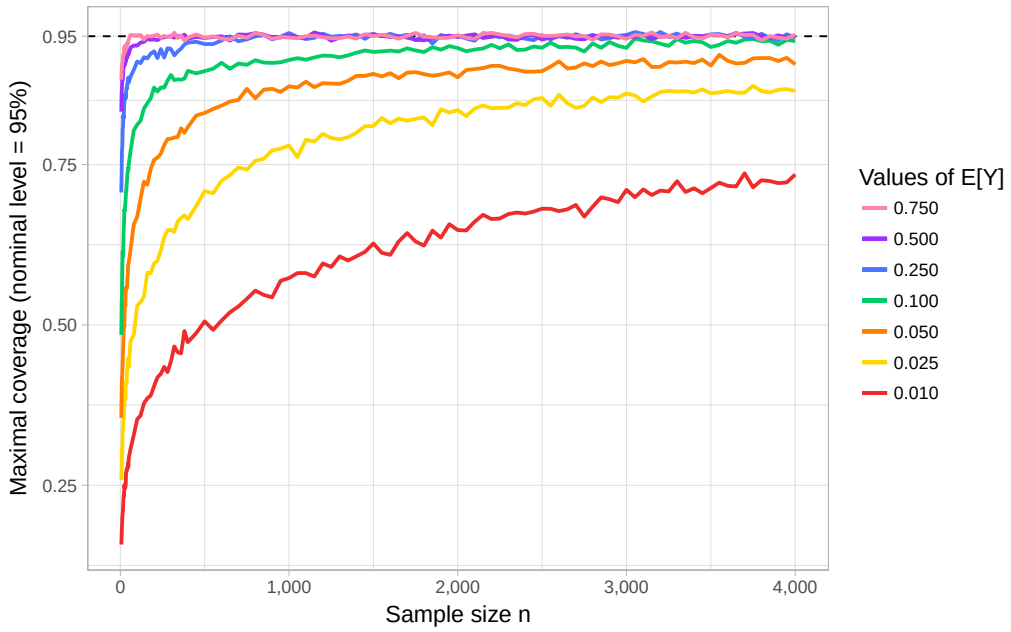


Figure 9.1: Maximal coverage of asymptotic CIs based on the delta method.
Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1,1) \otimes \mathcal{N}(\mathbb{E}[Y], 1)$. The nominal pointwise asymptotic level is set to 0.95. For a pair $(\mathbb{E}[Y], n)$, the coverage is obtained as the mean over 5,000 repetitions.

converge to the latter. Our simulations even suggest that the *maximal coverage* tends to 0 for $b > 1/2$. For $b < 1/2$, the *maximal coverage* of the delta method seems to tend to $1-\alpha$. Yet, in line with Figure 9.1, the validity of the asymptotic approximation requires very large sample sizes.

At this stage, Figure 9.2 present some evidence that the standard asymptotic CIs based on the delta method need to be adapted for sequences of models and that the speed of decrease toward 0 of the expectation $\mathbb{E}[Y_{1,n}]$ matters. The next subsection detail formal results to adapt the delta method for ratios of expectations in this setup.

9.3.3 Extension of the delta method for ratios of expectations in the sequence-of-models framework

We are interested in the asymptotic distribution, as n tends to infinity, of the real random variable $S_n := \sqrt{n}\left(\overline{X}_n/\overline{Y}_n - \mathbb{E}[X_{1,n}]/\mathbb{E}[Y_{1,n}]\right)$. The following theorem states the asymptotic behavior of S_n according to the comparison of $1/\sqrt{n}$ and $\mathbb{E}[Y_{1,n}]$, under a multivariate Lindeberg condition. As the distributions $P_{X,Y,n}$ change with n without any link from one to the next, it is not possible to obtain equivalents almost surely or in probability. To overcome this difficulty, we can only consider convergences in distribution, or here equivalents in distributions. We say that two sequences of random variables X_n and Y_n are equivalent in distribution if there exists a probability space $\tilde{\Omega}$ and two sequences of random variables $\tilde{X}_n, \tilde{Y}_n$ such that $\forall n \in \mathbb{N}$, $X_n \overset{d}{=} \tilde{X}_n$, and $Y_n \overset{d}{=} \tilde{Y}_n$, and $\tilde{X}_n$ is equivalent to $\tilde{Y}_n$ almost surely, as $n \to \infty$. This theorem is proved in Section 9.8.1.

Theorem 9.1. *Let Assumption 9.2.1 hold. Assume that* $\mathbb{V}[(X_{1,n}, Y_{1,n})] \to V$ *where* V *is a positive definite matrix, that* $\mathbb{P}(\overline{Y}_n = 0) \to 0$*, as* $n \to \infty$ *and that* $\sup_{n\in\mathbb{N}^*} \mathbb{E}\left[|X_{1,n}|^3\right]$ *and* $\sup_{n\in\mathbb{N}^*} \mathbb{E}\left[|Y_{1,n}|^3\right]$ *are finite.*

Then, the sequence of random variables $S_n := \sqrt{n}\Big(\overline{X}_n \,/\, \overline{Y}_n - \mathbb{E}[X_{1,n}] \,/\, \mathbb{E}[Y_{1,n}]\Big)$ *satisfies:*

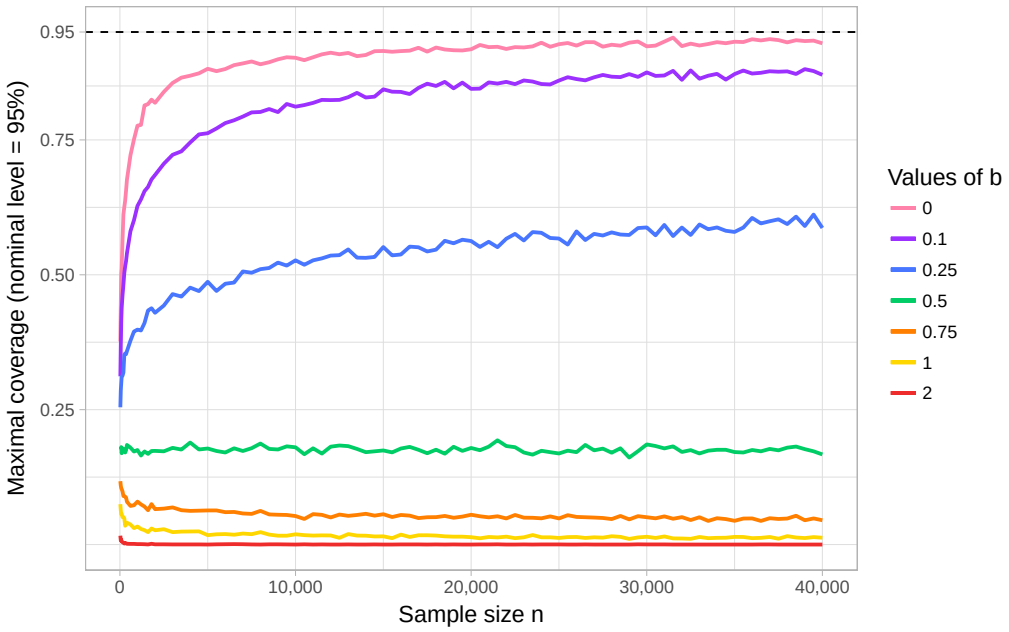


Figure 9.2: Coverage of asymptotic CIs based on the delta method.
Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1,1) \otimes \mathcal{N}(Cn^{-b}, 1)$, with $C = 0.025$. The nominal pointwise asymptotic level is set to 0.95. For a pair $(\mathbb{E}[Y], n)$, the coverage is obtained as the mean over 5,000 repetitions.

1. *If* $n^{1/2}\mathbb{E}\left[Y_{1,n}\right] \to 0$*, then* S_n *is equivalent in distribution to:*

$$\frac{\sqrt{n}\left(\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\right)}{\mathbb{E}\left[Y_{1,n}\right]} - \frac{\sqrt{n}\left(\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]\right)\mathbb{E}\left[X_{1,n}\right]}{\mathbb{E}\left[Y_{1,n}\right]^2}, \quad \textit{as } n \to \infty.$$

2. *If* $n^{1/2}\mathbb{E}\left[Y_{1,n}\right] \to +\infty$*, then* S_n *is equivalent in distribution to:*

$$\sqrt{n}\left(\frac{\sqrt{n}(\overline{X}_n - \mathbb{E}\left[X_{1,n}\right])}{\sqrt{n}(\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right])} - \frac{\mathbb{E}\left[X_{1,n}\right]}{\mathbb{E}\left[Y_{1,n}\right]}\right), \quad \textit{as } n \to \infty.$$

3. *If there exists a finite constant* $C \neq 0$ *such that* $\sqrt{n}\mathbb{E}\left[Y_{1,n}\right] \to C$ *as* $n \to \infty$*, then* S_n *is equivalent in distribution to:*

$$n\mathbb{E}\left[X_{1,n}\right]\left(\frac{1}{C + \sqrt{n}\left(\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]\right)} - \frac{1}{C}\right) + \frac{n\left(\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\right)}{C + \sqrt{n}\left(\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]\right)}, \quad \textit{as } n \to \infty.$$

Theorem 9.1 can thus be interpreted as a generalization of the result given by the CLT and the delta method for ratios of expectations. The sequence-of-models framework allows the expectation in the denominator to tend to 0. Figure 9.3 and its companion table highlight the different asymptotic regimes depending on the behaviors of $\{\mathbb{E}\left[X_{1,n}\right]\}_{n\in\mathbb{N}^*}$ and $\{\mathbb{E}\left[Y_{1,n}\right]\}_{n\in\mathbb{N}^*}$.

The main takeaway of the latter is that when $\mathbb{E}\left[X_{1,n}\right] = C_1/n^a$ and $\mathbb{E}\left[Y_{1,n}\right] = C_2/n^b$ for some constants $C_1 \neq 0$ and $C_2 \neq 0$, and $b < 1/2$ (namely the expectation in the denominator converges to 0 at a slower rate than the standard CLT one), S_n properly renormalized by n to a power that is a function of a and b, still converges in distribution to a Normal random variable. Asymptotically valid inference based on Normal approximation remains valid in that case, even if the length of such confidence intervals may not decrease with the sample size n. In all other cases (except when $a > b > 1/2$), up to a normalization of some power of n, S_n converges weakly to a non-Gaussian distribution, in some cases to generalized

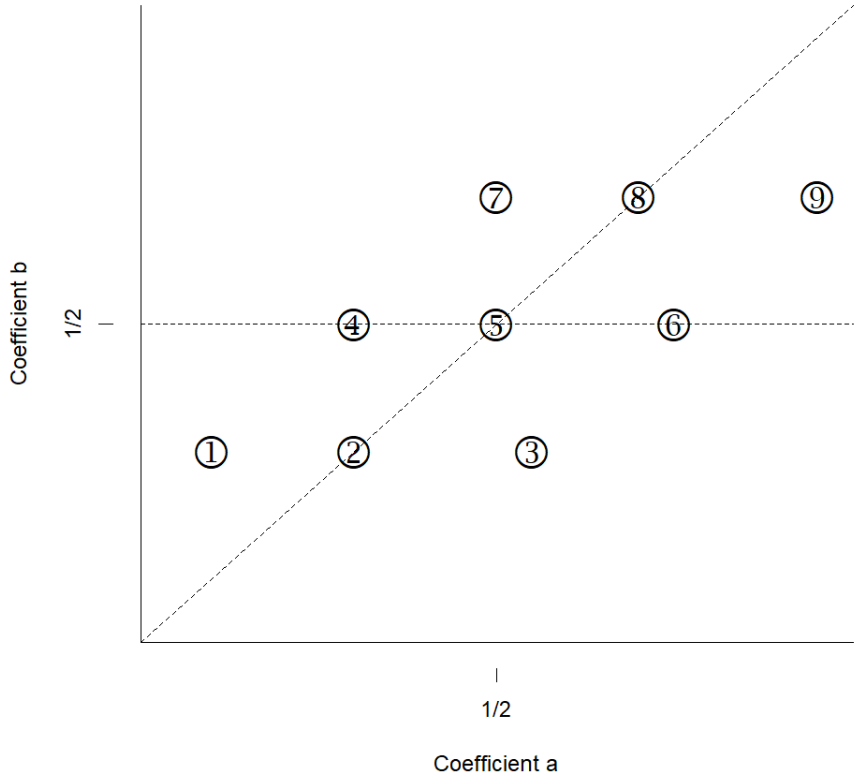


Figure 9.3: Separation between the different asymptotic regimes as a function of (a, b), where $\mathbb{E}[X_{1,n}] = C_1/n^a$ and $\mathbb{E}[Y_{1,n}] = C_2/n^b$, $(a, b) \in {\mathbb{R}_+^*}^2$.

	$a < b$	$a = b$	$a > b$
$b > 1/2$	$n^{1/2} W_1/W_2$	$n^{1/2}\big(W_1/W_2 - C_1/C_2\big)$	$-n^{b-a+1/2} C_1/C_2$
$b = 1/2$	$n^{1-a}\Big(C_1/(C_2+W_2) - C_1/C_2\Big)$	$n^{1/2}\big(C_1/(C_2+W_2) - C_1/C_2 + W_1/(C_2+W_2)\big)$	$n^{1/2}\big(W_1/(C_2+W_2)\big)$
$b < 1/2$	$n^{2b-a} C_1 W_2/C_2^2$	$n^b(W_1/C_1 - C_1 W_2/C_2^2)$	$n^b W_1/C_1$

Table 9.1: Limiting law of $S_n := \sqrt{n}\big(\overline{X}_n/\overline{Y}_n - \mathbb{E}[X_{1,n}]/\mathbb{E}[Y_{1,n}]\big)$ in the nine different regimes. The variables (W_1, W_2) follow the distribution $\mathcal{N}(0, V)$, where $V = \lim \mathbb{V}[(X_{1,n}, Y_{1,n})]$.

Cauchy distributions with parameters that depend on the data generating process. Theoretically, quantiles of the limiting distribution could be computed and thus asymptotic confidence intervals, as in the Normal case. Overall, the results of the table highlight that the method to conduct inference will depend on the speed of convergence of both numerators and denominators. In comparison, the nonasymptotic confidence intervals shown in Section 9.4 provide a unique way of conducting inference, once a class of distributions is defined.

9.4 Construction of nonasymptotic confidence intervals

To construct nonasymptotic confidence intervals for ratios of expectations, we rely on the possibility to ensure that, with large probability, (i) $\overline{X}_n$ is close to $\mathbb{E}[X_{1,n}]$ and (ii) $\overline{Y}_n$ is both close to $\mathbb{E}[Y_{1,n}]$ and bounded away from 0. Under Assumptions 9.2.1 and 9.2.2, the Bienaymé-Chebyshev inequality can be applied to obtain (i) and (ii). Without extra assumptions, we are only able to build nonasymptotic CIs at confidence levels that are not too close to 1 (see Section 9.4.2). This limitation does not arise when one builds a nonasymptotic CI for an expectation. In that sense, we can say that building a nonasymptotic CI for a ratio of expectations is more demanding. Intuitively, the extra difficulty of the latter task comes from the need to ensure (ii). To stress that point, we show in the next subsection that when $\overline{Y}_n$ is bounded away from 0 and positive almost surely, we can build a nonasymptotic CI for a ratio at every confidence level.

9.4.1 An easy case: the support of Y is well-separated from 0

We present a simple framework in which it is possible to build nonasymptotic CIs, valid for every $n \in \mathbb{N}^*$, and with coverage $1-\alpha$ for every $\alpha \in (0,1)$. To do so, we restrict further the set $\mathcal{P}$ of admissible distributions with the following assumption.

Assumption 9.4.1. *There exists $a_{Y,n} > 0$ such that $Y_{1,n} \geq a_{Y,n}$ almost surely.*

Under Assumption 9.4.1, for every $n \in \mathbb{N}^*$, $\overline{Y}_n \geq a_{Y,n} > 0$ almost surely under every distribution in $\mathcal{P}$ and $\overline{Y}_n^{-1}$ is bounded from above. This assumption obviously rules out binary $\{0,1\}$ random variables in the denominator of the ratio, which can be quite restrictive in practice. Under this assumption, the following theorem gives a concentration inequality for our ratio of expectations. It is proved in Section 9.8.2.

Theorem 9.2. *Under Assumptions 9.2.1, 9.2.2 and 9.4.1, we have for every $n \in \mathbb{N}^*$ and $\varepsilon \in \mathbb{R}_+^*$,*

$$\mathbb{P}\left(\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}[X_{1,n}]}{\mathbb{E}[Y_{1,n}]}\right| > \frac{\left(\varepsilon + \sqrt{u_{X,n}}\right)\varepsilon}{a_{Y,n} l_{Y,n}} + \frac{\varepsilon}{l_{Y,n}}\right) \leq \frac{u_{X,n}}{n\varepsilon^2} + \frac{u_{Y,n} - l_{Y,n}^2}{n\varepsilon^2}.$$

As a consequence, $\mathbb{P}\left(\mathbb{E}[X_{1,n}]/\mathbb{E}[Y_{1,n}] \in [\overline{X}_n/\overline{Y}_n \pm t]\right) \geq 1-\alpha$, with the choice

$$t := \frac{1}{l_{Y,n}}\sqrt{\frac{u_{X,n} + u_{Y,n} - l_{Y,n}^2}{n\alpha}}\left(1 + \frac{1}{a_{Y,n}}\left\{\sqrt{\frac{u_{X,n} + u_{Y,n} - l_{Y,n}^2}{n\alpha}} + \sqrt{u_{X,n}}\right\}\right),$$

for every $\alpha \in (0,1)$.

The theorem shows that it is possible to construct nonasymptotic CIs for ratios of expectations, with guaranteed coverage at every confidence level, that are almost surely bounded under every distribution in $\mathcal{P}$. In Section 9.4.2, we give an analogous result that only requires Assumptions 9.2.1 and 9.2.2 to hold, so that it encompasses the case of $\{0,1\}$-valued denominators. However, the cost to pay will be an upper bound on the achievable coverage of the confidence intervals.

9.4.2 Nonasymptotic confidence intervals with no assumption on the support of P_Y

We seek to build nontrivial nonasymptotic CIs under Assumptions 9.2.1 and 9.2.2 only. Under Assumption 9.2.1, $\mathbb{E}[Y_{1,n}] \neq 0$, so that there is no issue in considering the fraction $\mathbb{E}[X_{1,n}]/\mathbb{E}[Y_{1,n}]$. However, without Assumption 9.4.1, $\{\overline{Y}_n = 0\}$ has positive probability in general so that $\overline{X}_n/\overline{Y}_n$ is well-defined with probability less than one. Note that when $P_{Y,n}$ is continuous wrt to Lebesgue's measure, there is no issue in defining $\overline{X}_n/\overline{Y}_n$ anymore since the event $\{\overline{Y}_n = 0\}$ has probability zero. This is not an easier case from a theoretical point-of-view though, since without more restrictions, $\overline{Y}_n$ can still be arbitrarily close to zero with positive probability.

Theorem 9.3. *Assume that Assumptions 9.2.1 and 9.2.2 hold. For every $n \in \mathbb{N}^*$, $\varepsilon > 0, \tilde{\varepsilon} \in (0,1)$, we have*

$$\mathbb{P}\left(\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}[X_{1,n}]}{\mathbb{E}[Y_{1,n}]}\right| > \left(\frac{(\sqrt{u_{X,n}}+\varepsilon)\tilde{\varepsilon}}{(1-\tilde{\varepsilon})^2} + \varepsilon\right)\frac{1}{l_{Y,n}}\right) \leq \frac{u_{X,n}}{n\varepsilon^2} + \frac{u_{Y,n}-l_{Y,n}^2}{n\tilde{\varepsilon}^2 l_{Y,n}^2}.$$

As a consequence, $\mathbb{P}\Big(\big|\overline{X}_n/\overline{Y}_n - \mathbb{E}[X_{1,n}]/\mathbb{E}[Y_{1,n}]\big| > t\Big) \leq \alpha$, with the choice

$$t = \frac{1}{l_{Y,n}}\left(\frac{\left(\sqrt{u_{X,n}}+\sqrt{2u_{X,n}/n\alpha}\right)\sqrt{2(u_{Y,n}-l_{Y,n}^2)/n\alpha l_{Y,n}^2}}{\left(1-\sqrt{2(u_{Y,n}-l_{Y,n}^2)/n\alpha l_{Y,n}^2}\right)^2} + \sqrt{\frac{2u_{X,n}}{n\alpha}}\right),$$

for every $\alpha > \overline{\alpha}_n := \frac{2(u_{Y,n}-l_{Y,n}^2)}{nl_{Y,n}^2}$.[4]

This theorem is proved in Section 9.8.3. It states that when $l_{Y,n} > 0$, it is possible to build valid nonasymptotic CIs with finite length up to the confidence level $1-\overline{\alpha}_n$. This is a more positive result than [46] which states that it is not possible to build nontrivial nonasymptotic CIs when $l_{Y,n}$ is taken equal to 0, no matter the confidence level. On the other hand, it is more negative than Theorem 9.2. Note that Theorem 9.3 is not an impossibility theorem since it only claims that considering confidence levels smaller than $1-\overline{\alpha}_n$ is *sufficient* to build nontrivial CIs under Assumptions 9.2.1 and 9.2.2. The remaining question is to find out whether it is *necessary* to focus on confidence levels that do not exceed a certain threshold under Assumptions 9.2.1 and 9.2.2. We answer this in Section 9.5.1.

Theorem 9.3 has two other interesting consequences: for every confidence level up to $1-\overline{\alpha}_n$, a nonasymptotic CI of the form $\left[\overline{X}_n/\overline{Y}_n \pm \tilde{t}\right]$ with $\tilde{t} > t$ has coverage $1-\alpha$ but is conservative. Moreover, if the DGP does not depend on n (*i.e* in the standard i.i.d. setup), for every fixed α, the length of the confidence interval shrinks at the optimal rate $1/\sqrt{n}$. Note that the coefficient 2 in the definition of α_n defined above can be reduced to any number $w > 1$, at the expense of increasing the length of the confidence interval. In fact, the latter tends to infinity when $w \to 1$. This is due to the fact we equalize both terms in the bound of the probability above, and more subtle choices are possible indeed.

9.5 Nonasymptotic CIs: impossibility results and practical guidelines

In this section, we prove two impossibility results: a necessary lower bound on the length of nonasymptotic CIs and a maximum confidence level above which it is impossible to build nontrivial nonasymptotic CIs.

[4] Equivalently, it means that for a given level α, the above choice of t is valid for every integer $n > \overline{n}_\alpha := 2(u_{Y,n}-l_{Y,n}^2)/\alpha l_{Y,n}^2$.

9.5.1 An upper bound on testable confidence levels

Proposition 9.4. *For every $n \in \mathbb{N}^*$, and every $\alpha \in (0, \underline{\alpha}_n)$, where $\underline{\alpha}_n := \left(1-l_{Y,n}^2/u_{Y,n}\right)^n$, if $l_{Y,n}^2/u_{Y,n} < 1$, there is no finite $t > 0$ such that $\left[\overline{X}_n/\overline{Y}_n \pm t\right]$ has coverage $1-\alpha$ over $\mathcal{P}$, where $\mathcal{P}$ is the class of all distributions satisfying Assumptions 9.2.1 and 9.2.2 for fixed $l_{Y,n}$, $u_{X,n}$ and $u_{Y,n}$.*

This theorem asserts that confidence intervals of the form $\left[\overline{X}_n/\overline{Y}_n \pm t\right]$ with coverage higher than $1-\underline{\alpha}_n$ under Assumptions 9.2.1 and 9.2.2 are not defined (or are of infinite length) with positive probability for at least one distribution in $\mathcal{P}$. This is due to the fact that $\underline{\alpha}_n$ is the lower bound on $\mathbb{P}(\overline{Y}_n = 0)$ over all distributions in $\mathcal{P}$.

Remark that when $u_{Y,n}/l_{Y,n}^2 = 1$, there is no impossibility result anymore: assume that $u_{Y,n}/l_{Y,n}^2 = 1$ and let Q be a distribution on $\mathbb{R}^2$ that satisfies Assumptions 9.2.1 and 9.2.2. Let $(X_{i,n}, Y_{i,n})_{i=1}^n \overset{\text{i.i.d.}}{\sim} Q$. We have that $\mathbb{V}[Y_{1,n}] = 0$, which implies that $Y_{1,n} = \mathbb{E}[Y_{1,n}]$ almost surely. Assumption 9.2.1 further ensures that $Y_{1,n} \neq 0$ almost surely. Consequently, the results of Section 9.4.1 apply and allow us to conclude that under Assumptions 9.2.1, 9.2.2 and $u_{Y,n}/l_{Y,n}^2 = 1$, it is possible to build nontrivial nonasymptotic CIs at every confidence level. Indeed, in that case, we are in fact only estimating a simple mean, and therefore there is no constraint on α.

Proposition 9.4 is in fact a corollary of the more general Theorem 9.5 that states that it is impossible to construct confidence intervals that contain $\overline{X}_n/\overline{Y}_n$ almost surely and are almost surely bounded over $\mathcal{P}$ with coverage greater than $1-\underline{\alpha}_n$. It is proven in Section 9.8.5.

Theorem 9.5. *Let $n \in \mathbb{N}^*$, and a random set I_n that contains $\overline{X}_n/\overline{Y}_n$ almost surely whenever it is defined and is undefined if $\overline{Y}_n = 0$. Then $\sup_{P_n \in \mathcal{P}} \mathbb{P}\big(I_n \, undefined\big) \geq \underline{\alpha}_n$.*

Combining Theorems 9.3 and 9.5, we conclude that there exists some critical level $1-\alpha_n^c$ that belongs to the interval $[1-\overline{\alpha}_n, 1-\underline{\alpha}_n]$ such that it is impossible to build nontrivial nonasymptotic CIs if and only if their nominal level is above $1-\alpha_n^c$. Finally, it is worth remarking that with a sample of size n, CIs based on the delta method with a nominal level $1-\alpha > 1-\alpha_n^c$ cannot have coverage $1-\alpha$.

9.5.2 Practical methods and plug-in estimators

Nonasymptotic confidence intervals based on Theorem 9.3 require Assumptions 9.2.1 and 9.2.2. In particular, they require the knowledge of the constants $l_{Y,n}$, $u_{X,n}$ and $u_{Y,n}$ that determine the class of distributions we consider. In practice, we need to state values for $l_{Y,n}$, $u_{X,n}$ and $u_{Y,n}$ to build our nonasymptotic CIs and to compute $\overline{\alpha}_n$ (equivalently $\overline{n}_\alpha$)[5] or $\underline{\alpha}_n$. Note that constructing nontrivial and nonasymptotic CIs that overcome the limitations of having to choose some a priori class of distributions is not possible. Indeed, we would get back to [9] and [46] type impossibility results.

How to choose $l_{Y,n}$, $u_{X,n}$ and $u_{Y,n}$ depends on the specific application. In some cases, stating values can be sensible if researchers do have (some) control or expert knowledge of the variables. Resuming an example started in Section 9.1, if the variable in the denominator happens to be an indicator of participating to some treatment and in the setting of a Randomized Controlled Trial, researchers can have intuitions about reasonable values for the upper bound $u_{Y,n}$ of the probability of being treated, for instance.

Without prior information, a first approximation would be to replace the unknown constants with their empirical counterparts. Under i.i.d. sampling, the latter are consistent estimates of the former; though

[5] Actually, the computation of $\overline{\alpha}_n$ and $\overline{n}_\alpha$ only require knowledge of $l_{Y,n}$ and $u_{Y,n}$.

it ruins the exact nonasymptotic approach. On the other hand, it enables us to construct our CIs and the quantity $\overline{n}_\alpha$, which can be useful in practice as a rule of thumb. We stick to that principle in all our numerical applications (Section 9.6).

For a given level $1-\alpha$ and a class of distributions satisfying Assumptions 9.2.1 and 9.2.2 for some values $l_{Y,n}$, $u_{X,n}$ and $u_{Y,n}$, $\overline{n}_\alpha$ is the minimal sample size required to construct our nonasymptotic CIs. For a sample size $n < \overline{n}_\alpha$, the data is not rich enough to construct nonasymptotic CIs of Theorem 9.3 at this level. Heuristically, the comparison of $\overline{n}_\alpha$ and n can be used as a rule of thumb to quickly assess whether the nominal level of asymptotic CIs based on the delta method holds[6]. Several simulations tend to confirm the practical interest of that rule of thumb as the obtained $\overline{n}_\alpha$ turns out to be empirically very close to the sample size above which the coverage of asymptotic CIs converges to the nominal level $1-\alpha$ (see Section 9.6.1).

9.5.3 A lower bound on the length of nonasymptotic confidence intervals

The following theorem is an extension of [26][Proposition 6.2] to ratios. It is proved in Section 9.8.4.

Theorem 9.6. *For every integer $n \geq 7$, every $\alpha \in \left(0, 1 \wedge n/\left(l_{Y,n} + \sqrt{u_{Y,n} - l_{Y,n}^2}\right)^2\right)$, and every $\xi < 1$ there exists a distribution Q on $\mathbb{R}^2$ that satisfies Assumptions 9.2.1 and 9.2.2 such that for $(X_{i,n}, Y_{i,n})_{i=1}^n \overset{i.i.d}{\sim} Q$, we have*

$$\mathbb{P}\left(\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}\left[X_{1,n}\right]}{\mathbb{E}\left[Y_{1,n}\right]}\right| > \xi\sqrt{\frac{v_n}{3n\alpha}}\right) > \alpha,$$

where $v_n := u_{X,n}/\left(l_{Y,n} + \sqrt{u_{Y,n} - l_{Y,n}^2}\right)^2$.

With this theorem, we can claim that for every $\alpha \in \left(0, 1 \wedge n/\left(l_{Y,n} + \sqrt{u_{Y,n} - l_{Y,n}^2}\right)^2\right)$, confidence intervals of the form $\left[\overline{X}_n/\overline{Y}_n \pm t\right]$ cannot have uniform coverage $1-\alpha$ under Assumptions 9.2.1 and 9.2.2 if they are shorter than $\sqrt{v_n/(3n\alpha)}$. By a careful inspection of the proof (see Lemma 9.9), we can in fact replace the value 3 in the theorem by any number strictly larger than $e = \exp(1)$, at the price of assuming $n \geq n_0$ for n_0 large enough. It is interesting to note that the distributions Q that are built in the proof of the theorem are extremal in $\mathcal{P}$ in the sense that they satisfy $\mathbb{E}\left[X_n^2\right] = u_{X,n}$, $\mathbb{E}\left[Y_{1,n}\right] = l_{Y,n}$ and $\mathbb{E}\left[Y_n^2\right] = u_{Y,n}$.

9.6 Numerical applications

9.6.1 Simulations

This section presents simulations that support the use of $\overline{n}_\alpha$, or equivalently $\overline{\alpha}_n$, as a rule of thumb to give insight into the reliability of the asymptotic confidence intervals from the delta method. The simulations resume the setting of Figure 9.1.

In Figure 9.4, a nominal level $1-\alpha$ is fixed and we show the coverage of asymptotic CIs as a function of the sample size n, as well as $\overline{n}_\alpha$ derived in Theorem 9.3. It happens that the coverage converges toward its nominal level for sample sizes around $\overline{n}_\alpha$, which supports $\overline{n}_\alpha$ as a rule of thumb of interest in practice[7].

[6] Equivalently, we could compare $\overline{\alpha}_n$ and α. As a rule of thumb, $\overline{\alpha}_n$ can be seen as the lowest α (hence the highest nominal level $1-\alpha$) for which the asymptotic CIs based on the delta method are reliable given the sample size n.

[7] This fact holds across various specifications (see additional simulations in Appendix 9.10).

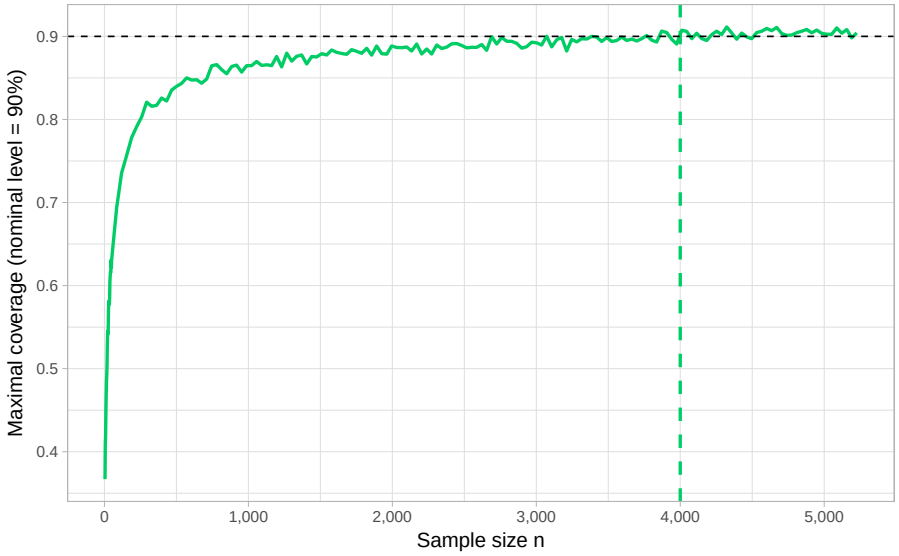


Figure 9.4: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}_2$ (bivariate Gaussian) with $\mathbb{E}[X] = 0.5$, $\mathbb{E}[Y] = 0.1$, $\mathbb{V}[X] = 1$, $\mathbb{V}[Y] = 2$, $\mathbb{C}orr(X, Y) = 0.5$. The nominal pointwise asymptotic level is set to 0.90. For a sample size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.1$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

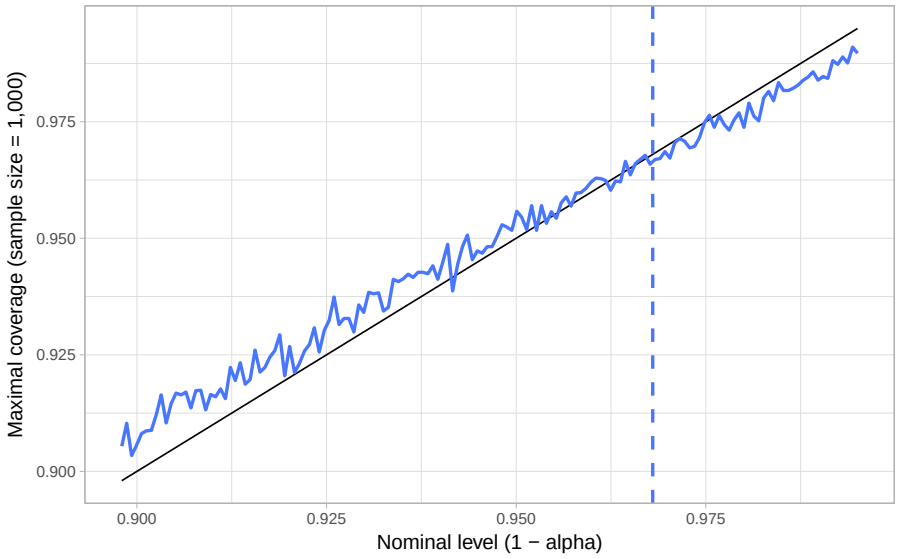


Figure 9.5: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb α_n. Specification: sample size $n = 1,000$, $P_{X,Y,n} = \mathcal{N}_2$ (bivariate Gaussian) with $\mathbb{E}[X] = 0.5$, $\mathbb{E}[Y] = 0.25$, $\mathbb{V}[X] = 2$, $\mathbb{V}[Y] = 1$, $\mathbb{C}orr(X, Y) = 0.5$. For each nominal level $1 - \alpha$ in the x-axis, we draw $10,000$ samples, compute the asymptotic CIs and see whether it covers or not the ratio of interest; we report the mean over the $10,000$ repetitions in the y-axis. The solid line is the first bisector $y = x$. The dashed vertical line shows $\alpha_n := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(n{l_{Y,n}}^2\right)$, setting here $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

In Figure 9.5, a sample size is fixed and we show the coverage[8] of asymptotic CIs for different nominal levels, as well as $\overline{\alpha}_n$. It is the converse of Figure 9.4.

The rule of thumb based on $\overline{\alpha}_n$ suggests that at nominal levels higher than $1-\overline{\alpha}_n$ (and for the sample size at hand) asymptotic CIs show some undercoverage. In the specification of Figure 9.5, $\overline{\alpha}_n$ turns out to fall close to the lowest α (hence highest $1-\alpha$) for which the coverage of asymptotic CIs attains their nominal level[9].

All in all, Figures 9.4 and 9.5 and additional simulations (see Appendix 9.10) advocate the use of the $\overline{n}_\alpha$ derived in Theorem 9.3 (or conversely $\overline{\alpha}_n$) as a rule of thumb to appraise the reliability of asymptotic CIs for ratios of expectations.

9.6.2 Application to real data

We illustrate the use of our confidence intervals with two applications on real data using French Labor Survey between 2010 and 2017.[10] Both applications resume our canonical example of conditional expectations and use $n = 204,246$ observations.

First, we compute the expected wage conditional on belonging to top wage brackets. Let W be a real random variable that denotes the wage of an employee. For a given top wage W_0, the parameter of interest is $\mathbb{E}[W \mid W \geq W_0]$, with $W_0 = q_{1-\tau}(W)$ the quantile at order $1-\tau$ of W. It can be written as a ratio with the variable in the numerator $X = W\mathbb{1}\{W \geq W_0\}$ and in the denominator $Y = \mathbb{1}\{W \geq W_0\}$. In what follows, we focus on the comparison between our nonasymptotic confidence intervals and the asymptotic CIs from the delta method. As explained in Section 9.5.2, we compute our confidence intervals using plug-in estimators to delimit the class of distributions we consider.

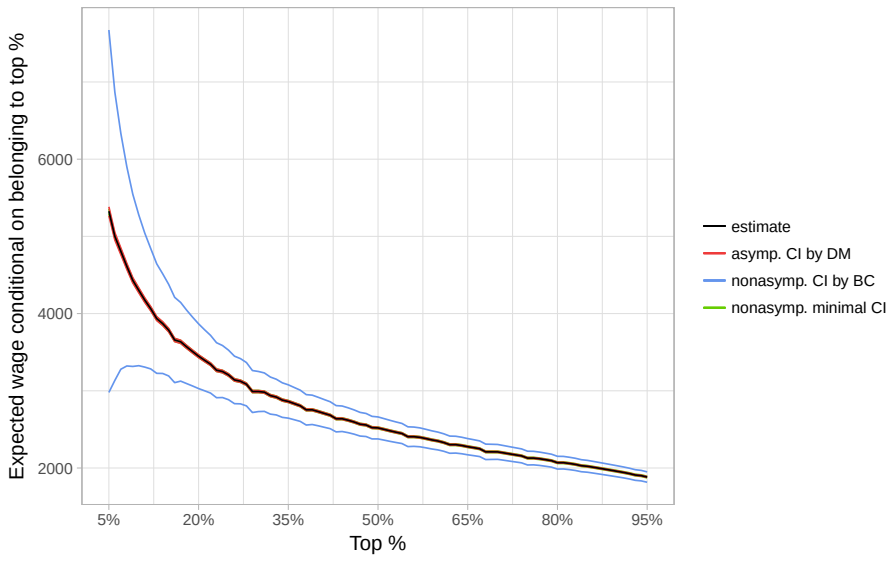


Figure 9.6: Expected wage conditional on belonging to top % as a function of τ.

Figure 9.6 compares the confidence intervals constructed using Theorem 9.3 with the confidence intervals constructed using the delta method. As expected, our nonasymptotic CIs are broader than

[8] The coverage shown is in fact an estimate obtained with Monte-Carlo simulations, namely the mean over a large number of repetitions, as usual in MC simulations.

[9] Again, this fact appears to hold for various specifications (cf. Appendix 9.10).

[10] The complete references of the databases are as follows: "*Enquête Emploi en continu (version FPR)*" - *year*, INSEE, ADISP, for *year* going through 2010 to 2017.

the ones from the delta method. Both CIs get narrower as the expectation in the denominator moves away from 0. Figure 9.6 also displays a plug-in estimate of nonasymptotic CIs based on the minimal half-length t^* presented in Theorem 9.6. As a reminder, no nonasymptotic CI of the form $\left[\overline{X}_n/\overline{Y}_n \pm t\right]$ can have good coverage if $t < t^*$. On Figure 9.6, we cannot distinguish CIs based on the delta method from nonasymptotic CIs with minimal half-length (see details below in Figures 9.7 and 9.8).

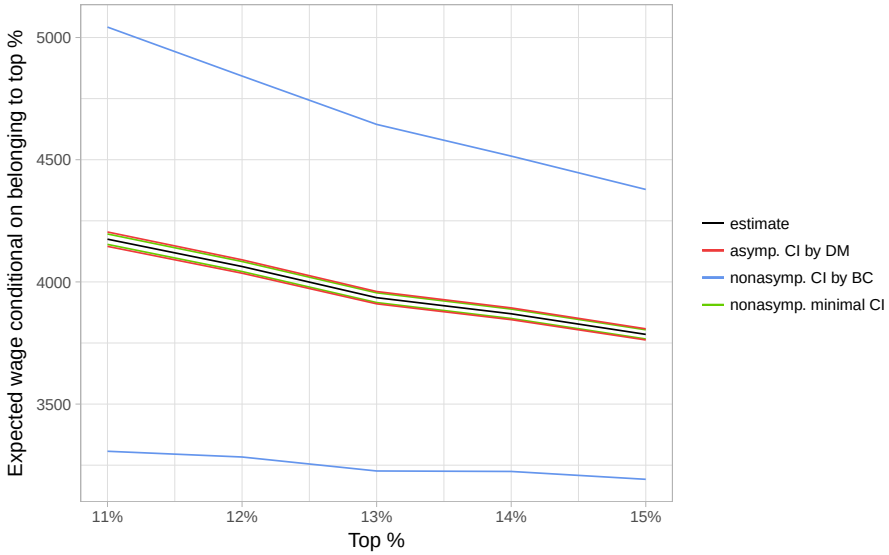


Figure 9.7: Expected wage conditional on belonging to top % as a function of τ (zoom).

Figures 9.7 and 9.8 zoom in subintervals of the previous graph in order to better compare the asymptotic confidence intervals from the delta method and the nonasymptotic CIs based on half-length t^* from Theorem 9.6. When focusing on the top of the distribution, the asymptotic CIs happen to be broader than the minimal one (Figure 9.7). On the contrary, with an expectation in the denominator further from 0, the length of the asymptotic CIs is lower than the minimal one (Figure 9.8).

Nevertheless, we do not have formal results to compare those lengths. At this stage, we can simply say that it is in some sense reassuring that the length of the asymptotic CIs increases relatively more than the length of the minimal one as we focus on thinner top percentage of the distribution.

Using the same data, our second application is an estimation of the proportion of women within the top brackets of the distribution of income. If G is an indicator variable equal to 1 for women and 0 for men, the conditional expectation of interest can be defined as $\mathbb{E}\left[G \mid W \geq q_{1-\tau}(W)\right]$. As in the first application, Figure 9.9 compares the different confidence intervals.

9.7 Conclusion

We provide an overview of the problem of constructing confidence intervals for a ratio of means in asymptotic and nonasymptotic frameworks. Using the delta method, asymptotic confidence intervals can be constructed but they have no coverage guarantee for a finite sample size. Nonasymptotic confidence intervals are proposed, but they depend on unknown parameters, which are functions of the data-generating process. To overcome such difficulties, we have proposed plug-in estimators based on a rule of thumb that allows the statistician to quantify whether the coverage of the delta-method should

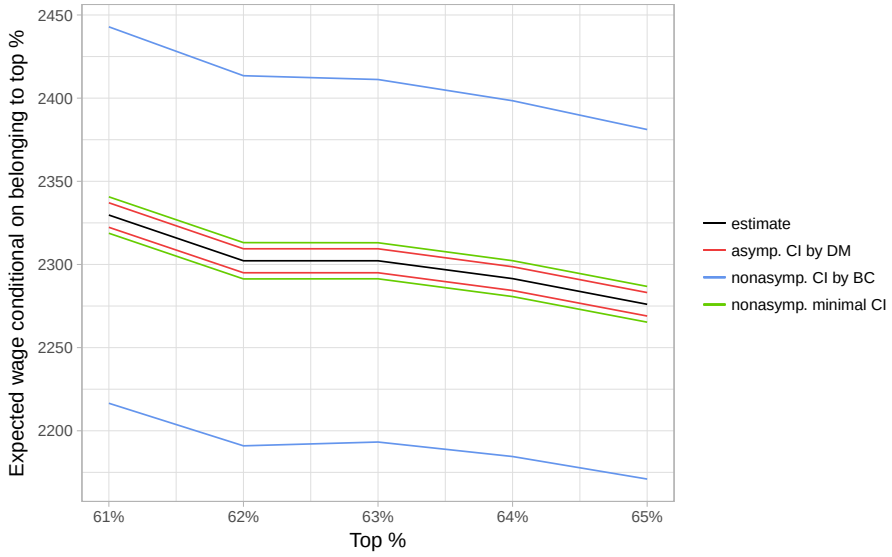


Figure 9.8: Expected wage conditional on belonging to top % (zoom)

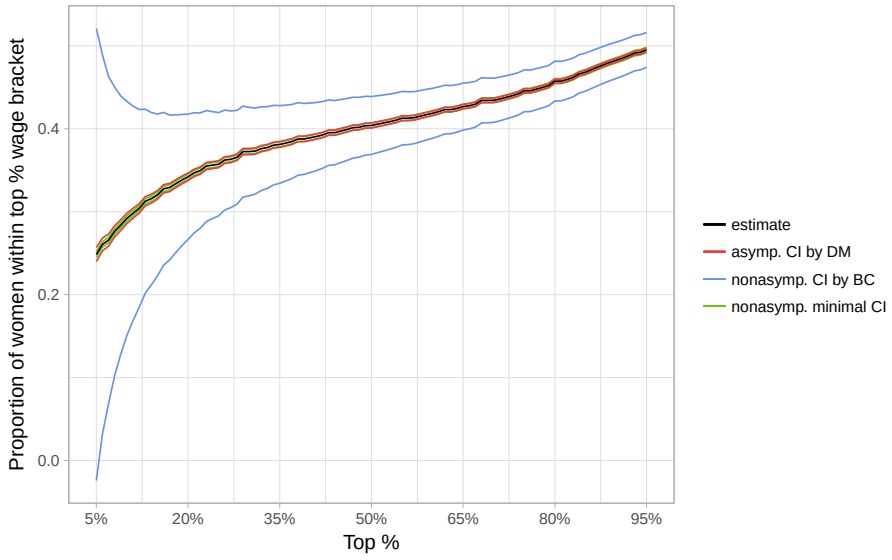


Figure 9.9: Proportion of women within top % wage bracket

be close to its nominal level (or not). Finally, we have illustrate our methods with simulated data, and we have applied it to a real dataset to construct confidence intervals for conditional expectations of wages.

9.8 Proofs of the results in Sections 9.3, 9.4 and 9.5

9.8.1 Proof of Theorem 9.1

Let $\theta_{X,n} := \mathbb{E}[X_{1,n}]$, $\theta_{Y,n} := \mathbb{E}[Y_{1,n}]$, $h_{X,n} := \sqrt{n}(\overline{X}_n - \theta_{X,n})$ and $h_{Y,n} := \sqrt{n}(\overline{Y}_n - \theta_{Y,n})$. We first rewrite Theorem 9.1 using this notation.

Theorem 9.7. *Let Assumption 9.2.1 hold. Assume that $\mathbb{E}[|X_{1,n}|^3]$ and $\mathbb{E}[|Y_{1,n}|^3]$ are bounded and $\mathbb{V}[(X_{1,n}, Y_{1,n})] \to V$ where V is a positive definite matrix and that $\mathbb{P}(\overline{Y}_n = 0) \to 0$, as $n \to \infty$.*

Then the sequence of random variables $A_n := \overline{X}_n/\overline{Y}_n - \theta_{X,n}/\theta_{Y,n}$ satisfies :

1. *If $n^{-1/2} = o(\theta_{Y,n})$, then A_n is equivalent to $n^{-1/2}\big(h_{X,n}/\theta_{Y,n} - h_{Y,n}\theta_{X,n}/\theta_{Y,n}^2\big)$, as $n \to \infty$.*
2. *If $\theta_{Y,n} = o(n^{-1/2})$, then A_n is equivalent to $h_{X,n}/h_{Y,n} - \theta_{X,n}/\theta_{Y,n}$.*
3. *If there exists a finite constant $C \neq 0$ such that $\sqrt{n}\theta_{Y,n} \to C$ as $n \to \infty$, then A_n is equivalent to $\sqrt{n}\theta_{X,n}\Big(1/(C + h_{Y,n}) - 1/C\Big) + h_{X,n}/(C + h_{Y,n})$.*

Let us define $W_n := \mathbb{1}\{\theta_{Y,n} + h_{Y,n}/\sqrt{n} = 0\}$ and remark that $W_n = 1$ whenever $\overline{Y}_n = 0$. By assumption $\mathbb{P}(\overline{Y}_n = 0) \to 0$, therefore $W_n \xrightarrow[n\to+\infty]{d} \delta_0$. Moreover, the positive-definiteness of V and the boundedness of $\mathbb{E}[|X_{1,n}|^3]$ and $\mathbb{E}[|Y_{1,n}|^3]$ ensure that the Cramer-Wold device applies and $(h_{Y,n}, h_{X,n}) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0, V)$. By Slutsky's Lemma, we also get $(h_{Y,n}, h_{X,n}, W_n) \xrightarrow[n\to+\infty]{d} \mathcal{N}(0, V) \otimes \delta_0$. We can therefore apply the almost sure representation theorem, see [138, Theorem 2.19]. It means that there exists a probability space $(\tilde{\Omega}, \tilde{\mathcal{U}}, \tilde{\mathbb{P}})$, a sequence of random vectors $(\tilde{h}_{Y,n}, \tilde{h}_{X,n}, \tilde{W}_n)$ such that for every $n \geq 1$, $(\tilde{h}_{Y,n}, \tilde{h}_{X,n}, \tilde{W}_n) \overset{d}{=} (h_{Y,n}, h_{X,n}, W_n)$, and a random vector $(\tilde{h}_{Y,\infty}, \tilde{h}_{X,\infty}, \tilde{W}_\infty)$ following the distribution $\mathcal{N}(0, V) \otimes \delta_0$ such that $(\tilde{h}_{Y,n}, \tilde{h}_{X,n}, \tilde{W}_n) \xrightarrow{a.s.} (\tilde{h}_{Y,\infty}, \tilde{h}_{X,\infty}, \tilde{W}_\infty)$, where the convergence is to be seen as of a sequence of random vectors defined on $(\tilde{\Omega}, \tilde{\mathcal{U}}, \tilde{\mathbb{P}})$. Let us define

$$\tilde{A}_n := \frac{\theta_{X,n} + \tilde{h}_{X,n}/\sqrt{n}}{\theta_{Y,n} + \tilde{h}_{Y,n}/\sqrt{n}} - \frac{\theta_{X,n}}{\theta_{Y,n}} \overset{d}{=} \frac{\theta_{X,n} + h_{X,n}/\sqrt{n}}{\theta_{Y,n} + h_{Y,n}/\sqrt{n}} - \frac{\theta_{X,n}}{\theta_{Y,n}} = \frac{\overline{X}_n}{\overline{Y}_n} - \frac{\theta_{X,n}}{\theta_{Y,n}} = A_n.$$

Moreover, we have $\tilde{W}_n = \mathbb{1}\{\theta_{Y,n} + \tilde{h}_{Y,n}/\sqrt{n} = 0\}$ and $\tilde{W}_\infty = 0$ almost surely. We can define

$$\tilde{\Omega}^* = \{\tilde{\omega} \in \tilde{\Omega} : \tilde{W}_n(\tilde{\omega}) \to 0 \text{ and } \exists N > 0, \forall n \geq N, \tilde{h}_{Y,n}(\tilde{\omega}) \neq 0\}.$$

By the almost sure convergence of $(\tilde{h}_{Y,n}, \tilde{W}_n)$, we get $\tilde{\mathbb{P}}(\tilde{\Omega}^*) = 1$, and for every $\tilde{\omega} \in \tilde{\Omega}^*$, $\tilde{W}_n(\tilde{\omega}) = 0$ and $\tilde{h}_{Y,n}(\tilde{\omega}) \neq 0$ for every n large enough. This means that for every given $\tilde{\omega} \in \tilde{\Omega}^*$, and for every n large enough, $\tilde{A}_n$ is well-defined. In the rest of the proof, we will fix such a $\tilde{\omega} \in \tilde{\Omega}^*$, so that all random variables may be considered as deterministic. By the almost sure representation theorem, this means that the equivalents and limits that will be obtained will still be valid in law in the original spaces Ω_n.

First case: We have

$$\begin{aligned}
\tilde{A}_n &= \frac{\overline{X}_n}{\overline{Y}_n} - \frac{\theta_{X,n}}{\theta_{Y,n}} = \frac{\theta_{X,n} + \tilde{h}_{X,n}/\sqrt{n}}{\theta_{Y,n} + \tilde{h}_{Y,n}/\sqrt{n}} - \frac{\theta_{X,n}}{\theta_{Y,n}} \\
&= \frac{\theta_{X,n} + \tilde{h}_{X,n}/\sqrt{n}}{\theta_{Y,n}}\left(1 - \frac{\tilde{h}_{Y,n}}{\sqrt{n}\theta_{Y,n}} + O\big((\sqrt{n}\theta_{Y,n})^{-2}\big)\right) - \frac{\theta_{X,n}}{\theta_{Y,n}} \\
&\sim \frac{-\theta_{X,n}\tilde{h}_{Y,n}}{\sqrt{n}\theta_{Y,n}^2} + \frac{\tilde{h}_{X,n}}{\sqrt{n}\theta_{Y,n}},
\end{aligned}$$

as claimed.

Second case: We have

$$\begin{aligned}\tilde{A}_n &= \frac{\theta_{X,n} + \tilde{h}_{X,n}/\sqrt{n}}{\theta_{Y,n} + \tilde{h}_{Y,n}/\sqrt{n}} - \frac{\theta_{X,n}}{\theta_{Y,n}} = \frac{\theta_{X,n} + \tilde{h}_{X,n}/\sqrt{n}}{\big(\tilde{h}_{Y,n} + o(1)\big)/\sqrt{n}} - \frac{\theta_{X,n}}{\theta_{Y,n}} \\ &= \frac{\sqrt{n}\theta_{X,n} + \tilde{h}_{X,n}}{\tilde{h}_{Y,n}} - \frac{\theta_{X,n}}{\theta_{Y,n}} + o\Big(\sqrt{n}\theta_{X,n} + 1\Big) \\ &\sim \theta_{X,n}\left(\frac{\sqrt{n}}{\tilde{h}_{Y,n}} - \frac{1}{\theta_{Y,n}}\right) + \frac{\tilde{h}_{X,n}}{\tilde{h}_{Y,n}},\end{aligned}$$

and the result follows from the fact that $\sqrt{n}/\tilde{h}_{Y,n}$ is negligible compared to $1/\theta_{Y,n}$.

Third case: We have

$$\tilde{A}_n \sim \frac{\theta_{X,n} + \tilde{h}_{X,n}/\sqrt{n}}{C/\sqrt{n} + \tilde{h}_{Y,n}/\sqrt{n}} - \frac{\theta_{X,n}}{C/\sqrt{n}} = \frac{\sqrt{n}\theta_{X,n} + \tilde{h}_{X,n}}{C + h_{Y,n}} - \frac{\sqrt{n}\theta_{X,n}}{C}.$$

We factorize by $\theta_{X,n}$ in the latter expression, which completes the proof.

□

9.8.2 Proof of Theorem 9.2

We fix arbitrary $n \in \mathbb{N}^*$ and $\varepsilon \in \mathbb{R}_+^*$. By the triangle inequality first, then using the bound $|\overline{X}_n| \leq |\overline{X}_n - \mathbb{E}[X_{1,n}]| + |\mathbb{E}[X_{1,n}]|$ and Assumptions 9.2.1 to 9.4.1 in the second inequality, we get:

$$\begin{aligned}\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}[X_{1,n}]}{\mathbb{E}[Y_{1,n}]}\right| &\leq |\overline{X}_n| \cdot \left|\frac{1}{\overline{Y}_n} - \frac{1}{\mathbb{E}[Y_{1,n}]}\right| + \frac{1}{\mathbb{E}[Y_{1,n}]}\left|\overline{X}_n - \mathbb{E}[X_{1,n}]\right| \\ &\leq \frac{\left(\left|\overline{X}_n - \mathbb{E}[X_{1,n}]\right| + \sqrt{u_{X,n}}\right)\left|\overline{Y}_n - \mathbb{E}[Y_{1,n}]\right|}{a_{Y,n} l_{Y,n}} + \frac{\left|\overline{X}_n - \mathbb{E}[X_{1,n}]\right|}{l_{Y,n}}.\end{aligned}$$

Consequently, the event considered in Theorem 9.2 is included in the event:

$$\frac{\left(\left|\overline{X}_n - \mathbb{E}[X_{1,n}]\right| + \sqrt{u_{X,n}}\right)\left|\overline{Y}_n - \mathbb{E}[Y_{1,n}]\right|}{a_{Y,n} l_{Y,n}} + \frac{\left|\overline{X}_n - \mathbb{E}[X_{1,n}]\right|}{l_{Y,n}} > \frac{\left(\varepsilon + \sqrt{u_{X,n}}\right)\varepsilon}{a_{Y,n} l_{Y,n}} + \frac{\varepsilon}{l_{Y,n}}. \tag{9.2}$$

If both $|\overline{X}_n - \mathbb{E}[X_{1,n}]|$ and $|\overline{Y}_n - \mathbb{E}[Y_{1,n}]|$ are inferior or equal to ε, event (9.2) cannot happen. By contraposition, we obtain:

$$\begin{aligned}&\mathbb{P}\left(\frac{\left(\left|\overline{X}_n - \mathbb{E}[X_{1,n}]\right| + \sqrt{u_{X,n}}\right)\left|\overline{Y}_n - \mathbb{E}[Y_{1,n}]\right|}{a_{Y,n} l_{Y,n}} + \frac{\left|\overline{X}_n - \mathbb{E}[X_{1,n}]\right|}{l_{Y,n}} > \frac{\left(\varepsilon + \sqrt{u_{X,n}}\right)\varepsilon}{a_{Y,n} l_{Y,n}} + \frac{\varepsilon}{l_{Y,n}}\right) \\ &\leq \mathbb{P}\left(\left\{\left|\overline{X}_n - \mathbb{E}[X_{1,n}]\right| > \varepsilon\right\} \cup \left\{\left|\overline{Y}_n - \mathbb{E}[Y_{1,n}]\right| > \varepsilon\right\}\right) \\ &\leq \mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[X_{1,n}]\right| > \varepsilon\right) + \mathbb{P}\left(\left|\overline{Y}_n - \mathbb{E}[Y_{1,n}]\right| > \varepsilon\right),\end{aligned}$$

where we use the union bound for the last inequality. The first conclusion follows from using twice Bienaymé-Chebyshev's inequality applied to the variables $\overline{X}_n$ and $\overline{Y}_n$ and the fact that under Assumptions 9.2.1 and 9.2.2 and Jensen's inequality, $\mathbb{V}[X_{1,n}] \leq u_{X,n}$ and $\mathbb{V}[Y_{1,n}] \leq u_{Y,n} - l_{Y,n}^2$. The second conclusion follows from solving $(u_{X,n} + u_{Y,n} - l_{Y,n}^2)/(n\varepsilon^2) = \alpha$.

□

9.8.3 Proof of Theorem 9.3

We start by introducing and proving an intermediate lemma that is also used to prove Theorem 9.12. For a random variable U, $\varepsilon > 0$, and $\tilde{\varepsilon} \in (0,1)$ we define the following events:

$$A_\varepsilon^U := \Big\{ \big|\overline{U}_n - \mathbb{E}[U]\big| \leq \varepsilon \Big\}, \text{ and } \tilde{A}_{\tilde{\varepsilon}}^U := \Big\{ \big|\overline{U}_n - \mathbb{E}[U]\big| \leq \tilde{\varepsilon} \big|\mathbb{E}[U]\big| \Big\}.$$

Lemma 9.8. *Assume that Assumption 9.2.1 holds. Then for every $n \in \mathbb{N}^*$, $\varepsilon > 0$ and $\tilde{\varepsilon} \in (0,1)$, we have*

$$\mathbb{P}\left(\left| \frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}\left[X_{1,n}\right]}{\mathbb{E}\left[Y_{1,n}\right]} \right| > \left(\frac{\left(\left|\mathbb{E}\left[X_{1,n}\right]\right| + \varepsilon\right)\tilde{\varepsilon}}{(1-\tilde{\varepsilon})^2} + \varepsilon \right) \frac{1}{\left|\mathbb{E}\left[Y_{1,n}\right]\right|} \right) \leq 1 - \mathbb{P}\left(A_\varepsilon^{X_{1,n}}\right) + 1 - \mathbb{P}\left(\tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}\right).$$

We use this lemma with different concentration inequalities, giving different speeds (namely Bienaymé-Chebychev or Hoeffding inequality).

Proof of Lemma 9.8: We fix arbitrary $\varepsilon > 0$ and $\tilde{\varepsilon} \in (0,1)$. Without loss of generality, we can assume that $\mathbb{E}\left[Y_{1,n}\right] > 0$ and $\mathbb{E}\left[X_{1,n}\right] \geq 0$.

First, using the union bound, note that the event $A_\varepsilon^{X_{1,n}} \cap \tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}$ holds with a probability bigger than $\mathbb{P}\left(A_\varepsilon^{X_{1,n}}\right) + \mathbb{P}\left(\tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}\right) - 1$. Hence, its complement is of probability lower than $1 - \mathbb{P}\left(A_\varepsilon^{X_{1,n}}\right) + 1 - \mathbb{P}\left(\tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}\right)$.

Second, we show that the event considered in Lemma 9.8 is included in the complement of $A_\varepsilon^{X_{1,n}} \cap \tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}$, which concludes the proof. To do so, we reason by contraposition and do the following computations on the event $A_\varepsilon^{X_{1,n}} \cap \tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}$.
By the triangle inequality, we get

$$\left| \frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}\left[X_{1,n}\right]}{\mathbb{E}\left[Y_{1,n}\right]} \right| \leq \left|\overline{X}_n\right| \cdot \left| \frac{1}{\overline{Y}_n} - \frac{1}{\mathbb{E}\left[Y_{1,n}\right]} \right| + \frac{1}{\mathbb{E}\left[Y_{1,n}\right]} \left| \overline{X}_n - \mathbb{E}\left[X_{1,n}\right] \right|.$$

We now bound the first term using the mean value theorem applied to the function $f(x) := 1/(x + \mathbb{E}\left[Y_{1,n}\right])$

$$\left| \frac{1}{\overline{Y}_n} - \frac{1}{\mathbb{E}\left[Y_{1,n}\right]} \right| = \left| f(\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]) - f(0) \right| \leq \frac{\left|\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]\right|}{(1-\tilde{\varepsilon})^2 \mathbb{E}\left[Y_{1,n}\right]^2} \leq \frac{\tilde{\varepsilon}\mathbb{E}\left[Y_{1,n}\right]}{(1-\tilde{\varepsilon})^2 \mathbb{E}\left[Y_{1,n}\right]^2},$$

where the first inequality uses that, on the event $\tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}$, a lower bound on $|x + \mathbb{E}\left[Y_{1,n}\right]|$ with x varying between 0 and $\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]$ is $(1-\tilde{\varepsilon})|\mathbb{E}\left[Y_{1,n}\right]|$. Therefore, on $A_\varepsilon^{X_{1,n}} \cap \tilde{A}_{\tilde{\varepsilon}}^{Y_{1,n}}$,

$$\begin{aligned}
\left| \frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}\left[X_{1,n}\right]}{\mathbb{E}\left[Y_{1,n}\right]} \right| &\leq \left|\overline{X}_n\right| \cdot \frac{\tilde{\varepsilon}\mathbb{E}\left[Y_{1,n}\right]}{(1-\tilde{\varepsilon})^2 \mathbb{E}\left[Y_{1,n}\right]^2} + \frac{\varepsilon}{\mathbb{E}\left[Y_{1,n}\right]} \\
&\leq \left(\left|\mathbb{E}\left[X_{1,n}\right]\right| + \left|\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\right| \right) \frac{\tilde{\varepsilon}}{(1-\tilde{\varepsilon})^2 \mathbb{E}\left[Y_{1,n}\right]} + \frac{\varepsilon}{\mathbb{E}\left[Y_{1,n}\right]} \\
&\leq \frac{\left(\left|\mathbb{E}\left[X_{1,n}\right]\right| + \varepsilon\right)\tilde{\varepsilon}}{(1-\tilde{\varepsilon})^2 \mathbb{E}\left[Y_{1,n}\right]} + \frac{\varepsilon}{\mathbb{E}\left[Y_{1,n}\right]},
\end{aligned}$$

where we use the triangle inequality to get the second line. It is indeed the complement of the event considered in the statement of Lemma 9.8.

□

Proof of Theorem 9.3:

We fix arbitrary $n \in \mathbb{N}^*$, $\varepsilon > 0$ and $\tilde{\varepsilon} \in (0,1)$. By Lemma 9.8, we have

$$\begin{aligned}
&\mathbb{P}\left(\left| \frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}\left[X_{1,n}\right]}{\mathbb{E}\left[Y_{1,n}\right]} \right| > \left(\frac{\left(\left|\mathbb{E}\left[X_{1,n}\right]\right| + \varepsilon\right)\tilde{\varepsilon}}{(1-\tilde{\varepsilon})^2} + \varepsilon \right) \frac{1}{\left|\mathbb{E}\left[Y_{1,n}\right]\right|} \right) \\
&\leq 1 - \mathbb{P}\Big(\big|\overline{X}_n - \mathbb{E}\left[X_{1,n}\right]\big| \leq \varepsilon \Big) + 1 - \mathbb{P}\Big(\big|\overline{Y}_n - \mathbb{E}\left[Y_{1,n}\right]\big| \leq \tilde{\varepsilon} \big|\mathbb{E}\left[Y_{1,n}\right]\big| \Big).
\end{aligned}$$

Using Jensen's inequality and Assumption 9.2.2, we have $|\mathbb{E}[X_{1,n}]| \leq (u_{X,n})^{1/2}$, and Assumption 9.2.1 entails $1/|\mathbb{E}[Y_{1,n}]| \leq 1/l_{Y,n}$. Consequently, we get

$$\mathbb{P}\left(\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}[X_{1,n}]}{\mathbb{E}[Y_{1,n}]}\right| > \left(\frac{\left(\sqrt{u_{X,n}}+\varepsilon\right)\tilde{\varepsilon}}{(1-\tilde{\varepsilon})^2} + \varepsilon\right)\frac{1}{l_{Y,n}}\right)$$
$$\leq 1 - \mathbb{P}\Big(\big|\overline{X}_n - \mathbb{E}[X_{1,n}]\big| \leq \varepsilon\Big) + 1 - \mathbb{P}\Big(\big|\overline{Y}_n - \mathbb{E}[Y_{1,n}]\big| \leq \tilde{\varepsilon}\big|\mathbb{E}[Y_{1,n}]\big|\Big).$$

Using Bienaymé-Chebyshev's inequality twice gives the bounds

$$1 - \mathbb{P}\Big(\big|\overline{X}_n - \mathbb{E}[X_{1,n}]\big| \leq \varepsilon\Big) \leq \frac{\mathbb{V}[X_{1,n}]}{n\varepsilon^2}$$
$$1 - \mathbb{P}\Big(\big|\overline{Y}_n - \mathbb{E}[Y_{1,n}]\big| \leq \tilde{\varepsilon}\big|\mathbb{E}[Y_{1,n}]\big|\Big) \leq \frac{\mathbb{V}[Y_{1,n}]}{n\tilde{\varepsilon}^2\left(\mathbb{E}[Y_{1,n}]\right)^2}.$$

For the numerator, $\mathbb{V}[X_{1,n}] = \mathbb{E}\left[X_{1,n}^2\right] - \left(\mathbb{E}[X_{1,n}]\right)^2 \leq \mathbb{E}\left[X_{1,n}^2\right] \leq u_{X,n}$ using Assumption 9.2.2. For the denominator, Assumption 9.2.1 immediately entails that $1/l_{Y,n}^2$ is an upper bound on $1/\left(\mathbb{E}[Y_{1,n}]\right)^2$ and $l_{Y,n}^2$ a lower bound on $\left(\mathbb{E}[Y_{1,n}]\right)^2$. Therefore

$$\frac{\mathbb{V}[Y_{1,n}]}{n\tilde{\varepsilon}^2\left(\mathbb{E}[Y_{1,n}]\right)^2} \leq \frac{\mathbb{E}\left[Y_{1,n}^2\right] - {l_{Y,n}}^2}{n\tilde{\varepsilon}^2{l_{Y,n}}^2} \leq \frac{{u_{Y,n}} - {l_{Y,n}}^2}{n\tilde{\varepsilon}^2{l_{Y,n}}^2},$$

where the second inequality uses Assumption 9.2.2.

Combining the two bounds yields the following upper bound on the probability considered in Theorem 9.3

$$\frac{u_{X,n}}{n\varepsilon^2} + \frac{{u_{Y,n}} - {l_{Y,n}}^2}{n\tilde{\varepsilon}^2{l_{Y,n}}^2}, \tag{9.3}$$

as claimed.

For the second part of Theorem 9.3, for a fixed α, we equalize each of the two terms in (9.3) to $\alpha/2$ and solve for ε and $\tilde{\varepsilon}$, which yields:

$$\varepsilon^2 = \frac{2u_{X,n}}{n\alpha} \text{ and } \tilde{\varepsilon}^2 = \frac{2\left({u_{Y,n}} - {l_{Y,n}}^2\right)}{n\alpha l_{Y,n}^2}.$$

The bound $\overline{\alpha}_n$ comes from the fact that $\tilde{\varepsilon}$ needs to be smaller than 1.

□

9.8.4 Proof of Theorem 9.6

To prove Theorem 9.6, we need the following lemma.

Lemma 9.9. *For every integer $n \geq 7$ and every $x \in (0,1)$, $x\left(1-x/n\right)^{n-1} \geq x/3$.*

Proof of Lemma 9.9: we write the lemma to be applied directly in the demonstration of Theorem 9.6, but the inequality is equivalent to $(1-x/n)^{n-1} \geq 1/3$. Under our assumptions on n and x, $\ln(1-x/n)$ is well-defined. Using Taylor-Lagrange formula on the function $[0,x] \ni t \mapsto \ln(1-t/n)$ yields:

$$\left(1-\frac{x}{n}\right)^{n-1} = \exp\left((n-1)\ln\left(1-\frac{x}{n}\right)\right) = \exp\left(-(n-1)\left(\frac{x}{n} + \frac{1}{2\left(1-\tau x/n\right)^2}\frac{x^2}{n^2}\right)\right)$$

for some $\tau \in (0,1)$. Using the fact that $\frac{n-1}{n} \leq 1$, $x \leq 1$ and $\frac{1}{2(1-\tau x/n)^2} \leq \frac{1}{2(1-n^{-1})^2}$, we get that under our assumptions $\left(1-\frac{x}{n}\right)^{n-1} \geq \exp\left(-\left(1+\frac{1}{2n(1-n^{-1})^2}\right)\right)$. This bound is actually valid for every $x \in (0,1)$ and every $n \in \mathbb{N}^*$. The computation of $\exp\left(-\left(1+\frac{1}{2n(1-n^{-1})^2}\right)\right)$ shows that the latter is larger than $1/4$ whenever $n \geq 3$ and larger than $1/3$ whenever $n \geq 7$.

□

Proof of Theorem 9.6: we start using arguments developed in the proof of [26][Proposition 6.2]. We detail those for the sake of clarity. For every $n \in \mathbb{N}^*$ and $\eta > \sqrt{u_{X,n}}/n$, let us define the following distribution on $\mathbb{R}$, which will be used for the variable in the numerator[11]:

$$P_{n,u_{X,n},\eta} := \frac{u_{X,n}}{2n^2\eta^2}\delta_{\{-n\eta\}} + \left(1 - \frac{u_{X,n}}{n^2\eta^2}\right)\delta_{\{0\}} + \frac{u_{X,n}}{2n^2\eta^2}\delta_{\{n\eta\}}.$$

This distribution is symmetric, centered and has variance $u_{X,n}$. As shown in [26], every i.i.d. sample $(X_{i,n})_{i=1}^n$ drawn from $P_{n,u_{X,n},\eta}$ satisfies

$$\begin{aligned}&\mathbb{P}\left(\overline{X}_n \leq -\eta\right) = \mathbb{P}\left(\overline{X}_n \geq \eta\right) \geq \mathbb{P}\left(\overline{X}_n = \eta\right)\\ &\geq \sum_{i=1}^n \mathbb{P}\left(X_{i,n} = n\eta, X_{j,n} = 0,\ \forall j \neq i\right) = \frac{u_{X,n}}{2n\eta^2}\left(1 - \frac{u_{X,n}}{\eta^2 n^2}\right)^{n-1}.\end{aligned}$$

Note further that for every integer $n \geq 2$, the inequality $\mathbb{P}\left(\overline{X}_n \geq \eta\right) \geq \mathbb{P}\left(\overline{X}_n = \eta\right)$ becomes strict and for every $\xi \in (0,1)$ $\left\{\left|\overline{X}_n\right| \geq \eta\right\} \subseteq \left\{\left|\overline{X}_n\right| > \xi\eta\right\}$. As a result, if $(X_{i,n})_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{n,u_{X,n},\eta}$, for every $\eta > 0$, we have

$$\mathbb{P}\left(\left|\overline{X}_n\right| > \xi\eta\right) > \frac{u_{X,n}}{n\eta^2}\left(1 - \frac{u_{X,n}}{\eta^2 n^2}\right)^{n-1}. \tag{9.4}$$

The following steps do not show up in [26] since they are specific to controlling ratios of expectations and sample averages. For every $n \in \mathbb{N}^*$, let us define the following distribution on $\mathbb{R}$, which will be used for the variable in the denominator

$$P_{n,l_{Y,n},u_{Y,n}} := \frac{1}{2}\delta_{\left\{l_{Y,n} - \sqrt{u_{Y,n} - l_{Y,n}^2}\right\}} + \frac{1}{2}\delta_{\left\{l_{Y,n} + \sqrt{u_{Y,n} - l_{Y,n}^2}\right\}}.$$

Let $(X_{i,n}, Y_{i,n})_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_n := P_{n,u_{X,n},\eta} \otimes P_{n,l_{Y,n},u_{Y,n}}$. Observe that $\mathbb{E}\left[Y_{1,n}\right] = l_{Y,n}$ and $\mathbb{E}\left[Y_{1,n}^2\right] = u_{Y,n}$. Furthermore, $\left|\overline{Y}_n\right| \leq l_{Y,n} + \sqrt{u_{Y,n} - l_{Y,n}^2}$ almost surely. This implies that for every $\eta > 0$ and $\xi \in (0,1)$, the following holds

$$\left\{\left|\overline{X}_n\right| > \left(l_{Y,n} + \sqrt{u_{Y,n} - l_{Y,n}^2}\right)\xi\eta\right\} \subseteq \left\{\left|\frac{\overline{X}_n}{\overline{Y}_n}\right| > \xi\eta\right\}.$$

For fixed $n \geq 7$ and $\alpha \in \left(0, 1 \wedge n/\left(l_{Y,n} + \sqrt{u_{Y,n} - l_{Y,n}^2}\right)^2\right)$, we choose $\eta = \eta(\alpha) = \sqrt{v_n/3n\alpha}$. Combining the above inclusion with (9.4), and Lemma 9.9 (with the choice $x = 3\alpha$), we conclude that there exists a distribution on $\mathbb{R}^2$, namely P_n, that fulfills Assumptions 9.2.1 and 9.2.2 such that

$$\mathbb{P}\left(\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}\left[X_{1,n}\right]}{\mathbb{E}\left[Y_{1,n}\right]}\right| > \xi\sqrt{\frac{v_n}{3n\alpha}}\right) > \alpha,$$

which completes the proof.

□

9.8.5 Proof of Theorem 9.5

By Lemma 9.10, for every $\xi < 1 \wedge \left(u_{Y,n}/l_{Y,n}^2 - 1\right)$, there exists a distribution $P_{n,\xi}$ such that $\mathbb{P}\left(\overline{Y}_n = 0\right) \geq \tilde{\alpha}_n(\xi)$. Taking the supremum over ξ, we deduce that

$$\sup_{P_n \in \mathcal{P}} \mathbb{P}\left(\overline{Y}_n = 0\right) \geq \sup_{\xi} \tilde{\alpha}_n(\xi) = \underline{\alpha}_n.$$

Using the assumption that I_n is undefined whenever $\overline{X}_n/\overline{Y}_n$ is undefined (which is equivalent to $\overline{Y}_n = 0$), we deduce that $\mathbb{P}\big(I_n\, undefined\big) \geq \underline{\alpha}_n$.

[11] The notation δ denotes the Dirac distribution.

Lemma 9.10. *For each ξ in the interval $\left(0, 1 \wedge \left(u_{Y,n}/l_{Y,n}^2 - 1\right)\right)$, there exists a distribution $P_{n,\xi} \in \mathcal{P}$ such that $\mathbb{P}\left(\overline{Y}_n = 0\right) \geq \tilde{\alpha}_n$, where $\tilde{\alpha}_n := \left(1 - (1+\xi) l_{Y,n}^2/u_{Y,n}\right)^n$.*

Note that the interval $\left(0, 1 \wedge \left(u_{Y,n}/l_{Y,n}^2 - 1\right)\right)$ is not empty since we have assumed $u_{Y,n}/l_{Y,n}^2 > 1$.

Proof of Lemma 9.10: We consider the following distribution on $\mathbb{R}$

$$P_{n,l_{Y,n},u_{Y,n},c,\xi} := \left(\frac{c}{n}\right)^{1/n} \delta_{\{0\}} + \frac{1}{2}\left(1 - \left(\frac{c}{n}\right)^{1/n}\right)\delta_{\{y_{c-}\}} + \frac{1}{2}\left(1 - \left(\frac{c}{n}\right)^{1/n}\right)\delta_{\{y_{c+}\}},$$

where $c \in (0, n)$ is some constant to be chosen later, $y_{c-} := l_{Y,n}(1 - \sqrt{\xi})/(1 - (c/n)^{1/n})$ and $y_{c+} := l_{Y,n}(1 + \sqrt{\xi})/(1 - (c/n)^{1/n})$. Let $Y_{1,n} \sim P_{n,l_{Y,n},u_{Y,n},c,\xi_n}$. Observe that $\mathbb{E}[Y_{1,n}] = l_{Y,n}$ and $\mathbb{E}\left[Y_{1,n}^2\right] = l_{Y,n}^2(1+\xi_n)/\left(1 - (c/n)^{1/n}\right)$. With the choice

$$c = c_n := n\left(1 - \frac{l_{Y,n}^2}{u_{Y,n}}(1+\xi)\right)^n,$$

we have $\mathbb{E}\left[Y_{1,n}^2\right] = u_{Y,n}$. Note that c_n is strictly positive, because $1 - \frac{l_{Y,n}^2}{u_{Y,n}}(1+\xi_n) > 0$. This is equivalent to $u_{Y,n}/l_{Y,n}^2 > 1 + \xi_n$, which is true by assumption.

Consider now the following product measure on $\mathbb{R}^2$ defined by $P_n := \delta_{\left\{\sqrt{u_{X,n}}\right\}} \otimes P_{n,l_{Y,n},u_{Y,n},c_n,\xi}$. Let $(X_{i,n}, Y_{i,n})_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_n$. These random vectors satisfy $\mathbb{E}\left[X_{1,n}^2\right] = u_{X,n}$, $\mathbb{E}[Y_{1,n}] = l_{Y,n}$ and $\mathbb{E}\left[Y_{1,n}^2\right] = u_{Y,n}$. The next step is to build a lower bound on the event $\{\overline{Y}_n = 0\}$.

The assumption that $(X_{i,n}, Y_{i,n})_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_n$ and the construction of $P_{n,l_{Y,n},u_{Y,n},c_n,\xi}$ imply that

$$\mathbb{P}\left(\overline{Y}_n = 0\right) = \frac{c_n}{n} = \left(1 - \frac{l_{Y,n}^2}{u_{Y,n}}(1+\xi)\right)^n = \tilde{\alpha}_n$$

□

9.9 Adapted results for Hoeffding framework

Assumption 9.9.1. *For every $n \in \mathbb{N}^*$, there exist finite constants $a_{X,n}, b_{X,n}, a_{Y,n}, b_{Y,n}$ such that $X_{1,n}$ (respectively $Y_{1,n}$) lies $P_{X,Y,n}$-almost surely in the interval $[a_{X,n}, b_{X,n}]$ (resp. $[a_{Y,n}, b_{Y,n}]$).*

The support of $X_{1,n}$ and $Y_{1,n}$ is allowed to change with n, even though in many examples of interest, the former can be chosen independent from n. Assumptions 9.2.1 and 9.9.1 together correspond to the *Hoeffding case* because under these two assumptions, we can use the Hoeffding inequality to build nonasymptotic CIs.

9.9.1 Concentration inequality in the easy case

Assumption 9.9.2. *For every $n \in \mathbb{N}^*$, the lower bound $a_{Y,n}$ is strictly positive.*

Theorem 9.11. *Let $u_{X,n} := (b_{X,n} - a_{X,n})^2$ and $u_{Y,n} := (b_{Y,n} - a_{Y,n})^2$. Under Assumptions 9.2.1, 9.9.1 and 9.9.2, we have for every $n \in \mathbb{N}^*$ and $\varepsilon \in \mathbb{R}_+^*$*

$$\mathbb{P}\left(\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}[X_{1,n}]}{\mathbb{E}[Y_{1,n}]}\right| > \frac{\varepsilon}{l_{Y,n}}\left\{1 + \frac{1}{a_{Y,n}}\left(|a_{X,n}| \vee |b_{X,n}| + \varepsilon\right)\right\}\right) \leq 4\exp\left(-\frac{2n\varepsilon^2}{u_{X,n} \vee u_{Y,n}}\right).$$

As a consequence, $\mathbb{P}\left(\left|\overline{X}_n/\overline{Y}_n - \mathbb{E}\left[X_{1,n}\right]/\mathbb{E}\left[Y_{1,n}\right]\right| > t\right) \leq \alpha$*, with the choice*

$$t := \frac{1}{l_{Y,n}}\sqrt{\frac{(u_{X,n} \vee u_{Y,n})\ln(4/\alpha)}{2n}}\left(1 + \frac{1}{a_{Y,n}}\left(|a_{X,n}| \vee |b_{X,n}| + \sqrt{\frac{(u_{X,n} \vee u_{Y,n})\ln(4/\alpha)}{2n}}\right)\right),$$

for every $\alpha \in (0,1)$*.*

The theorem shows that it is possible to construct nonasymptotic CIs for ratios of expectations at every confidence level that are almost surely bounded. However, it requires the additional Assumption 9.9.2, that in particular does not allow for binary $\{0,1\}$ random variables in the denominator which may limit its applicability for various applications. In Section 9.9.2, we give an analogous result that only requires Assumptions 9.2.1 and 9.9.1 to hold, so that it encompasses the case of $\{0,1\}$-valued denominators. However, the cost to pay will be an upper bound on the achievable coverage of the confidence intervals.

9.9.2 Concentration inequality in the general case

We seek to build nontrivial nonasymptotic CIs under Assumptions 9.2.1 and 9.9.1 only. Under Assumption 9.2.1, $\mathbb{E}\left[Y_{1,n}\right] \neq 0$, so that there is no issue in considering the fraction $\mathbb{E}\left[X_{1,n}\right]/\mathbb{E}\left[Y_{1,n}\right]$. However, without Assumption 9.9.2, $\left\{\overline{Y}_n = 0\right\}$ has positive probability in general so that $\overline{X}_n/\overline{Y}_n$ is well-defined with probability less than one and undefined else. Note that when $P_{Y,n}$ is continuous wrt to Lebesgue's measure, there is no issue in defining $\overline{X}_n/\overline{Y}_n$ anymore since the event $\left\{\overline{Y}_n = 0\right\}$ has probability zero. This is not an easier case to establish concentration inequalities though, since without more restrictions, $\overline{Y}_n$ can still be arbitrarily close to zero with positive probability.

Theorem 9.12. *Assume that Assumptions 9.2.1 and 9.9.1 hold. For every* $n \in \mathbb{N}^*$*,* $\varepsilon > 0, \tilde{\varepsilon} \in (0,1)$*, we have*

$$\mathbb{P}\left(\left|\frac{\overline{X}_n}{\overline{Y}_n} - \frac{\mathbb{E}\left[X_{1,n}\right]}{\mathbb{E}\left[Y_{1,n}\right]}\right| > \left(\frac{(|a_{X,n}| \vee |b_{X,n}| + \varepsilon)\tilde{\varepsilon}}{(1-\tilde{\varepsilon})^2} + \varepsilon\right)\frac{1}{l_{Y,n}}\right) \leq 2\exp(-n\varepsilon^2\gamma\left(X_{1,n}\right)) + 2\exp(-n\tilde{\varepsilon}^2\gamma\left(Y_{1,n}\right)),$$

where $\gamma\left(X_{1,n}\right) = 2/(b_{X,n} - a_{X,n})^2$ *and* $\gamma\left(Y_{1,n}\right) = 2l_{Y,n}^2/(b_{Y,n} - a_{Y,n})^2$*.*

As a consequence, $\mathbb{P}\left(\left|\overline{X}_n/\overline{Y}_n - \mathbb{E}\left[X_{1,n}\right]/\mathbb{E}\left[Y_{1,n}\right]\right| > t\right) \leq \alpha$*, with the choice*

$$t := \sqrt{\frac{\ln(4/\alpha)}{n\gamma\left(X_{1,n}\right) \wedge \gamma\left(Y_{1,n}\right)}}\left(\frac{|a_{X,n}| \vee |b_{X,n}| + \sqrt{\ln(4/\alpha)/n\gamma\left(X_{1,n}\right)}}{(1 - \sqrt{\ln(4/\alpha)/n\gamma\left(Y_{1,n}\right)})^2} + 1\right)\frac{1}{l_{Y,n}},$$

for every $\alpha > \overline{\alpha}_{n,H} := 4e^{-n\gamma(Y_{1,n})}$*.*[12]

This theorem is proven in Section 9.9.4. It states that when $l_{Y,n} > 0$, it is possible to build valid nonasymptotic CIs with finite length up to the confidence level $1 - \overline{\alpha}_{n,H}$. This is a more positive result than [46] which claims that it is not possible to build nontrivial nonasymptotic CIs when $l_{Y,n}$ is taken equal to 0, no matter the confidence level. On the other hand, it is more negative than Theorem 9.11. Note that Theorem 9.12 is not an impossibility theorem since it only claims that considering confidence levels smaller than $1 - \overline{\alpha}_{n,H}$ is *sufficient* to build nontrivial CIs under Assumptions 9.2.1 and 9.9.1. The remaining question is to find out whether it is *necessary* to focus on confidence levels that do not exceed a certain threshold under Assumptions 9.2.1 and 9.9.1. We answer this in Section 9.9.3.

Theorem 9.12 has two other interesting consequences: for every confidence level up to $1 - \overline{\alpha}_{n,H}$, a nonasymptotic CI of the form $\left[\overline{X}_n/\overline{Y}_n \pm \tilde{t}\right]$ with $\tilde{t} > t$ has good coverage but is too conservative. What is more, if the DGP does not depend on n (*i.e* in the standard i.i.d. setup), for every fixed $\alpha > \overline{\alpha}_{n,H}$, the length of the confidence interval shrinks at the optimal rate $1/\sqrt{n}$.

[12] Equivalently, it means that for a given level α, the choice of t is valid for every integer $n > \overline{n}_\alpha := \ln(4/\alpha)/\gamma\left(Y_{1,n}\right)$.

9.9.3 An upper bound on testable confidence levels

Theorem 9.13. *For every $n \in \mathbb{N}^*$, and every $\alpha \in \left(0, \underline{\alpha}_{n,H}\right)$, where $\underline{\alpha}_{n,H} := \left(1 - l_{Y,n}/(b_{Y,n} - a_{Y,n})\right)^n$, if $(b_{Y,n} - a_{Y,n})/l_{Y,n} > 1$, there is no finite $t > 0$ such that $\left[\overline{X}_n/\overline{Y}_n \pm t\right]$ has coverage $1 - \alpha$ over $\mathcal{P}_H$, where $\mathcal{P}_H$ is the class of all distributions satisfying Assumptions 9.2.1 and 9.9.1 for a fixed lower bound $l_{Y,n}$ and fixed lengths $b_{X,n} - a_{X,n}$ and $b_{Y,n} - a_{Y,n}$.*

This theorem asserts that confidence intervals of the form $\left[\overline{X}_n/\overline{Y}_n \pm t\right]$ with coverage higher than $1 - \underline{\alpha}_{n,H}$ under Assumptions 9.2.1 and 9.9.1 are not defined (or are of infinite length) with positive probability for at least one distribution in $\mathcal{P}_H$. The additional restriction $(b_{Y,n} - a_{Y,n})/l_{Y,n} > 1$ is rather mild in practice: it is equivalent to $b_{Y,n} - a_{Y,n} > l_{Y,n}$ and is satisfied as soon as $a_{Y,n} \leq 0$ and $b_{Y,n} > l_{Y,n} > 0$. This encompasses all DGPs where the denominator is $\{0,1\}$-valued and the probability that the denominator equals 1 is bounded from below by $l_{Y,n} \in (0,1)$.

Note that for Theorems 9.11 and 9.12, it is required to know not only the length $b_{X,n} - a_{X,n}$ but also the actual endpoints of the support, $a_{X,n}$ and $b_{X,n}$. On the contrary, Theorem 9.13 does not require the latter. In that respect, the class of Theorem 9.13 is larger than the one of the two preceding theorems.

9.9.4 Proof of Theorems 9.11 and 9.12

The proofs are identical to those of Theorems 9.2 and 9.3, except for the Bienaymé-Chebyshev inequality that has to be replaced with the Hoeffding inequality. The latter can be used under Assumption 9.9.1. Note also that $\mathbb{E}\left[X_{1,n}\right]$ is now bounded by $|a_{X,n}| \vee |b_{X,n}|$.

9.9.5 Proof of Theorem 9.13

By Lemma 9.14, for every $\xi < 1 \wedge \left((b_{Y,n} - a_{Y,n})/l_{Y,n} - 1\right)$, there exists a distribution $P_{n,\xi} \in \mathcal{P}_H$ satisfying Assumptions 9.2.1 and 9.9.1 such that $\mathbb{P}\left(\overline{Y}_n = 0\right) \geq \tilde{\alpha}_{n,H}(\xi)$. Denote its marginal distributions by $P_{X,n,\xi}$ and $P_{Y,n,\xi}$. Therefore, $P_{n,\xi}$ satisfies Assumptions 9.2.1 and 9.9.1, and $\overline{X}_n/\overline{Y}_n$ is undefined with probability greater than $\tilde{\alpha}_{n,H}(\xi)$. Taking the supremum over ξ, we deduce that

$$\sup_{P_n \in \mathcal{P}_H} \mathbb{P}\left(\overline{Y}_n = 0\right) \geq \sup_{\xi} \tilde{\alpha}_n(\xi) = \underline{\alpha}_{n,H}.$$

This means that the random interval $I_n^* := \left[\overline{X}_n/\overline{Y}_n \pm t\right]$ cannot have coverage higher than $1 - \underline{\alpha}_{n,H}$ since it may be undefined with a probability higher than $\underline{\alpha}_{n,H}$.

□

Lemma 9.14. *For each ξ in the interval $\left(0, 1 \wedge \left((b_{Y,n} - a_{Y,n})/l_{Y,n} - 1\right)\right)$, there exists a distribution $P_{n,\xi} \in \mathcal{P}_H$ such that $\mathbb{P}\left(\overline{Y}_n = 0\right) \geq \tilde{\alpha}_{n,H}$, where $\tilde{\alpha}_{n,H} := \left(1 - (1+\xi) l_{Y,n}/(b_{Y,n} - a_{Y,n})\right)^n$.*

Note that the interval $\left(0, 1 \wedge \left((b_{Y,n} - a_{Y,n})/l_{Y,n} - 1\right)\right)$ is not empty since we have assumed $(b_{Y,n} - a_{Y,n})/l_{Y,n} > 1$.

Proof of Lemma 9.14: We consider the following distribution on $\mathbb{R}$

$$P_{n,l_{Y,n},c,\xi} := \left(\frac{c}{n}\right)^{1/n} \delta_{\{0\}} + \frac{1}{2}\left(1 - \left(\frac{c}{n}\right)^{1/n}\right) \delta_{\{y_{c-}\}} + \frac{1}{2}\left(1 - \left(\frac{c}{n}\right)^{1/n}\right) \delta_{\{y_{c+}\}},$$

where $c \in (0,n)$ is some constant to be chosen later, $y_{c-} := l_{Y,n}(1-\xi)/(1-(c/n)^{1/n})$ and $y_{c+} := l_{Y,n}(1+\xi)/(1-(c/n)^{1/n})$. Let $Y_{1,n} \sim P_{n,l_{Y,n},c,\xi_n}$. Observe that $\mathbb{E}\left[Y_{1,n}\right] = l_{Y,n}$. With the choice

$$c = c_n := n\left(1 - \frac{l_{Y,n}}{b_{Y,n} - a_{Y,n}}(1+\xi)\right)^n,$$

we have $y_{c+} = b_{Y,n} - a_{Y,n}$. Note that c_n is strictly positive, because $1 - \frac{l_{Y,n}}{b_{Y,n}-a_{Y,n}}(1+\xi_n) > 0$. This is equivalent to $b_{Y,n} - a_{Y,n}/l_{Y,n} > 1 + \xi_n$, which is true by assumption.

Consider now the following product measure on $\mathbb{R}^2$ defined by $P_n := \left(0.5\delta_{\{0\}} + 0.5\delta_{\{b_{X,n}-a_{X,n}\}}\right) \otimes P_{n,l_{Y,n},c_n,\xi}$. Let $(X_{i,n}, Y_{i,n})_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_n$. These random vectors satisfy $\mathbb{E}[Y_{1,n}] = l_{Y,n}$, $(\max - \min)[Y_{1,n}] = b_{Y,n} - a_{Y,n}$ and $(\max - \min)[X_{1,n}] = b_{X,n} - a_{X,n}$. The next step is to build a lower bound on the event $\{\overline{Y}_n = 0\}$.

The assumption that $(X_{i,n}, Y_{i,n})_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_n$ and the construction of $P_{n,l_{Y,n},c_n,\xi}$ imply that

$$\mathbb{P}\left(\overline{Y}_n = 0\right) = \frac{c_n}{n} = \left(1 - \frac{l_{Y,n}}{b_{X,n} - a_{X,n}}(1+\xi)\right)^n = \tilde{\alpha}_{n,H}$$

□

9.10 Additional simulations

This section complements the simulations presented in the main body of the article. Figures show the same objects but with different specifications of the distribution used. We use those different specifications as the order of presentation in this section.

In this setting of simulations, we use the best bounds by setting the constants $l_{Y,n}$ and $u_{Y,n}$ that define our class of distributions equal to the actual corresponding moments (respectively the expectation for $l_{Y,n}$ and the second moment i.e. the expectation of the square for $u_{Y,n}$). That is we use $\overline{n}_\alpha = 2(\mathbb{E}[Y]^2 + \mathbb{V}[Y])/(\alpha \mathbb{E}[Y]^2)$.

In practical settings, the rule-of-thumb will be to replace the theoretical and unknown moments by their empirical counterparts and use the $\overline{n}_\alpha$ obtained to appraise the reliability of the asymptotic CIs from the delta method.

9.10.1 Gaussian distributions

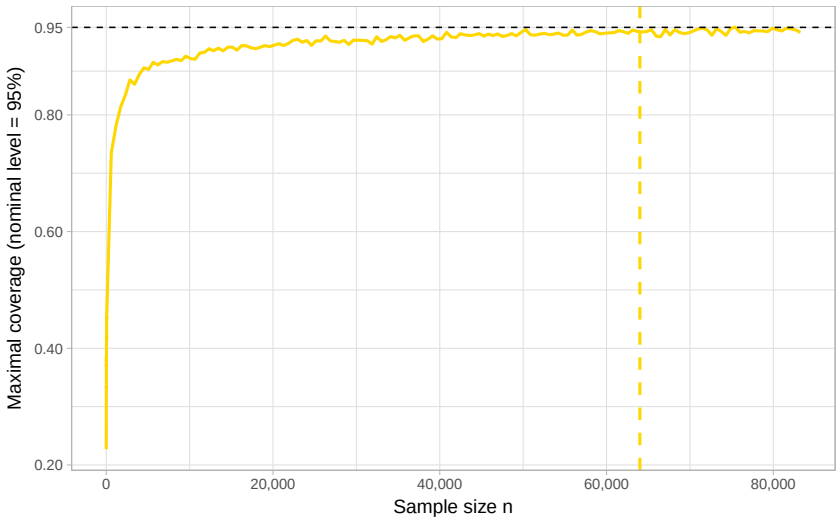


Figure 9.10: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1,1) \otimes \mathcal{N}(0.025, 1)$. The nominal pointwise asymptotic level is set to 0.95. For a sample size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

9.10.2 Rule of thumb using $\overline{\alpha}_n$

Symmetrically, we can consider the rule of thumb for a fixed sample size n and compare the desired nominal level $1-\alpha$ of the test to the rule of thumb $1 - \overline{\alpha}_n$, with the $\overline{\alpha}_n$ derived in Theorem 9.3.

9.10.3 Student distributions

The specification here is two Student distributions, both in the numerator and in the denominator. Standard Student distributions are centered. We use therefore translated versions by simply adding the expectations in order to avoid a null denominator for the ratio of expectations of interest. Below, $\mathcal{T}(\mu, \nu)$ denotes the distribution of a translated standard Student variable: $\mu + T$ where T is distributed according

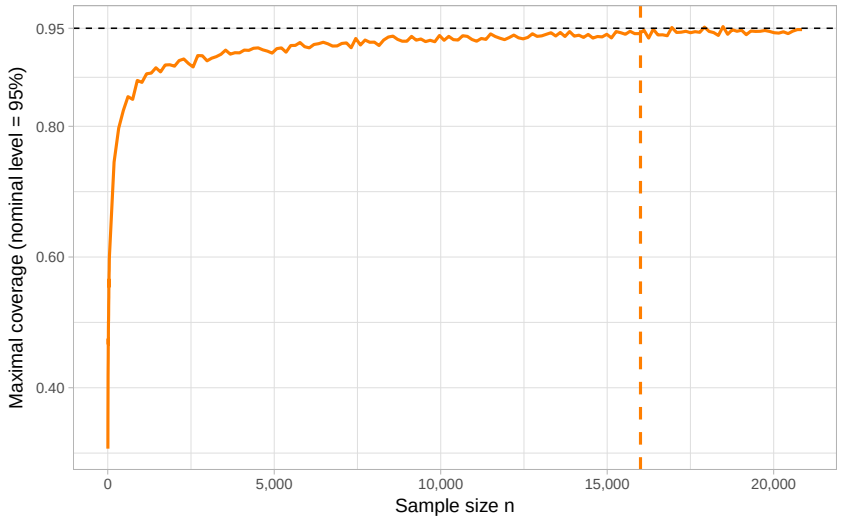


Figure 9.11: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$.
Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1,1) \otimes \mathcal{N}(0.05,1)$. The nominal pointwise asymptotic level is set to 0.95. For a sample size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

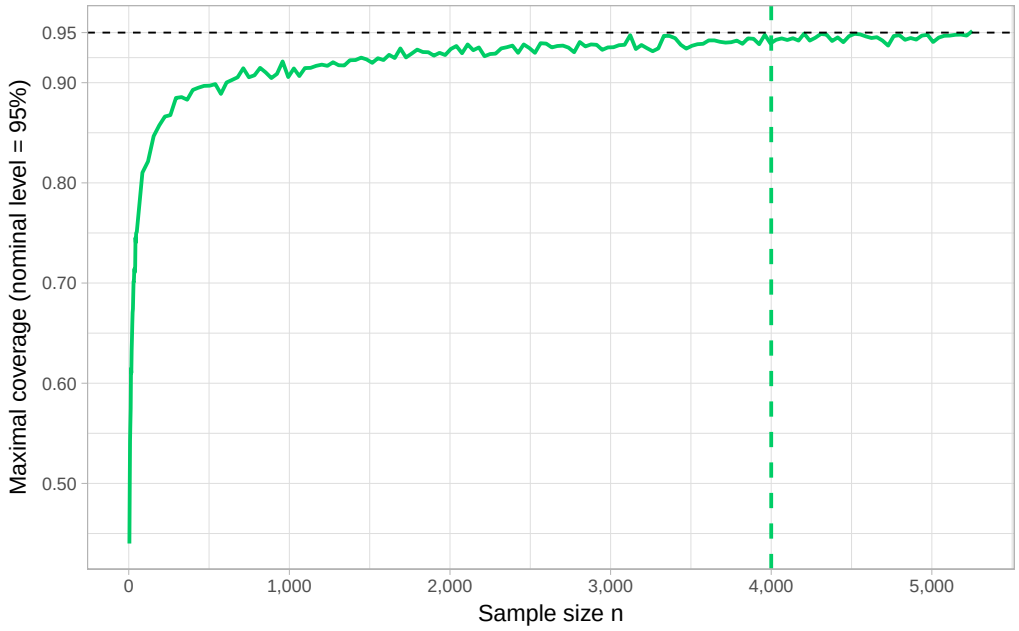


Figure 9.12: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$.
Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1,1) \otimes \mathcal{N}(0.1,1)$. The nominal pointwise asymptotic level is set to 0.95. For a sample size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{V}[Y] + \mathbb{E}[Y]^2$.

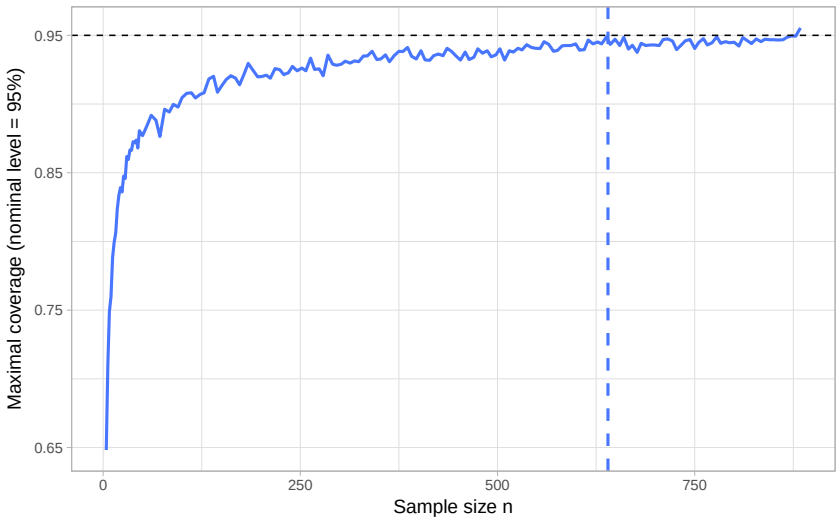


Figure 9.13: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1,1) \otimes \mathcal{N}(0.25,1)$. The nominal pointwise asymptotic level is set to 0.95. For a sample size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.5$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

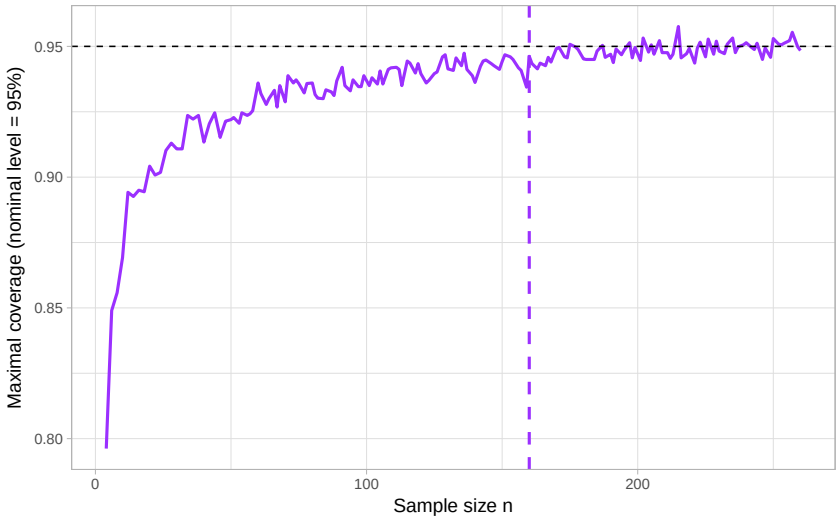


Figure 9.14: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1,1) \otimes \mathcal{N}(0.5,1)$. The nominal pointwise asymptotic level is set to 0.95. For a sample size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.5$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

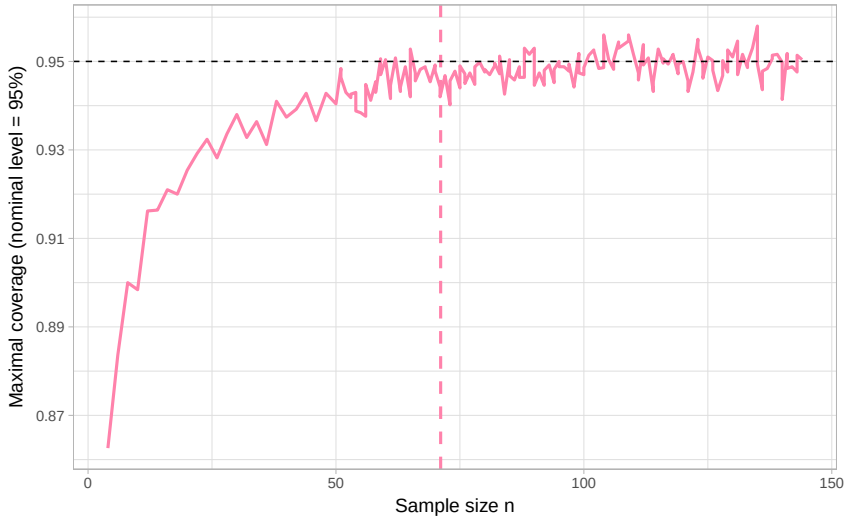


Figure 9.15: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$.
Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}(1,1) \otimes \mathcal{N}(0.75,1)$. The nominal pointwise asymptotic level is set to 0.95. For a sample size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.5$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

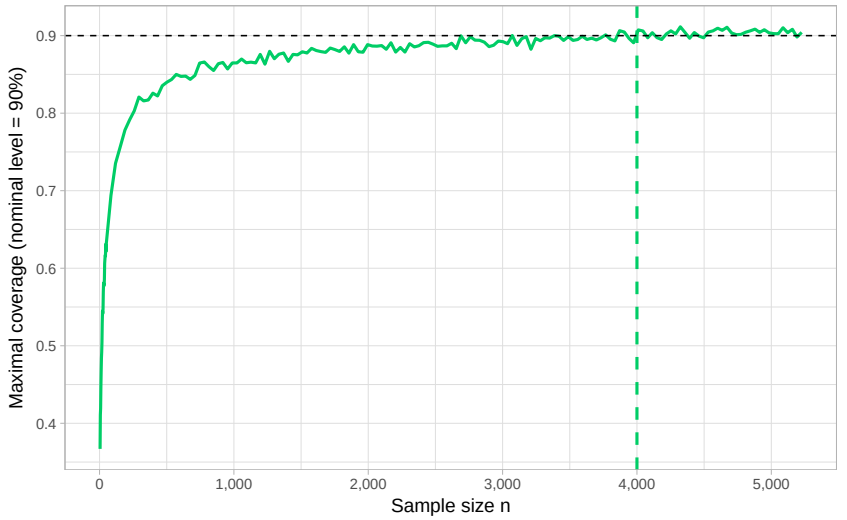


Figure 9.16: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$.
Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}_2$ (bivariate Gaussian) with $\mathbb{E}[X] = 0.5$, $\mathbb{E}[Y] = 0.1$, $\mathbb{V}[X] = 1$, $\mathbb{V}[Y] = 2$, $\mathbb{C}orr(X,Y) = 0.5$. The nominal pointwise asymptotic level is set to 0.90. For a sample size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.1$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

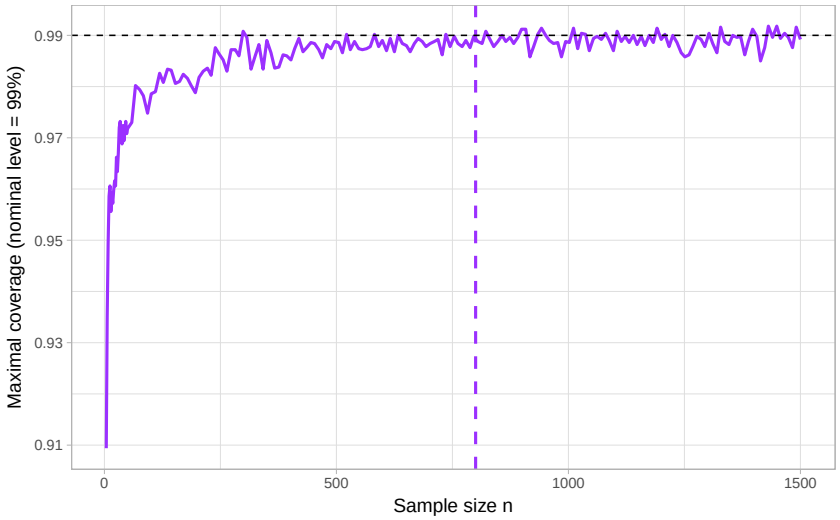


Figure 9.17: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}_2$ (bivariate Gaussian) with $\mathbb{E}[X] = 0.5$, $\mathbb{E}[Y] = 0.5$, $\mathbb{V}[X] = 2$, $\mathbb{V}[Y] = 1$, $\mathbb{C}orr(X, Y) = -0.3$. The nominal pointwise asymptotic level is set to 0.99. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.01$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

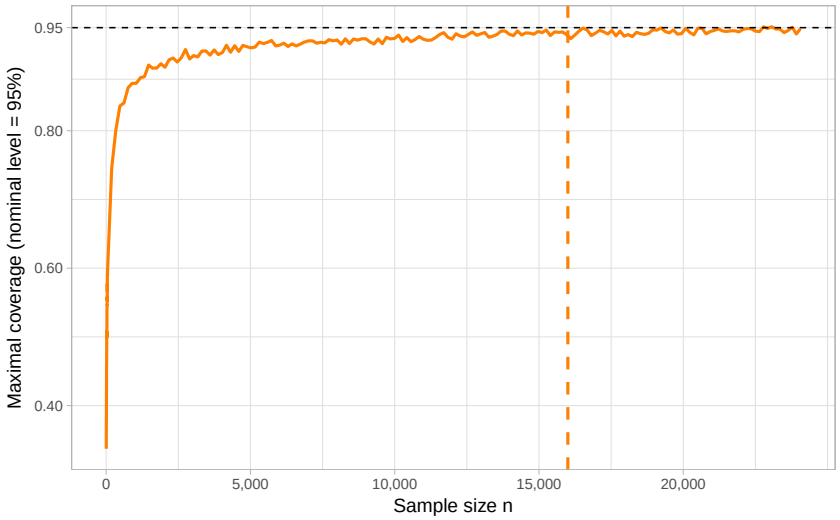


Figure 9.18: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{N}_2$ (bivariate Gaussian) with $\mathbb{E}[X] = 0.5$, $\mathbb{E}[Y] = 0.05$, $\mathbb{V}[X] = 2$, $\mathbb{V}[Y] = 1$, $\mathbb{C}orr(X, Y) = -0.7$. The nominal pointwise asymptotic level is set to 0.95. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

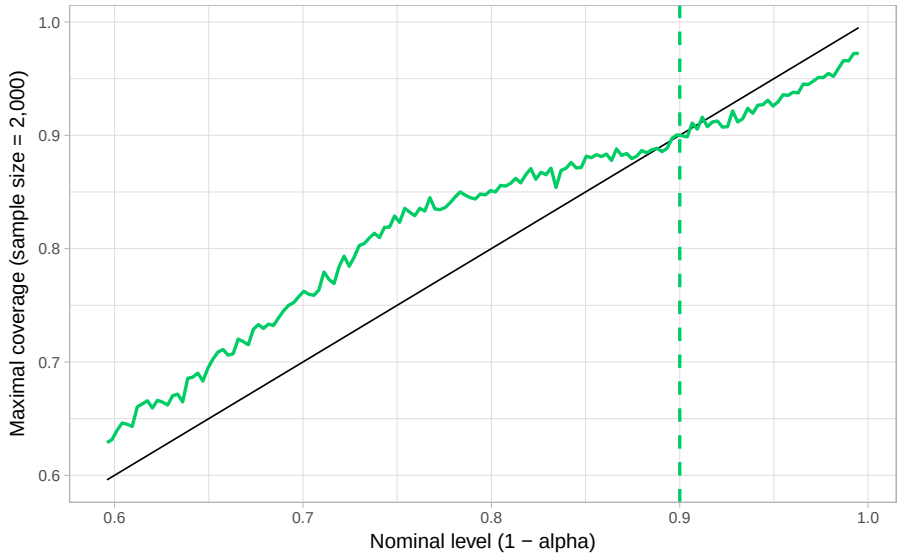


Figure 9.19: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{\alpha}_n$.
Specification: sample size $n = 2,000$, $P_{X,Y,n} = \mathcal{N}_2$ (bivariate Gaussian) with $\mathbb{E}[X] = 1$, $\mathbb{E}[Y] = 0.1$, $\mathbb{V}[X] = 1$, $\mathbb{V}[Y] = 1$, $\mathbb{C}orr(X, Y) = 0$. For each nominal level $1 - \alpha$ in the x-axis, we draw $5,000$ samples, compute the asymptotic CIs and see whether it covers or not the ratio of interest; we report the mean over the $5,000$ repetitions in the y-axis. The solid line is the first bisector $y = x$. The dashed vertical line shows $\overline{\alpha}_n := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(n{l_{Y,n}}^2\right)$, setting here $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

to a Student distribution with ν degrees of freedom. To satisfy Assumption 9.2.1, we need finite variance: we use degrees of freedom strictly higher than 2 for this purpose.

9.10.4 Exponential distributions

The specification here is two exponential distributions, both in the numerator and in the denominator.

In this setting of simulations, we use as previously the best bounds by setting the constants $l_{Y,n}$ and $u_{Y,n}$ that define our class of distributions equal to the actual corresponding moments (respectively the expectation for $l_{Y,n}$ and the second moment i.e. expectation of the square for $u_{Y,n}$). In other words, we use $\overline{n}_\alpha = 2(\mathbb{E}[Y]^2 + \mathbb{V}[Y])/(\alpha\mathbb{E}[Y]^2)$.

For reminder, in practical settings, the rule-of-thumb will be to replace the theoretical and unknown moments by their empirical counterparts and use the $\overline{n}_\alpha$ obtained to appraise the reliability of the asymptotic CIs from the delta method.

However, for exponential distributions, the variance is equal to the square of the expectation. Consequently, whatever the parameter of the exponential distribution in the denominator, we have $\overline{n}_\alpha = 4/\alpha$. Previous simulations suggest that the closer the expectation in the denominator to 0, the larger the sample size required for the asymptotic approximation to hold. At first sight, We might be worried for the usefulness of our rule-of-thumb to get $\overline{n}_\alpha$ independent of $\mathbb{E}[Y]$. However, in the special case of exponential distributions, the lower the expectation, the lower too is the variance. Intuitively, the lower variance will compensate having an expectation closer to 0. The previous statement that links the closeness to 0 of the expectation in the denominator and the sample size required to reach the asymptotic approximation presupposes keeping fixed the variance. It cannot be anymore for exponential distributions.

The following simulations reveal that the convergence of the coverage of the asymptotic confidence

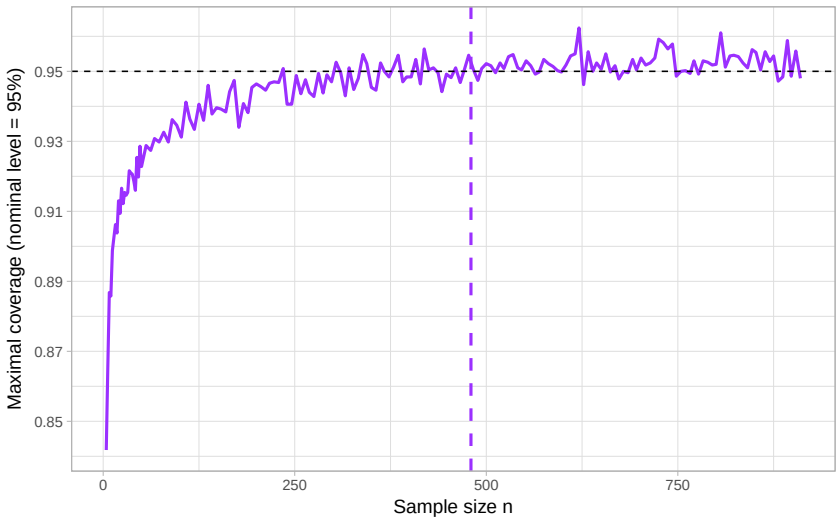


Figure 9.20: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{T}(\mathbb{E}[X], 3) \otimes \mathcal{T}(\mathbb{E}[Y], 3)$ with $\mathbb{E}[X] = 0.5$, $\mathbb{E}[Y] = 0.5$. The nominal pointwise asymptotic level is set to 0.95. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

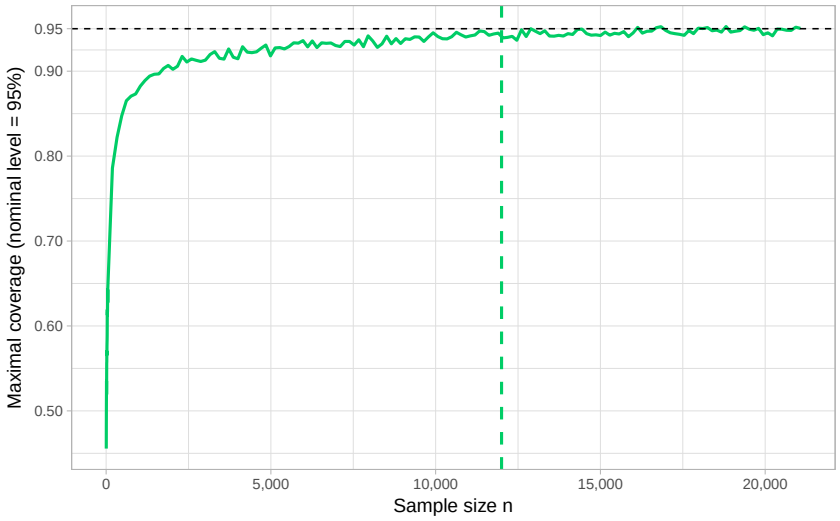


Figure 9.21: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{T}(\mathbb{E}[X], 3) \otimes \mathcal{T}(\mathbb{E}[Y], 3)$ with $\mathbb{E}[X] = 0.5$, $\mathbb{E}[Y] = 0.1$. The nominal pointwise asymptotic level is set to 0.95. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

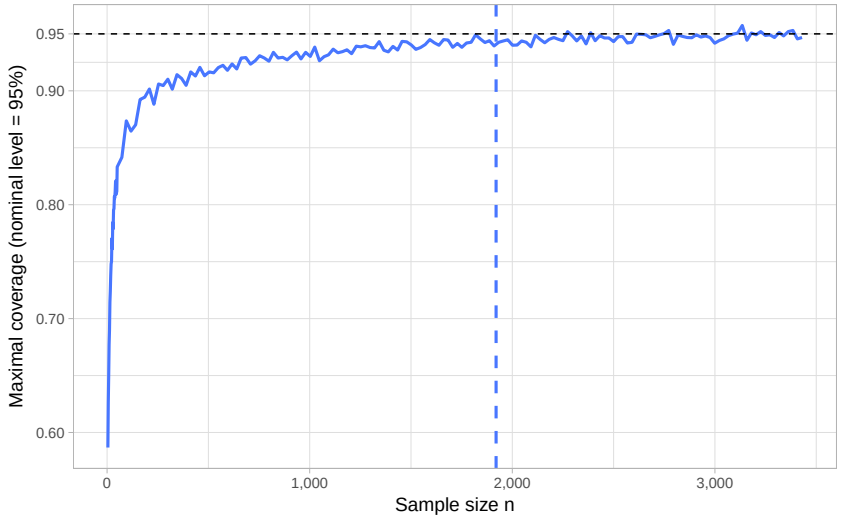


Figure 9.22: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*$, the marginal distributions of X and Y are $\mathcal{T}(\mathbb{E}[X], 3)$ and $\mathcal{T}(\mathbb{E}[Y], 3)$, with $\mathbb{E}[X] = 1$, $\mathbb{E}[Y] = 0.25$ and simulated using a Gaussian copula to have $\mathbb{C}orr(X, Y) \approx 0.5$. The nominal pointwise asymptotic level is set to 0.95. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

intervals toward their nominal level happens for n around one hundred and more and, furthermore, has the same pattern whatever the expectation of the exponential distribution in the denominator. Our rule-of-thumb $\overline{n}_\alpha$ appears to be a bit small. Nonetheless, it is coherent that it is constant across the value of $\mathbb{E}[Y]$.

9.10.5 Pareto distributions

The specification here is two Pareto distributions, both in the numerator and in the denominator. Pareto distributions have support in $\mathbb{R}_+^*$. They would fall in the easier case when the support of the denominator is well separated from 0. To assess the dependability of our rule-of-thumb in the general case, we use translated Pareto distributions. In what follows, the notation $\mathcal{P}areto(\mathbb{E}[Y], \tau, \gamma)$ denotes the distribution of a random variable that follows a Pareto distribution with shape parameter equal to γ translated such that its support is $(\tau, +\infty)$ and its expectation is $\mathbb{E}[Y]$. A variable that is distributed according to $\mathcal{P}areto(\mathbb{E}[Y], \tau, \gamma)$ is equal in distribution to $P + (\mathbb{E}[Y] - \gamma t_Y)/(\gamma - 1)$ with $t_Y = (\mathbb{E}[Y] - \tau) \times (\gamma - 1)$ and P a usual Pareto distribution with support or scale parameter t_Y and shape parameter γ, that is P has the density $x \mapsto \mathbb{1}\{x \geq t_Y\} \times \gamma t_Y^\gamma / x^{\gamma+1}$ with respect to Lebesgue measure.

9.10.6 Bernoulli distributions

With discrete distributions for the variable at the denominator, it may happen that $\overline{Y}_n = 0$, all the more so as the expectation in the denominator and the sample size are low in the case of Bernoulli distributions. In that situation, the confidence interval is said to be *undefined* and, for any arbitrary value, the statement that the CI contains that value is considered as false. Consequently, in the simulations with discrete distributions in the denominator, whenever the sample drawn is such that $\overline{Y}_n = 0$, we count the draw as a no coverage occurrence in the Monte Carlo estimation of the (maximal) coverage. Concretely, the maximal coverage displayed in the graph is computed as the mean over B repetitions. The repetitions

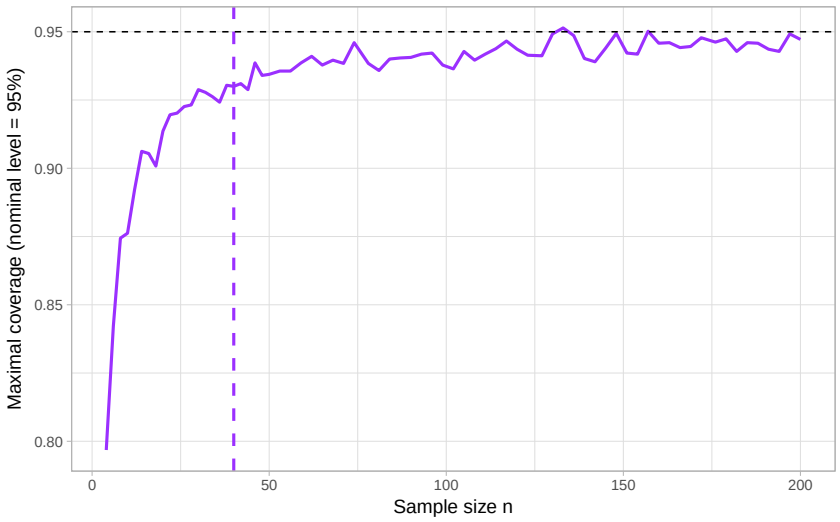


Figure 9.23: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{E} \otimes \mathcal{E}$ with $\mathbb{E}[X] = 1$, $\mathbb{E}[Y] = 0.5$. The nominal pointwise asymptotic level is set to 0.95. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

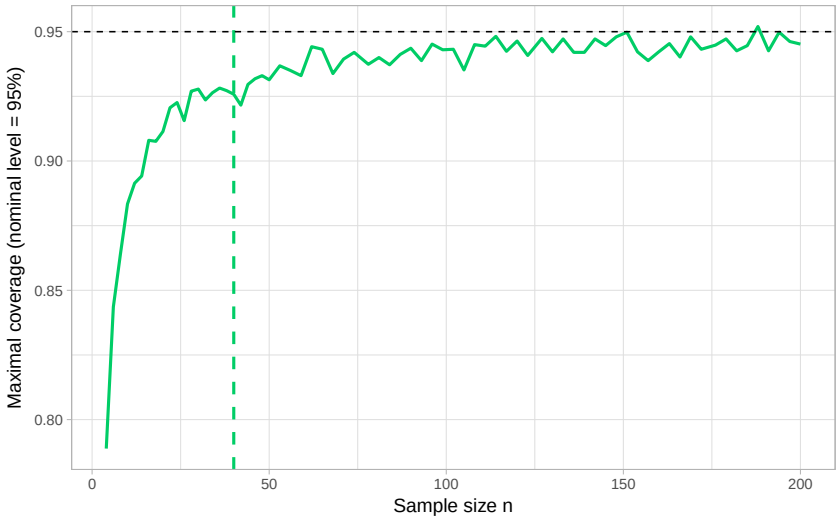


Figure 9.24: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{E} \otimes \mathcal{E}$ with $\mathbb{E}[X] = 1$, $\mathbb{E}[Y] = 0.1$. The nominal pointwise asymptotic level is set to 0.95. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

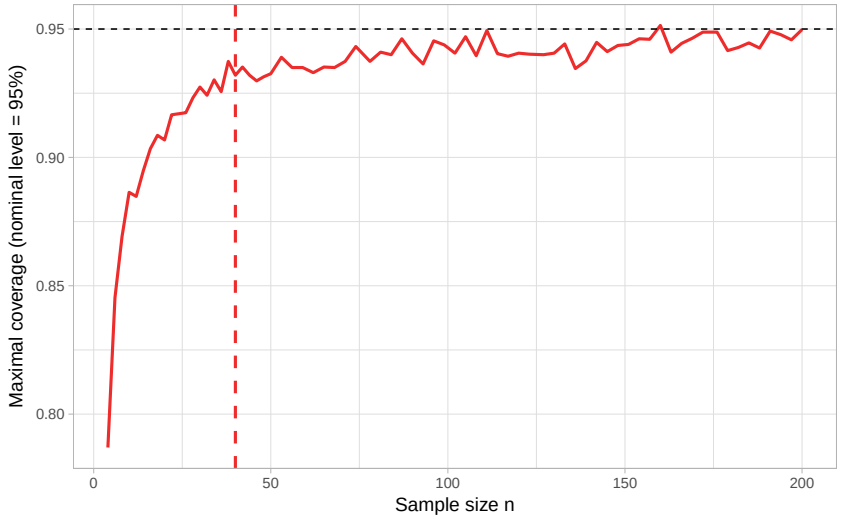


Figure 9.25: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{E} \otimes \mathcal{E}$ with $\mathbb{E}[X] = 1$, $\mathbb{E}[Y] = 0.01$. The nominal pointwise asymptotic level is set to 0.95. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

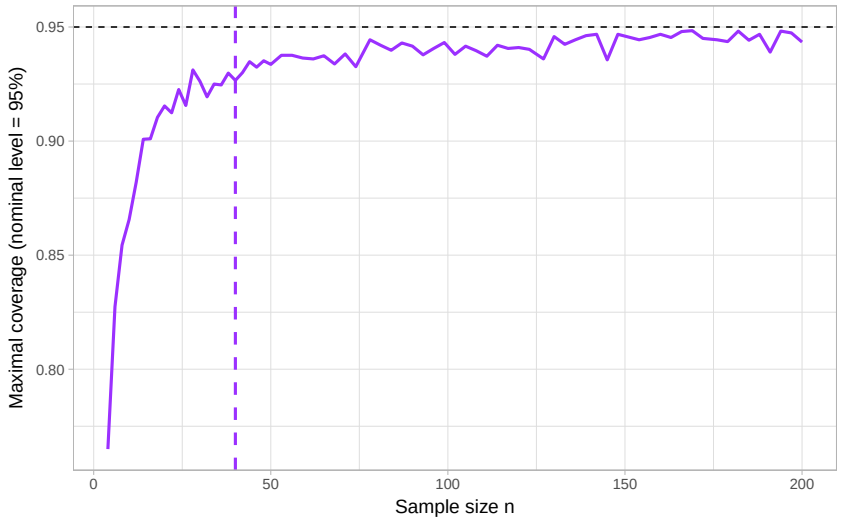


Figure 9.26: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*$, the marginal distributions of X and Y are two exponentials, with $\mathbb{E}[X] = 1$, $\mathbb{E}[Y] = 0.5$ and simulated using a Gaussian copula to have $\mathbb{C}orr(X, Y) \approx 0.75$. The nominal pointwise asymptotic level is set to 0.95. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

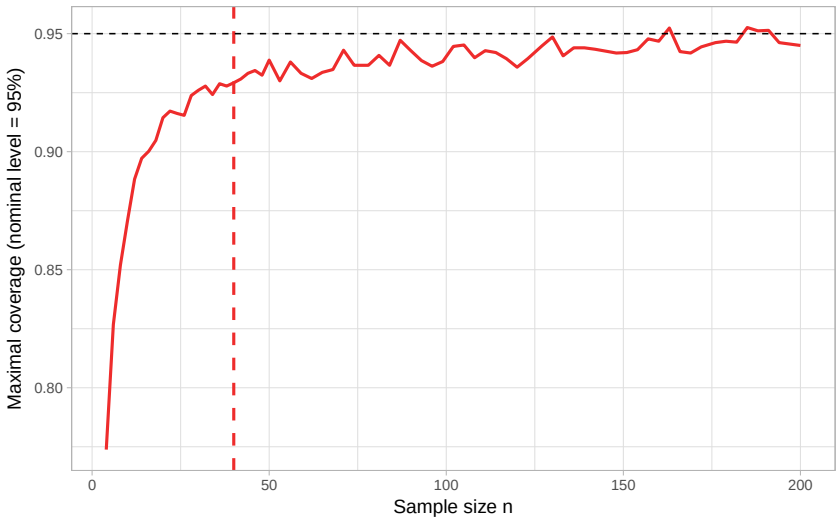


Figure 9.27: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*$, the marginal distributions of X and Y are two exponentials, with $\mathbb{E}[X] = 1$, $\mathbb{E}[Y] = 0.01$ and simulated using a Gaussian copula to have $\mathbb{C}orr(X, Y) \approx 0.75$. The nominal pointwise asymptotic level is set to 0.95. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

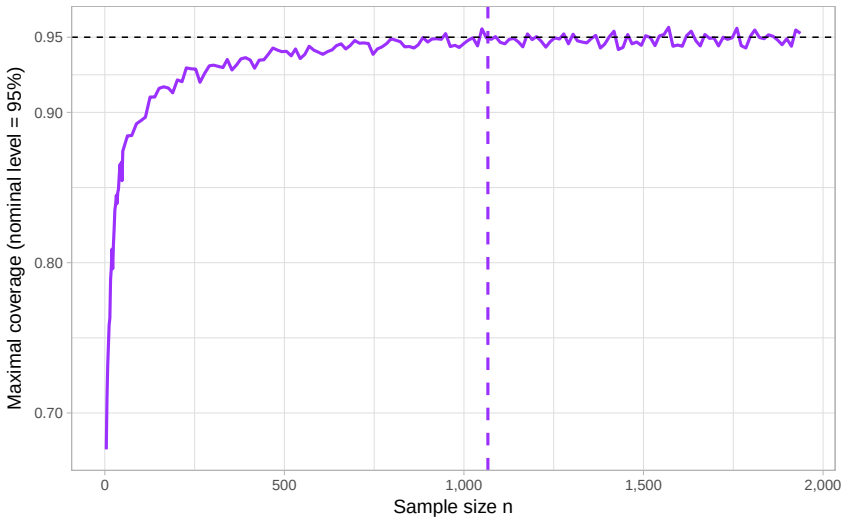


Figure 9.28: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{P}areto(1, -1.5, 5) \otimes \mathcal{P}areto(\mathbb{E}[Y], -1.5, 5)$, with $\mathbb{E}[Y] = 0.5$. The nominal pointwise asymptotic level is set to 0.95. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

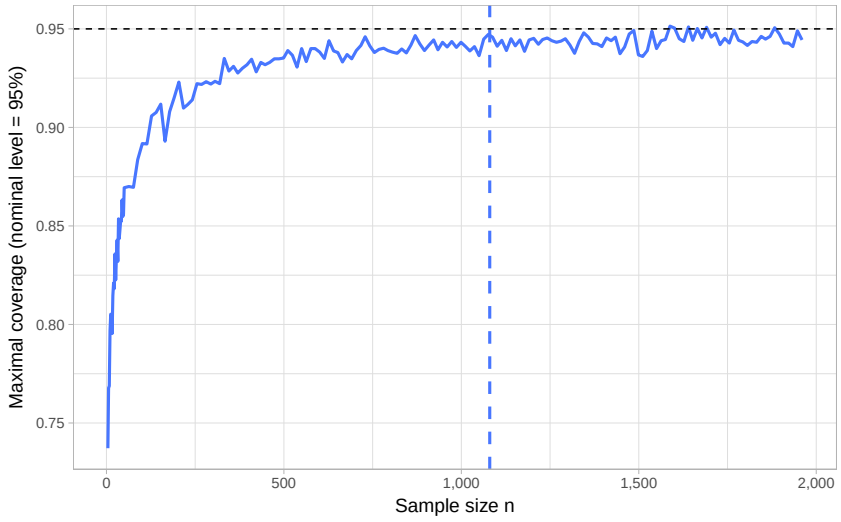


Figure 9.29: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{P}areto(1, -1.5, 3) \otimes \mathcal{P}areto(\mathbb{E}[Y], -0.5, 3)$, with $\mathbb{E}[Y] = 0.25$. The nominal pointwise asymptotic level is set to 0.95. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

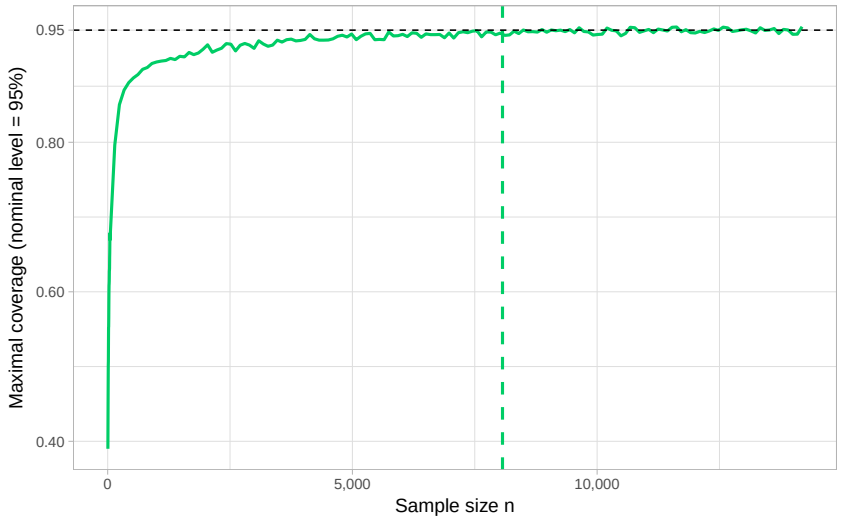


Figure 9.30: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{P}areto(1, -1.5, 5) \otimes \mathcal{P}areto(\mathbb{E}[Y], -1, 5)$, with $\mathbb{E}[Y] = 0.1$. The nominal pointwise asymptotic level is set to 0.95. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

for which $\overline{Y}_n = 0$ account for 0 in the mean.

Note that in some specifications, a substantial part of the repetitions yield $\overline{Y}_n = 0$. For instance, with Bernoulli distributions, for n smaller than 10 and the expectation at the denominator equal to 0.01, around 10% only of the repetitions display $\overline{Y}_n \neq 0$.

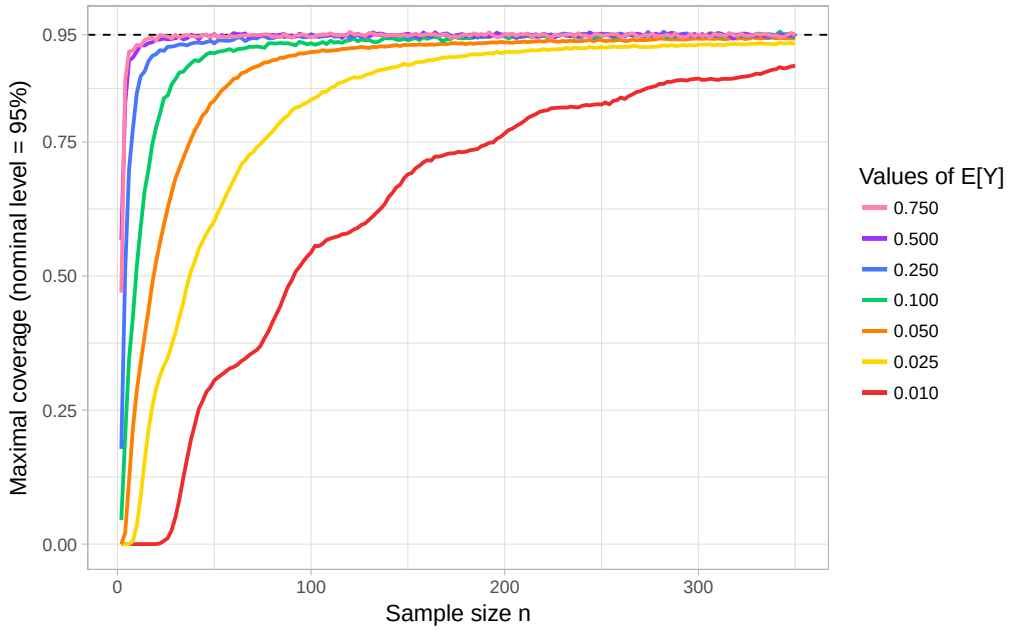


Figure 9.31: Maximal coverage of asymptotic CIs based on the delta method.
Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{B}(0.5) \otimes \mathcal{B}(\mathbb{E}[Y])$. The nominal pointwise asymptotic level is set to 0.95. For a pair $(\mathbb{E}[Y], n)$, the coverage is obtained as the mean over 10,000 repetitions.

With two Bernoulli variables in the numerator and the denominator, we are both in the BC and the Hoeffding cases. The following graphs illustrate the use of $\overline{n}_\alpha$ to appraise the reliability of the asymptotic confidence based on the delta method. We show both the one obtained in the BC case (Theorem 9.3) and the one obtained in the Hoeffding case (Theorem 9.12). Again, as in the main body of the paper, we follow a plug-in strategy to compute in practice $\overline{n}_\alpha$ and, in the setting of simulations, we simply use the known moments and bounds of the DGP used in the simulation.

9.10.7 Poisson distributions

The specification here considers two variables distributed according to a Poisson distribution, both in the numerator and in the denominator.

A Poisson distribution is entirely defined by its positive real parameter, which is equal to both its expectation and its variance. Consequently, to have denominator close to 0, we would need small variance too, as in the exponential specification (see Section 9.10.4). In order to disentangle expectation and variance, we use below translated Poisson variables. Precisely, the notation $\mathcal{P}oisson(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+^*$, denotes a distribution alike to a Poisson, with parameter and variance equal to σ^2 but translated such that its expectation is μ. That is a variable distributed according to $\mathcal{P}oisson(\mu, \sigma^2)$ is equal in distribution to $P + (\mu - \sigma^2)$ with P a standard Poisson distribution with parameter σ^2 - that is with density with respect to the counting measure equal to $(\sigma^2)^k \exp(-\sigma^2)/(k!)$ for every $k \in \mathbb{N}$. Thus, a $\mathcal{P}oisson(\mu, \sigma^2)$ has expectation μ and variance σ^2.

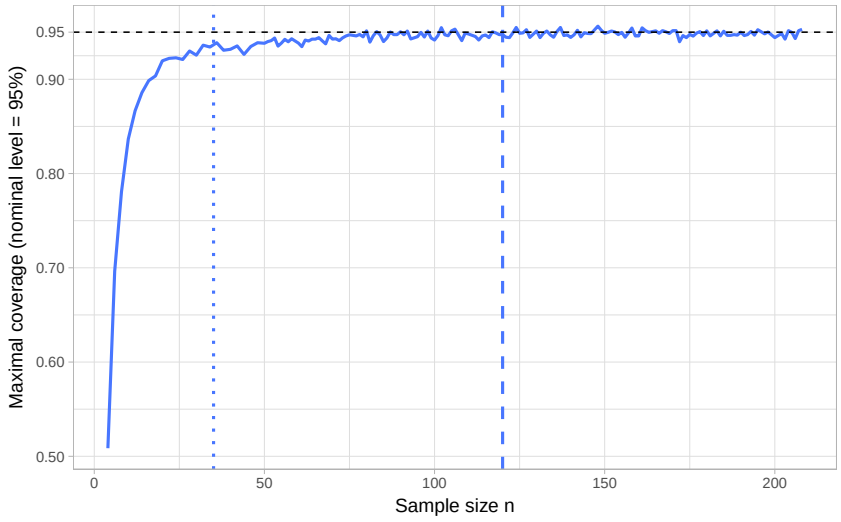


Figure 9.32: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$.
Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{B}(0.5) \otimes \mathcal{B}(0.25)$. The nominal pointwise asymptotic level is set to 0.95. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$ (BC case). The dotted one shows $\overline{n}_\alpha := \ln(4/\alpha)/\gamma(Y_{1,n})$, setting here $\alpha = 0.05$, $a_{Y,n} = 0$, $b_{Y,n} = 1$ and $l_{Y,n} = \mathbb{E}[Y]$ (Hoeffding case).

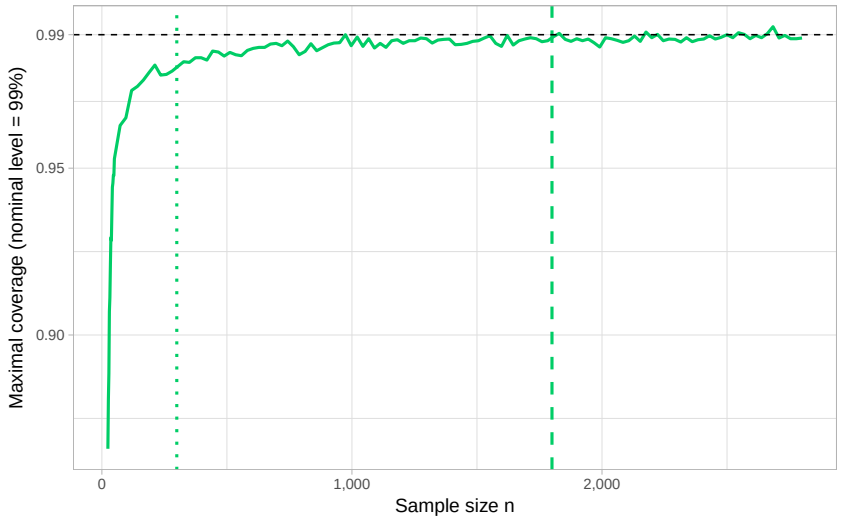


Figure 9.33: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$.
Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{B}(0.5) \otimes \mathcal{B}(0.1)$. The nominal pointwise asymptotic level is set to 0.99. For a size n, the coverage is obtained as the mean over 10,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.01$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$ (BC case). The dotted shows $\overline{n}_\alpha := \ln(4/\alpha)/\gamma(Y_{1,n})$, setting here $\alpha = 0.01$, $a_{Y,n} = 0$, $b_{Y,n} = 1$ and $l_{Y,n} = \mathbb{E}[Y]$ (Hoeffding case).

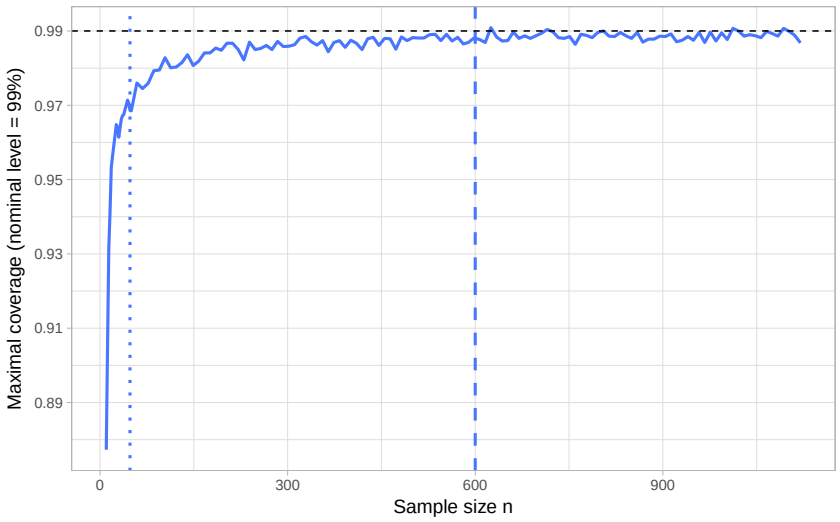


Figure 9.34: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$ (zoom: starting from higher n compared to Figure 9.32).
Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{B}(0.5) \otimes \mathcal{B}(0.25)$. The nominal pointwise asymptotic level is set to 0.99. For a size n, the coverage is obtained as the mean over 10,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.01$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$ (BC case). The dotted shows $\overline{n}_\alpha := \ln(4/\alpha)/\gamma(Y_{1,n})$, setting here $\alpha = 0.01$, $a_{Y,n} = 0$, $b_{Y,n} = 1$ and $l_{Y,n} = \mathbb{E}[Y]$ (Hoeffding case).

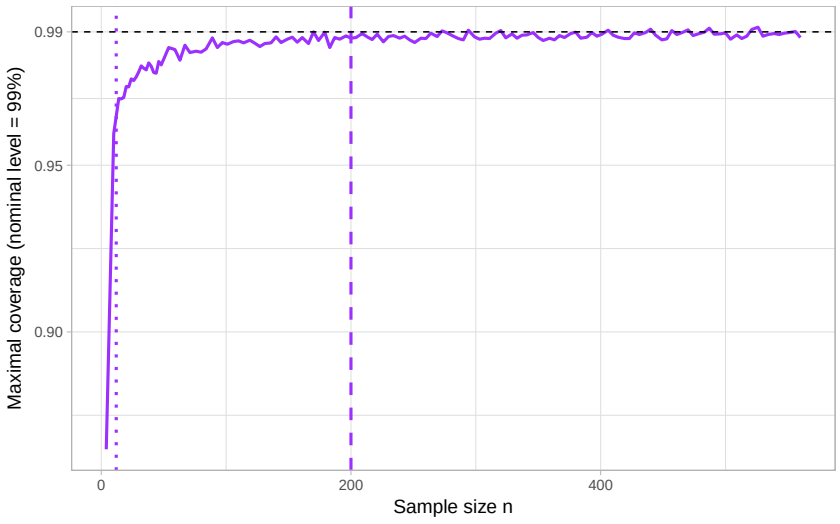


Figure 9.35: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$.
Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{B}(0.5) \otimes \mathcal{B}(0.5)$. The nominal pointwise asymptotic level is set to 0.99. For a size n, the coverage is obtained as the mean over 10,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.01$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$ (BC case). The dotted shows $\overline{n}_\alpha := \ln(4/\alpha)/\gamma(Y_{1,n})$, setting here $\alpha = 0.01$, $a_{Y,n} = 0$, $b_{Y,n} = 1$ and $l_{Y,n} = \mathbb{E}[Y]$ (Hoeffding case).

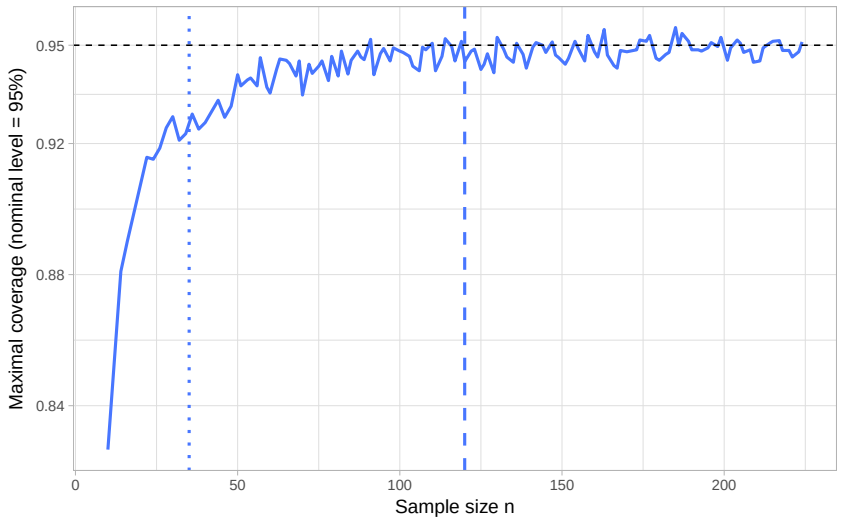


Figure 9.36: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*$, the marginal distributions of X and Y are two Bernoulli, with $\mathbb{E}[X] = 0.5$ and $\mathbb{E}[Y] = 0.25$ and simulated using a Gaussian copula to have $\mathbb{C}orr(X, Y) \approx 0.35$. The nominal pointwise asymptotic level is set to 0.95. For a size n, the coverage is obtained as the mean over 10,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$ (BC case). The dotted shows $\overline{n}_\alpha := \ln(4/\alpha)/\gamma(Y_{1,n})$, setting here $\alpha = 0.05$, $a_{Y,n} = 0$, $b_{Y,n} = 1$ and $l_{Y,n} = \mathbb{E}[Y]$ (Hoeffding case).

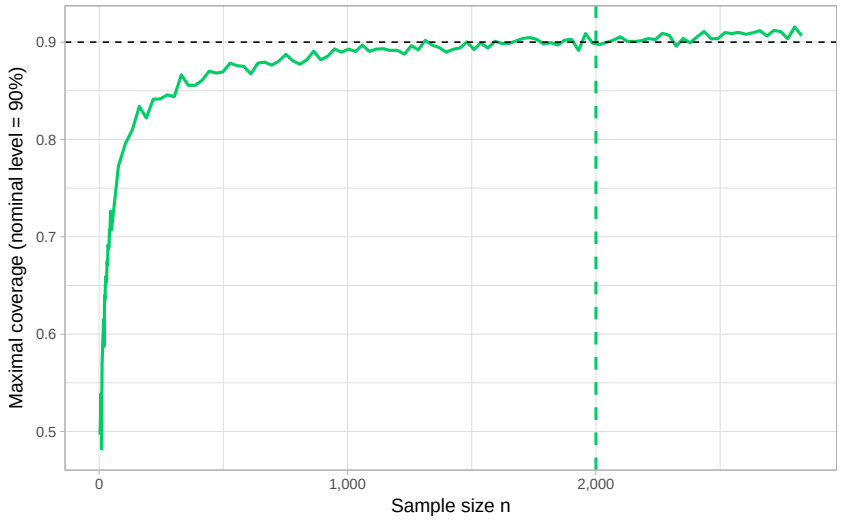


Figure 9.37: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{P}oisson(0.5, 2) \otimes \mathcal{P}oisson(0.1, 1)$. The nominal pointwise asymptotic level is set to 0.9. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.1$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

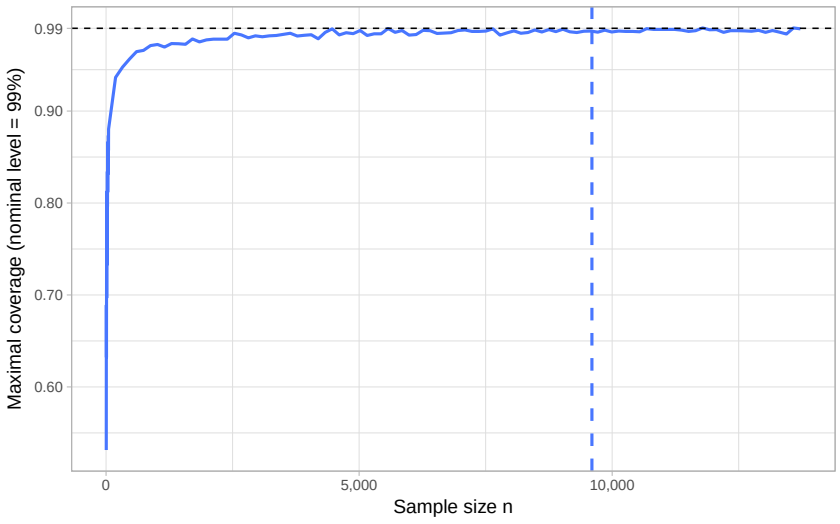


Figure 9.38: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{P}oisson(1,4) \otimes \mathcal{P}oisson(0.25,3)$. The nominal pointwise asymptotic level is set to 0.99. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.01$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

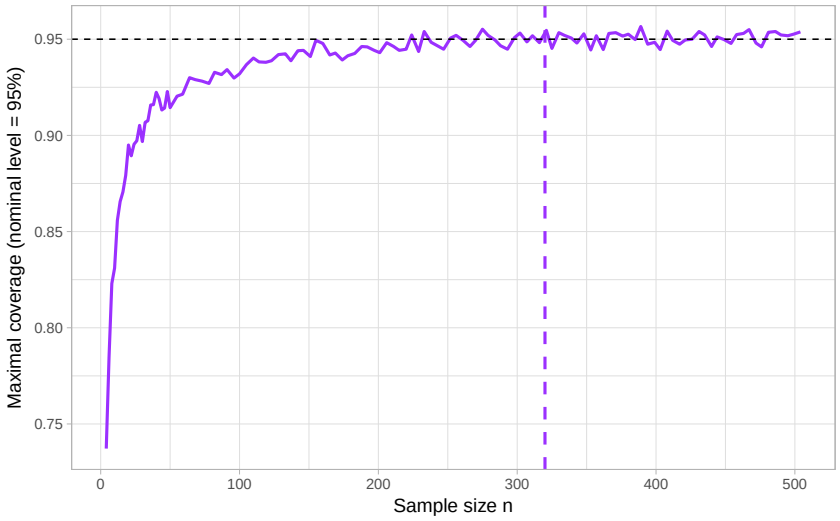


Figure 9.39: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$. Specification: $\forall n \in \mathbb{N}^*, P_{X,Y,n} = \mathcal{P}oisson(0.5,2) \otimes \mathcal{P}oisson(0.5,2)$. The nominal pointwise asymptotic level is set to 0.95. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.05$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

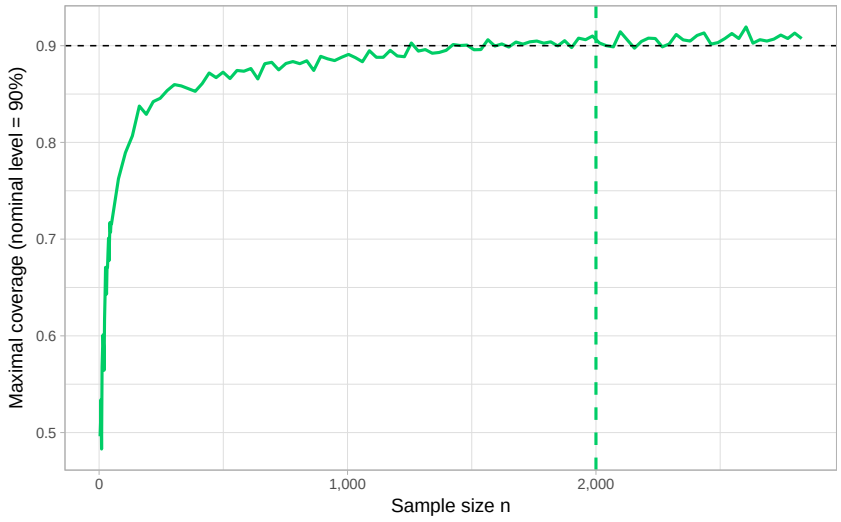


Figure 9.40: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$.
Specification: $\forall n \in \mathbb{N}^*$, the marginal distributions of X and Y are $\mathcal{P}oisson(0.5, 2)$ and $\mathcal{P}oisson(0.1, 1)$, and simulated using a Gaussian copula to have $\mathbb{C}orr(X, Y) \approx 0.7$. The nominal pointwise asymptotic level is set to 0.90. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.1$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

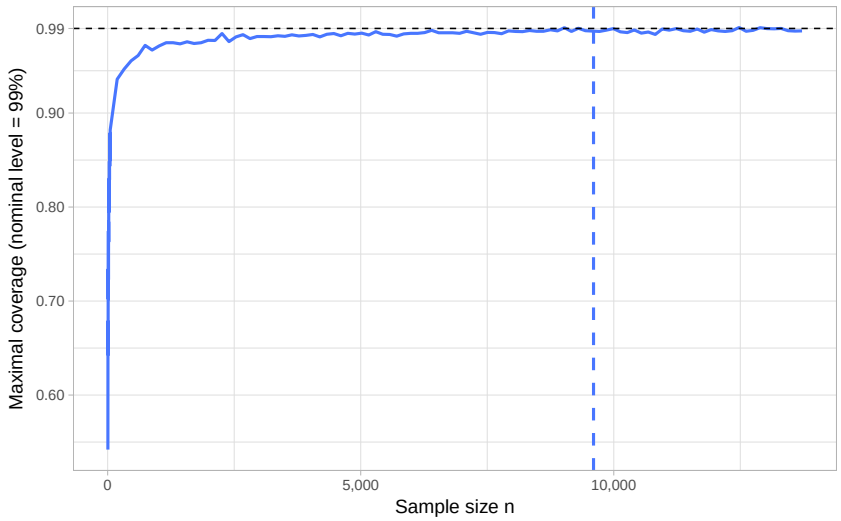


Figure 9.41: Maximal coverage of asymptotic CIs based on the delta method and rule-of-thumb $\overline{n}_\alpha$.
Specification: $\forall n \in \mathbb{N}^*$, the marginal distributions of X and Y are $\mathcal{P}oisson(1, 4)$ and $\mathcal{P}oisson(0.25, 3)$, and simulated using a Gaussian copula to have $\mathbb{C}orr(X, Y) \approx 0.4$. The nominal pointwise asymptotic level is set to 0.99. For a size n, the coverage is obtained as the mean over 5,000 repetitions. The dashed vertical line shows $\overline{n}_\alpha := 2\left(u_{Y,n} - {l_{Y,n}}^2\right) / \left(\alpha l_{Y,n}^2\right)$, setting here $\alpha = 0.01$, $l_{Y,n} = \mathbb{E}[Y]$, $u_{Y,n} = \mathbb{E}[Y]^2 + \mathbb{V}[Y]$.

Remerciements

Je voudrais adresser tout d'abord mes remerciements à mon directeur de thèse, Jean-David Fermanian. Je me rappelle encore du début, quand on s'est rencontré alors que je n'étais qu'un étudiant en début de deuxième année de l'ENSAE. À l'époque, le projet de Statistique appliquée m'apparaissait comme le cours le plus intéressant de l'année, et c'est grâce à ce projet que je me suis engagé dans la recherche. Je te remercie beaucoup pour tous les conseils et toutes les discussions que nous avons eues depuis, et j'ai appris beaucoup grâce à toi.

Je voudrais ensuite remercier Alexandre Tsybakov d'avoir accepté de co-superviser ma thèse. Merci beaucoup de m'avoir accepté au sein du laboratoire de Statistique du CREST, et pour ton aide tout au long de cette thèse. Je remercie aussi Jean-Michel Zakoïan pour son accueil au laboratoire de Finance-Assurance. Je suis très reconnaissant pour tous les échanges que j'ai pu avoir avec les professeurs du CREST, en particulier Pierre Alquier, Guillaume Lecué, Cristina Butucea, Arnak Dalalyan, Nicolas Chopin, Marco Cuturi et Christian Francq, et pour tous ces moments passés avec les autres doctorants, en particulier mes co-auteurs Yannick et Lucas, ainsi que Gautier, Lionel, Badr, Geoffrey, Phillipe, Boris, Clara, Alexander, Vincent, Edwin, Anna, Nicolas, Léna, Mohammed, Amir, Avo, Gabriel, Ophélie et Sébastien. Enfin, je souhaite remercier beaucoup tous les membres du jury: Ivan Kojadinovic et Marten Wegkamp pour leurs rapports très positifs sur cette thèse, la présidente du jury, Anne-Laure Fougères ainsi que les examinateurs Dominique Picard et Matthieu Lerasle.

Je remercie mes amis Julie-Anne, James, Fabien, Bhawna, Guillaume, Ali, Yannis, Alexis, Sylvain et David pour leur soutien durant cette longue période. Je remercie ma famille et en particulier mes parents, qui m'ont offert le meilleur cadre de travail possible pour ma thèse, ainsi que mon frère Nicolas et ma soeur Héloïse. Finalement, j'ai une pensée spéciale pour Anh Thư et Mỹ Phương, qui ont toujours été là pour me soutenir durant ces dernières années.

Bibliography

[1] K. Aas, C. Czado, A. Frigessi, and H. Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, 44(2):182–198, 2009.

[2] F. Abegaz, I. Gijbels, and N. Veraverbeke. Semiparametric estimation of conditional copulas. *Journal of Multivariate Analysis*, 110:43–73, 2012.

[3] E. F. Acar, R. V. Craiu, and F. Yao. Dependence calibration in conditional copulas: A nonparametric approach. *Biometrics*, 67(2):445–453, 2011.

[4] E. F. Acar, R. V. Craiu, and F. Yao. Statistical testing of covariate effects in conditional copula models. *Electronic Journal of Statistics*, 7:2822–2850, 2013.

[5] E. F. Acar, C. Genest, and J. Nešlehová. Beyond simplified pair-copula constructions. *Journal of Multivariate Analysis*, 110:74–90, 2012.

[6] C. Almeida and C. Czado. Efficient bayesian inference for stochastic time-varying copula models. *Comput. Statist. Data Anal.*, 56:1511–1527, 2012.

[7] D. W. Andrews. A conditional Kolmogorov test. *Econometrica: Journal of the Econometric Society*, pages 1097–1128, 1997.

[8] A. V. Asimit, R. Gerrard, Y. Hou, and L. Peng. Tail dependence measure for examining financial extreme co-movements. *Journal of Econometrics*, 194(2):330–348, 2016.

[9] R. R. Bahadur and L. J. Savage. The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics*, 27(4):1115–1122, 1956.

[10] T. Bedford and R. M. Cooke. Vines: A new graphical model for dependent random variables. *Annals of Statistics*, pages 1031–1068, 2002.

[11] P. C. Bellec, G. Lecué, and A. B. Tsybakov. Towards the study of least squares estimators with convex penalty. *Séminaires et Congrès*, 39, 2017.

[12] P. C. Bellec, G. Lecué, and A. B. Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018.

[13] P. C. Bellec and A. B. Tsybakov. Bounds on the prediction error of penalized least squares estimators with convex penalty. In V. Panov, editor, *Modern Problems of Stochastic Analysis and Statistics, Selected Contributions In Honor of Valentin Konakov*. Springer, 2017.

[14] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[15] A. Belloni, V. Chernozhukov, and L. Wang. Pivotal estimation via square-root lasso in nonparametric regression. *Annals of Statistics*, 42(2):757–788, 2014.

[16] D. Berg. Copula goodness-of-fit testing: an overview and power comparison. *The European Journal of Finance*, 15(7-8):675–701, 2009.

[17] W. Bergsma. Nonparametric testing of conditional independence by means of the partial copula. *Arxiv preprint, arXiv:1101.4607*, 2011.

[18] P. Bickel and P. Millar. Uniform convergence of probability measures on classes of functions. *Statistica Sinica*, pages 1–15, 1992.

[19] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[20] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès. Slope - adaptive variable selection via convex optimization. *Annals of Applied Statistics*, 9(3):1103, 2015.

[21] D. Bosq and J.-P. Lecoutre. *Théorie de l'estimation fonctionnelle*. Economica, 1987.

[22] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[23] S. Bouzebda, A. Keziou, and T. Zari. K-sample problem using strong approximations of empirical copula processes. *Mathematical Methods of Statistics*, 20(1):14–29, 2011.

[24] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning*, 3(1):1–122, 2011.

[25] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Republished by CRC Press., Wadsworth, Belmont, CA., 1984.

[26] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.

[27] L. L. Chaieb, L.-P. Rivest, and B. Abdous. Estimating survival under a dependent truncation. *Biometrika*, 93(3):655–669, 2006.

[28] V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597, 2014.

[29] U. Cherubini, E. Luciano, and W. Vecchiato. *Copula methods in finance*. Wiley, 2004.

[30] G. Chinot, L. Guillaume, and L. Matthieu. Statistical learning with lipschitz and convex loss functions. *arXiv preprint arXiv:1810.01090*, 2018.

[31] L. Chollete, A. Heinen, and A. Valdesogo. Modeling international financial returns with a multivariate regime-switching copula. *Journal of financial econometrics*, 7(4):437–480, 2009.

[32] V. R. Craiu and A. Sabeti. In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes. *Journal of Multivariate Analysis*, 110:106–120, 2012.

[33] P. Deheuvels. La fonction de dependence empirique et ses proprietes, un test non parametrique d'independance. *Bulletin de la classe des sciences, Academie Royale de Belgique, 5e serie*, 65:274–292, 1979.

[34] P. Deheuvels. A kolmogorov-smirnov type test for independence and multivariate samples. *Revue roumaine de mathématiques pures et appliquées*, 26(2):213–226, 1981.

[35] A. Derumigny. Improved bounds for square-root Lasso and square-root Slope. *Electronic Journal of Statistics*, 12(1):741–766, 2018.

[36] A. Derumigny. Estimation of a regular conditional functional by conditional U-statistics regression. *ArXiv preprint, arXiv:1903.10914*, 2019.

[37] A. Derumigny. Robust-to-outliers simultaneous inference and noise level estimation using a MOM approach. In progress, 2019.

[38] A. Derumigny and J.-D. Fermanian. About tests of the "simplifying" assumption for conditional copulas. *Dependence Modeling*, 5(1):154–197, 2017.

[39] A. Derumigny and J.-D. Fermanian. About Kendall's regression. *Arxiv preprint, arXiv:1802.07613*, 2018.

[40] A. Derumigny and J.-D. Fermanian. About kernel-based estimation of the conditional Kendall's tau: finite-distance bounds and asymptotic behavior. *Arxiv preprint, arXiv:1810.06234*, 2018.

[41] A. Derumigny and J.-D. Fermanian. A classification point-of-view about conditional Kendall's tau. *Computational Statistics & Data Analysis*, 135:70–94, 2019.

[42] A. Derumigny, L. Girard, and Y. Guyonvarch. On the construction of confidence intervals for ratios of expectations. *ArXiv preprint, arXiv:1904.07111*, 2019.

[43] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

[44] D. L. Donoho and P. J. Huber. The notion of breakdown point. *A festschrift for Erich L. Lehmann*, 157184, 1983.

[45] J. Dony and D. Mason. Uniform in bandwidth consistency of conditional U-statistics. *Bernoulli*, 14(4):1108–1133, 2008.

[46] J.-M. Dufour. Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica*, 65(6):1365–1387, 1997.

[47] U. Einmahl and D. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.*, 33(3):1380–1403, 2005.

[48] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

[49] J.-D. Fermanian. Goodness-of-fit tests for copulas. *Journal of multivariate analysis*, 95(1):119–152, 2005.

[50] J.-D. Fermanian and O. Lopez. Single-index copulas. *J. Multivariate Anal.*, 165:27–55, 2018.

[51] J.-D. Fermanian, D. Radulovic, and M. Wegkamp. Weak convergence of empirical copula processes. *Bernoulli*, 10(5):847–860, 2004.

[52] J.-D. Fermanian and M. Wegkamp. Time-dependent copulas. *J. Multivariate Anal.*, 110:19–29, 2012.

[53] H. Fink, Y. Klimova, C. Czado, and J. Stöber. Regime switching vine copula models for global equity and volatility indices. *Econometrics*, 5(1):3, 2017.

[54] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[55] J. Friedman, T. Hastie, R. Tibshirani, and N. Simon. glmnet: Lasso and elastic-net regularized generalized linear models. R package version 2.0–2, 2017.

[56] C. Genest, K. Ghoudi, and L.-P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, 1995.

[57] C. Genest and B. Rémillard. Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 44(6):1096–1127, 2008.

[58] C. Genest, B. Rémillard, and D. Beaudoin. Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and economics*, 44(2):199–213, 2009.

[59] C. J. Geyer. On the asymptotics of convex stochastic optimization. Technical report, Dept. Statistics, Univ. Minnesota, 1996.

[60] I. Gijbels, M. Omelka, and N. Veraverbeke. Estimation of a copula when a covariate affects only marginal distributions. *Scandinavian Journal of Statistics*, 42(4):1109–1126, 2015.

[61] I. Gijbels, M. Omelka, and N. Veraverbeke. Partial and average copulas and association measures. *Electr. J. Statist.*, 9:2420–2474, 2015.

[62] I. Gijbels, M. Omelka, and N. Veraverbeke. Nonparametric testing for no covariate effects in conditional copulas. *Statistics*, 51(3):475–509, 2017.

[63] I. Gijbels, N. Veraverbeke, and M. Omelka. Conditional copulas, association measures and their applications. *Comput. Statist. Data Anal.*, 55(5):1919–1932, 2011.

[64] I. Gijbels, N. Veraverbeke, and M. Omelka. Estimation of a conditional copula and association measures. *Scandin. J. Statist.*, 38:766–780, 2011.

[65] I. Gijbels, N. Veraverbeke, and M. Omelka. Multivariate and functional covariates and conditional copulas. *Electr. J. Statist.*, 6:1273–1306, 2012.

[66] E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6):907–921, 2002.

[67] C. Giraud. *Introduction to high-dimensional statistics*, volume 138. CRC Press, 2014.

[68] I. Hobæk Haff, K. Aas, and A. Frigessi. On the simplified pair-copula construction–simply useful or too simplistic? *J. Multivariate Anal.*, 101:1296–1310, 2010.

[69] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

[70] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58(301):13–30, 1963.

[71] M. Hollander and D. Wolfe. *Nonparametric Statistical Methods*. Wiley, 1973.

[72] S. B. Hopkins. Sub-gaussian mean estimation in polynomial time. *Arxiv preprint, arXiv:1809.07425*, 2018.

[73] J.-J. Hsieh and W.-C. Huang. Nonparametric estimation and test of conditional kendall's tau under semi-competing risks data and truncated data. *Journal of Applied Statistics*, 42(7):1602–1616, 2015.

[74] H. Joe. Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. *Lecture Notes-Monograph Series*, pages 120–141, 1996.

[75] H. Joe. *Multivariate models and multivariate dependence concepts*. Chapman and Hall/CRC, 1997.

[76] H. Joe. *Dependence Modeling with copulas*. Chapman & Hall, 2015.

[77] H. Joe and D. Kurowicka. *Dependence modeling: vine copula handbook*. World Scientific, 2011.

[78] E. Jondeau and M. Rockinger. The copula-garch model of conditional dependencies: An international stock market application. *J. of Internat. Money and Finance*, 25:827–853, 2006.

[79] K. Kato. Asymptotics for argmin processes: Convexity arguments. *Journal of Multivariate Analysis*, 100(8):1816–1829, 2009.

[80] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[81] Y.-J. Kim. Estimation of conditional kendall's tau for bivariate interval censored data. *Communications for Statistical Applications and Methods*, 22(6):599–604, 2015.

[82] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000.

[83] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

[84] V. S. Korolyuk and Y. V. Borovskich. *Theory of U-statistics*. Springer, 1994.

[85] J. Kowalski and X. M. Tu. *Modern applied U-statistics*, volume 714. John Wiley & Sons, 2008.

[86] W. Kruskal. Ordinal measures of association. *J. Amer. Statist. Ass.*, 53(284):814–861, 1958.

[87] M. S. Kurz and F. Spanhel. Testing the simplifying assumption in high-dimensional vine copulas. *Arxiv preprint, arXiv:1706.02338*, 2017.

[88] J. Kwon, G. Lecué, and M. Lerasle. Median of means principle as a divide-and-conquer procedure for robustness, sub-sampling and hyper-parameters tuning. *Arxiv preprint, arXiv:1812.02435*, 2018.

[89] L. Lakhal, L.-P. Rivest, and B. Abdous. Estimating survival and association in a semicompeting risks model. *Biometrics*, 64(1):180–188, 2008.

[90] G. Lecué and M. Lerasle. Learning from mom's principles: Le cam's approach. *Arxiv preprint, arXiv:1701.01961*, 2017.

[91] G. Lecué and M. Lerasle. Robust machine learning by median-of-means: theory and practice. *Arxiv preprint, arXiv:1711.10306*, 2017.

[92] G. Lecué, M. Lerasle, and T. Mathieu. Robust classification via mom minimization. *Arxiv preprint, arXiv:1808.03106*, 2018.

[93] G. Lecué and S. Mendelson. Regularization and the small-ball method i: sparse recovery. *The Annals of Statistics*, 46(2):611–641, 2018.

[94] E. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, 1975.

[95] O. V. Lepski and V. G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, pages 2512–2546, 1997.

[96] M. Lerasle, Z. Szabó, G. Massiot, and E. Moulines. Monk–outlier-robust mean embedding estimation by median-of-means. *arXiv preprint arXiv:1802.04784*, 2018.

[97] A. Liu, Y. Hou, and L. Peng. Interval estimation for a measure of tail dependence. *Insurance: Mathematics and Economics*, 64:294–305, 2015.

[98] G. Lugosi and S. Mendelson. Sub-gaussian estimators of the mean of a random vector. *ArXiv preprint, arXiv:1702.00482*, 2017.

[99] P. Major. An estimate on the supremum of a nice class of stochastic integrals and U-statistics. *Probability Theory and Related Fields*, 134(3):489–537, 2006.

[100] A. K. Manatunga and D. Oakes. A measure of association for bivariate frailty distributions. *Journal of Multivariate Analysis*, 56(1):60–74, 1996.

[101] E. C. Martin and R. A. Betensky. Testing quasi-independence of failure and truncation times via conditional kendall's tau. *Journal of the American Statistical Association*, 100(470):484–492, 2005.

[102] S. Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

[103] S. Mendelson. On multiplier processes under weak moment assumptions. In *Geometric Aspects of Functional Analysis*, pages 301–318. Springer, 2017.

[104] S. Minsker. Uniform bounds for robust mean estimators. *Arxiv preprint, arXiv:1812.03523*, 2018.

[105] T. Nagler and C. Czado. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89, 2016.

[106] R. Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[107] D. Oakes. Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84(406):487–493, 1989.

[108] M. Omelka, N. Veraverbeke, and I. Gijbels. Bootstrapping the conditional copula. *Journal of Statistical Planning and Inference*, 143(1):1–23, 2013.

[109] A. B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007.

[110] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

[111] A. Patton. Estimation of multivariate models for time series of possibly different lengths. *J. Appl. Econometrics*, 21(2):147–173, 2006.

[112] A. Patton. Modelling asymmetric exchange rate dependence. *Internat. Econom. Rev.*, 47(2):527–556, 2006.

[113] I. Pinelis and R. Molzon. Optimal-order bounds on the rate of convergence to normality in the multivariate delta method. *Electronic Journal of Statistics*, 10(1):1001–1063, 2016.

[114] F. Portier and J. Segers. On the weak convergence of the empirical conditional copula under a simplifying assumption. *ArXiv:1511.06544*, 2015.

[115] B. Rémillard and O. Scaillet. Testing for equality between two copulas. *Journal of Multivariate Analysis*, 100(3):377–386, 2009.

[116] A. Rinaldo, L. Wasserman, et al. Generalized density clustering. *The Annals of Statistics*, 38(5):2678–2722, 2010.

[117] B. Ripley. Tree: classification and regression trees. R package version 1.0-39, 2018.

[118] J. P. Romano and M. Wolf. Finite sample nonparametric inference and large sample efficiency. *The Annals of Statistics*, 28(3):756–778, 2000.

[119] A. Sabeti, M. Wei, and R. V. Craiu. Additive models for conditional copulas. *Stat*, 3(1):300–312, 2014.

[120] C. Schellhase and F. Spanhel. Estimating non-simplified vine copulas using penalized splines. *Statistics and Computing*, 28(2):387–409, 2018.

[121] B. Schweizer. Thirty years of copulas. In *Advances in probability distributions with given marginals*, pages 13–50. Springer, 1991.

[122] D. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.

[123] R. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 1980.

[124] A. Shiryaev. *Probability*, volume 95. Springer, 1984.

[125] A. Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

[126] A. Sklar. Random variables, distribution functions, and copulas: a personal look backward and forward. *Lecture notes-monograph series*, pages 1–14, 1996.

[127] M. K. So and C. Y. Yeung. Vine-copula garch model with dynamic conditional dependence. *Comput. Statist. Data Anal.*, 76:655–671, 2014.

[128] F. Spanhel and M. S. Kurz. The partial vine copula: A dependence measure and approximation based on the simplifying assumption. *Arxiv preprint, arXiv:1510.06971*, 2015.

[129] J. Stöber and C. Czado. Regime switches in the dependence structure of multidimensional financial data. *Computational Statistics & Data Analysis*, 76:672–686, 2014.

[130] J. Stoeber, H. Joe, and C. Czado. Simplified pair copula constructions—limitations and extensions. *Journal of Multivariate Analysis*, 119:101–118, 2013.

[131] B. Stucky and S. van de Geer. Sharp oracle inequalities for square root regularization. *Journal of Machine Learning Research*, 18:1–29, 2017.

[132] W. Stute. Conditional U-statistics. *Ann. Probab.*, 19(2):812–825, 1991.

[133] W. Su and E. Candes. Slope is adaptive to unknown sparsity and asymptotically minimax. *Annals of Statistics*, 44(3):1038–1068, 2016.

[134] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, pages 1–20, 2012.

[135] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[136] W.-Y. Tsai. Testing the assumption of independence of truncation time and failure time. *Biometrika*, 77(1):169–177, 1990.

[137] H. Tsukahara. Semiparametric estimation in copula models. *Canadian Journal of Statistics*, 33(3):357–375, 2005.

[138] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[139] A. W. Van Der Vaart and J. A. Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

[140] T. Vatter and V. Chavez-Demoulin. Generalized additive models for conditional dependence structures. *Journal of Multivariate Analysis*, 141:147–167, 2015.

[141] N. Veraverbeke, M. Omelka, and I. Gijbels. Estimation of a conditional copula and association measures. *Scand. J. Stat.*, 38(4):766–780, 2011.

[142] Y.-C. Wang, J.-L. Wu, and Y.-H. Lai. A revisit to the dependence structure between the stock and foreign exchange markets: A dependence-switching copula approach. *Journal of Banking & Finance*, 37(5):1706–1719, 2013.

[143] M. J. Wurm, P. J. Rathouz, and B. M. Hanlon. Regularized ordinal regression and the ordinalNet R package. *Arxiv preprint„ arXiv:1706.05003*, 2017.

[144] X. Zeng and M. A. T. Figueiredo. The ordered weighted ℓ_1 norm: Atomic formulation, projections, and algorithms. *Arxiv preprint„ arXiv:1409.4271*, 2014.

[145] J. X. Zheng. A consistent test of conditional parametric distributions. *Econometric Theory*, 16(5):667–691, 2000.

[146] H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre: Contributions à l'analyse statistique des modèles de dépendance en grande dimension

Mots clés: Copules conditionnelles, statistique en grande dimension, distribution conditionnelle

Résumé: Cette thèse peut être divisée en trois parties. Dans la première partie, nous étudions des méthodes d'adaptation au niveau de bruit dans le modèle de régression linéaire en grande dimension. Nous prouvons que deux estimateurs à racine carrée, peuvent atteindre les vitesses minimax d'estimation et de prédiction. Nous montrons qu'une version similaire construite à partir de médianes de moyennes, peut encore atteindre les mêmes vitesses optimales en plus d'être robuste vis-à-vis de l'éventuelle présence de données aberrantes.

La seconde partie est consacrée à l'analyse de plusieurs modèles de dépendance conditionnelle. Nous proposons plusieurs tests de l'hypothèse simplificatrice qu'une copule conditionnelle est constante vis-à-vis de son évènement conditionnant, et nous prouvons la consistance d'une technique de rééchantillonage semi-paramétrique. Si la copule conditionnelle n'est pas constante par rapport à sa variable conditionnante, alors elle peut être modélisée via son tau de Kendall conditionnel. Nous étudions donc l'estimation de ce paramètre de dépendance conditionnelle sous 3 approches différentes : les techniques à noyaux, les modèles de type régression et les algorithmes de classification.

La dernière partie regroupe deux contributions dans le domaine de l'inférence. Nous comparons et proposons différents estimateurs de fonctionnelles conditionnelles régulières en utilisant des U-statistiques. Finalement, nous étudions la construction et les propriétés théoriques d'intervalles de confiance pour des ratios de moyennes sous différents choix d'hypothèses et de paradigmes.

Title: Some statistical results in high-dimensional dependence modeling

Keywords: Conditional copulas, high-dimensional statistics, conditional distribution

Abstract: This thesis can be divided into three parts. In the first part, we study adaptivity to the noise level in the high-dimensional linear regression framework. We prove that two square-root estimators attains the minimax rates of estimation and prediction. We show that a corresponding median-of-means version can still attains the same optimal rates while being robust to outliers in the data.

The second part is devoted to the analysis of several conditional dependence models. We propose some tests of the simplifying assumption that a conditional copula is constant with respect to its conditioning event, and prove the consistency of a semiparametric bootstrap scheme. If the conditional copula is not constant with respect to the conditional event, then it can be modelled using the corresponding Kendall's tau. We study the estimation of this conditional dependence parameter using 3 different approaches : kernel techniques, regression-type models and classification algorithms.

The last part regroups two different topics in inference. We review and propose estimators for regular conditional functionals using U-statistics. Finally, we study the construction and the theoretical properties of confidence intervals for ratios of means under different sets of assumptions and paradigms.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France