

Multivariate statistics for dietary risk analysis

Emilie Chautru

▶ **To cite this version:**

Emilie Chautru. Multivariate statistics for dietary risk analysis. Statistics [math.ST]. Télécom ParisTech, 2013. English. NNT : 2013ENST0045 . tel-01306948

HAL Id: tel-01306948
https://pastel.hal.science/tel-01306948

Submitted on 25 Apr 2016

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.

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2013-ENST-0045

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et Image »

présentée et soutenue publiquement par

Emilie CHAUTRU

le 6 septembre 2013

Statistiques multivariées pour l'analyse du risque alimentaire

Directeur de thèse : **Stéphan CLÉMENÇON**
Co-encadrement de la thèse : **Jean-Luc VOLATIER**

Jury

Mme Anne RUIZ-GAZEN, Professeur, GREMAQ, Toulouse School of Economics	Rapporteur
M. Johan SEGERS, Professeur, ISBA, Université catholique de Louvain	Rapporteur
Mme Liliane BEL, Professeur, MIA, AgroParisTech	Examinateur
Mme Anne-Laure FOUGERES, Professeur, CNRS UMR 5208, Université Lyon 1	Examinateur
M. Valentin PATILEA, Professeur, LSM Crest, Ensai	Examinateur
M. Paul DOUKHAN, Professeur, AGM, Université de Cergy-Pontoise	Invité

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

THÈSE

FINANCEMENTS INRA-ANSES-CREST







Inra - Unité Met@risk
Anses - Direction de l'évaluation des risques
Crest - Ensai - Laboratoire de statistiques d'enquête

ABSTRACT - RÉSUMÉ

ABSTRACT

Dietary risk analysis is a multidisciplinary field par excellence, requiring in particular a mixture of biological, sociological, chemical, cultural, economic, statistical and sanitary expertise to answer practical issues. In a global system where international exchanges are encouraged, where mass production favors cheap, profitable production strategies (GMO, pesticides, enriched animal food, addition of colorants, preservatives or artificial flavors, etc.), it is necessary to quantify the risks that result from such economic behaviors. The focus here is on the chronic (year-long) exposure to a set of food contaminants, the long-term toxicity of which is already well-known. Since food consumption is also the privileged way of supplying the body with necessary nutrients, nutritional benefits (or deficiencies) are taken into account as well. This thesis is dedicated to a collection of mathematical problems arising from both the analysis of such dietary risks and the nature of the data. First of all, the adequacy of the classical univariate long-term models that are recommended by Efsa (European Food Safety Authority) is discussed at length. They usually assume that observations have a log-normal distribution, thereby neglecting the possibility that they are in fact heavy-tailed. In terms of food risks, this would mean that a very high exposure to a single chemical element is a rare enough event to be neglected. When the component of interest clearly violates this strong assumption, extreme value theory is proposed as a relevant alternative, as revealed by a set of illustrations based on real data. Then, this sub-field of theoretical statistics is adapted to the analysis of the simultaneous exposure to many nutrients and contaminants. Following in the footsteps of classical machine learning techniques, using in particular the recent "Principal Nested Spheres" algorithm of S. Jung, I.L. Dryden and J.S. Marron, we construct a new model that identifies extreme dependencies in high dimension. This allows us to define some cocktails of chemicals that are jointly ingested in very high quantities. Remaining in a multivariate framework, we then move to the issue of dietary recommendations. In the line of the "minimum volume set" approach of C. Scott and R. Nowak, we propose an algorithm that selects food baskets realizing a compromise between toxicological risk and nutritional benefit. Finally, as consumption databases often result from complex survey schemes, the estimators constructed under the hypothesis that observations are independent and identically distributed can produce severely biased outcomes. In an attempt to take into account this preliminary sampling phase, we mimic the approaches of J. Hàjek and Y. Berger and focus on the specific family of Poisson-like survey plans. Under this framework, the asymptotic properties of Horvitz-Thompson empirical processes are inspected, before concluding this thesis on the introduction of a weighted version of the widely celebrated Hill estimator for the heavy-tail analysis of sampled observations.

RÉSUMÉ

Véritable carrefour de problématiques économiques, biologiques, sociologiques, culturelles et sanitaires, l'alimentation suscite de nombreuses polémiques. Dans un contexte où les échanges mondiaux facilitent le transport de denrées alimentaires produites dans des conditions environnementales diverses, où la consommation de masse encourage les stratégies visant à réduire les coûts et maximiser le volume de production (OGM, pesticides, nourriture enrichie donnée aux animaux en élevage, ajout de substances chimiques tels les colorants et les arômes artificiels, etc.) il devient nécessaire de quantifier les risques sanitaires que de tels procédés engendrent. Notre intérêt se place ici sur l'étude de l'exposition chronique, de l'ordre de l'année, à un ensemble quelconque de contaminants dont la nocivité à long terme est d'ores et déjà établie. Les dangers et bénéfices de l'alimentation ne se restreignant pas à l'ingestion ou non de substances toxiques, nous ajoutons à nos objectifs l'étude de certains apports nutritionnels. Les travaux réalisés au cours de cette thèse répondent à plusieurs problématiques mathématiques engendrées par l'analyse de ce type de risque alimentaire et par la nature des données utilisées à ces fins. Dans un premier temps, nous discutons la portée des modèles classiques dédiés à l'estimation de l'exposition de long terme à un unique composant alimentaire. Ces derniers, dont l'utilisation est recommandée par l'Efsa (European Food Safety Authority), requièrent généralement la log-normalité des observations et interdisent par là-même les distributions à queue plus épaisse, où l'élément chimique étudié aurait une probabilité non négligeable d'être ingéré en de très grandes quantités. Nous montrons alors, exemples à l'appui, comment la théorie des valeurs extrêmes peut être utilisée dans de telles situations. Ce pan de théorie statistique est ensuite adapté à l'étude des fortes expositions à un nombre quelconque (potentiellement très grand) de nutriments et contaminants. En nous inspirant de techniques traditionnelles de l'apprentissage statistique, notamment le récent algorithme des "Principal Nested Spheres" développé par S. Jung, I.L. Dryden et J.S. Marron, nous construisons un modèle ouvrant l'étude des dépendances extrêmes à la grande dimension, qui nous permet en particulier de définir des groupes d'éléments auxquels les consommateurs sont simultanément sur-exposés. Toujours dans une optique multivariée, nous nous éloignons ensuite des expositions extrêmes pour nous placer du côté des consommations alimentaires. En nous basant sur une approche de type "ensemble de volume minimum" comme introduite par C. Scott et R. Nowak, nous proposons un algorithme pour identifier des paniers de produits réalisant un compromis entre risque toxicologique et bénéfice nutritionnel. Enfin, les données alimentaires disponibles résultant souvent de plans de sondage non triviaux, les estimateurs construits sous l'hypothèse que les données sont indépendantes et identiquement distribuées peuvent produire des résultats biaisés. Tentant de prendre en compte cette étape préliminaire d'échantillonnage, nous nous concentrons dans la lignée des travaux de J. Hàjek et Y. Berger sur la famille des plans de sondage de type Poisson et étudions le comportement asymptotique des processus empiriques pondérés à la manière Horvitz-Thompson. Suivant la même approche, nous proposons finalement une variante de l'estimateur de Hill pour l'analyse des distributions à queue épaisse adaptée au cadre des données de sondage.

PUBLICATIONS

submitted for publication

E. Chautru, S. Clémençon, Dimension reduction in multivariate heavy-tail analysis, 2013

P. Bertail, E. Chautru, S. Clémençon, Empirical Processes in Survey Sampling, 2013

P. Bertail, E. Chautru, S. Clémençon, On Tail Index Estimation Based on Survey Data, 2013

to be submitted

E. Chautru, S. Clémençon, J. Tressou, Determination of an optimal diet by minimum volume sets, 2013

E. Chautru, J.L. Volatier, Heavy-tail modeling of Usual Intakes, 2013

REMERCIEMENTS

En premier lieu, je souhaite vivement remercier les membres de mon jury d'avoir accepté d'évaluer mon travail et de me faire profiter par là-même de leurs conseils avisés. Je suis tout particulièrement reconnaissante à Johan Segers et Anne Ruiz-Gazen d'avoir sacrifié une partie de leur mois d'août pour éplucher mon mémoire en détail. Merci de l'intérêt que vous avez porté à ma thèse et merci pour les critiques constructives qui m'ont permis d'en améliorer la substance. Je suis de même très honorée par la présence dans mon jury de thèse de Liliane Bel, Anne-Laure Fougères, Valentin Patilea et Paul Doukhan, que j'admire personnellement tant sur le plan scientifique que sur le plan humain.

Il va sans dire que les lauriers devraient revenir à Jean-Luc et Stéphan, mes deux directeurs de thèse. Merci à tous deux de m'avoir introduite au risque alimentaire, l'un des champs d'étude les plus interdisciplinaires que j'aie eu l'occasion de découvrir, ainsi que de m'avoir suivie sans relâche durant ces quatre longues années. Votre collaboration aura généré un équilibre entre thématiques théoriques et appliquées dans ma recherche, point qui me tenait à cœur lorsque je me suis décidée à démarrer un doctorat et que j'ai pu réaliser grâce à vous.

Je me dois aussi de reconnaître la partie prenante qu'ont pris dans mon parcours ceux avec qui j'ai écrit la plus grande partie des papiers présentés dans cette thèse : Patrice et Jessica. Merci à tous deux pour votre gentillesse et votre patience, pour tout le temps et l'énergie que vous m'avez consacrés et qui m'ont permis d'avancer.

Bien entendu, cette thèse n'aurait pu avoir lieu sans le soutien financier de l'Inra, de l'Anses et du Crest-Ensai. Etre ainsi rattachée à un nombre non négligeable d'unités de recherche s'est avéré des plus stimulant et enrichissant d'un point de vue intellectuel. A commencer par Met@risk, où j'ai pu profiter des conseils pratiques de l'équipe informatique, découvrir l'aspect sociologique de la nutrition au contact des sociologues, partager avec les statisticiens, etc. . Un grand merci des plus enthousiastes à tous ceux que j'ai rencontrés au sein de cette unité formidable, Laura, Lydie, Damien, Stéphane, Juliette, Sandrine, Eve, Isabelle, Jessica, Mélanie, Clémence, Max et les autres.

Du côté de l'Anses, je souhaite en particulier remercier Ariane, Carine, Véronique, Amélie et Camille, non seulement pour leur accueil chaleureux dans un milieu aux codes quelque peu différents de ceux de la recherche universitaire, mais aussi pour tous leurs conseils et leur expertise sur les bases de données que j'ai utilisées dans ma thèse et les enjeux mathématiques inhérents au risque alimentaire.

Plus loin en Bretagne, dans la campagne Bruzoise, où les Ensaiens ont fait preuve d'une hospitalité remarquable, je tiens à dire un grand merci à Eric et Valentin pour m'avoir soutenue lors des auditions et permis de rejoindre les Rennais en cette dernière année de thèse. Merci aussi à François, Guillaume, Brigitte, David, Nicolas, Nicolas, Gaspar, Laurent, Stéphane, Lionel, Magalie, Myriam et tous les permanents du Crest auprès de qui j'ai passé une excellente année. N'oublions pas les Insee, tout particulièrement Philippe et Jean-Michel,

grâce à qui j'ai pu payer mon loyer chaque mois. Je finirai cette liste par mes deux compagnons de route, Cyril et Samuel, sans qui Rennes aurait été bien morne. Merci à vous deux pour votre amitié et votre soutien inconditionnel durant cette période éprouvante.

Enfin, à Télécom, je souhaite remercier une fois de plus Eric pour son aide lors de la recherche de financements supplémentaires pour terminer mon doctorat. Merci à Jérémie, co-bureau des débuts qui m'aura fait prendre un grand recul sur la place que je peux occuper dans la recherche et aux permanents qui m'ont aiguillée ; à Sarah, Steffen, Olaf, Malika, Marine, Anne-Laure, Alexandre, Jeoffrey, Tabea, Onur, les vieux de la vieille qui sont partis beaucoup trop tôt à mon goût ; à toute la bande des thésards, Emilie, Amandine, Olivier, Cristina, Sylvain, Yao, Nicolas, Benjamin, Aymeric, Sébastien et des post-doc, Charanpal, Julien, Romaric, j'en oublie et des meilleurs. Une pensée toute particulière pour mes frères de thèse, Sylvain, Ruocong, Antoine, Eric, ainsi que pour Billy et Billy ; Dareau manque sérieusement de chevelus déjantés parfumés à la bière belge depuis votre départ.

Beaucoup de personnes extérieures à mes nombreux lieux de travail ont aussi eu un impact non négligeable sur ma recherche. Eric, qui m'a écoutée des heures durant parler de mathématiques un kir à la main, qui m'a épaulée, conseillée, guidée tout au long de ce parcours, je ne te dirai jamais assez merci ; Olivier, qui m'a introduite auprès de nombreux chercheurs et encouragée dans ma recherche ; Xavier, sans qui je me serais sentie bien seule lors de ma toute première conférence ; Julyan, Ronan, Emmanuelle, avec qui j'ai eu le plaisir de donner des TD et qui m'ont aidée à prendre confiance en moi ; M. Mano, mon professeur de mathématiques en hypokhâgne et khâgne, qui m'a donné le goût de la matière. Merci aussi à Céline et Meriem et à tous les anciens de l'Ensae, Miki, Adélaïde, Marion, Arnaud, Max, Nathanaël, Maha, Sophie, Alexandra, Simon, la JE et le forum, avec qui j'ai découvert et pris goût aux statistiques. Tout particulièrement merci à Stéphanie, une des meilleures partenaires de mémoire (et de sorties), que j'encourage vivement à quitter l'Insee pour monter une agence matrimoniale. Sans oublier les anciens d'Abu Dhabi, Nassime, Antoine, Pohe, Pierre, Cédric, toujours fidèles au poste plus de dix ans après nos aventures lycéennes.

Je remercie également toute ma famille, et plus encore mes parents, mon frère et ma sœur (à qui je dois quelques uns des dessins de mes slides), pour m'avoir toujours soutenue et encouragée, que ce soit dans mes études, ma vie professionnelle ou ma vie privée. Ce travail n'existe que grâce à vous (notamment parce que vous avez pensé à regarder les résultats d'admissibilité à l'Ensae alors que j'avais oublié que je m'étais inscrite au concours). Isabelle et Mélanie, je vous dis merci d'être mes amies depuis plus de 20 ans et de continuer encore et toujours d'occuper une place importante dans ma vie.

Enfin, à Guillaume, je dis merci et re-merci. Merci de m'avoir redonné le goût des mathématiques à un moment où je n'y croyais plus, merci d'avoir relu l'intégralité de mon mémoire, merci de m'avoir supportée, clope au bec, durant les moments les plus stressants, la liste est longue... But most of all, thank you for being yourself.

CONTENTS

Preface 1

p régimes alimentaires, exposition individuelle et statistique 3

p.1 Qu'est-ce que le risque alimentaire? . . . 4

p.1.1 Evaluation du risque . . . 4

p.1.2 Gestion du risque . . . 5

p.1.3 Communication autour du risque . . . 5

p.2 Des sources multiples d'information . . . 6

p.2.1 Données de consommation . . . 6

p.2.2 Données de composition . . . 9

p.3 Estimation des risques alimentaires : petit aperçu de l'état de l'art . . . 11

p.3.1 Exposition extrême de long terme . . . 12

p.3.2 Exposition simultanée à un ensemble de nutriments et contaminants . . . 12

p.3.3 Données de sondage . . . 13

p.4 Objectifs et contributions principales . . . 13

p.4.1 Estimation de l'exposition extrême de long terme à un unique élément chimique . . . 14

p.4.2 Evaluation de la sur-exposition simultanée à un cocktail de composants alimentaires . . . 15

p.4.3 Ensembles de volume minimum et approche risque-bénéfice . . . 16

p.4.4 Processus empiriques dans le cadre des sondages . . . 16

p.4.5 Estimation de l'indice de valeurs extrêmes à partir de données de sondage . . . 17

Thesis 19

1 introduction: of dietary habits and exposure 21

1.1 Getting a taste of dietary risk analysis . . . 22

1.1.1 Risk evaluation . . . 22

1.1.2 Risk management . . . 22

1.1.3 Risk communication . . . 23

1.2 Collecting information: multiple sources, multiple difficulties . . . 23

1.2.1 Consumption data . . . 24

1.2.2 Composition data . . . 26

1.3 Statistical assessment of dietary risks: a review of the literature . . . 28

1.3.1 Extreme long-term exposure . . . 29
1.3.2 Simultaneous exposure to multiple food components . . . 29
1.3.3 Survey data . . . 30
1.4 Objectives and main contributions . . . 30
1.4.1 Extreme chronic exposure to one chemical . . . 31
1.4.2 Simultaneous over-exposure to many food chemicals . . . 31
1.4.3 A minimum volume set approach to dietary risk-benefit analysis 32
1.4.4 Empirical processes in survey sampling . . . 32
1.4.5 Tail index estimation based on survey data . . . 33
2 extreme chronic exposure to one chemical 35
2.1 Introduction . . . 35
2.2 Mixed-effects models to assess usual intakes . . . 36
2.2.1 The BetaBinomial-Normal model . . . 37
2.2.2 The LogisticNormal-Normal model . . . 38
2.3 Heavy-tail modeling of usual intakes . . . 39
2.3.1 Extreme value theory and heavy-tailed distributions . . . 39
2.3.2 Testing the heavy-tail assumption . . . 41
2.3.3 Assessing extreme quantities . . . 43
2.4 Case study: tails of iron, zinc, calcium and retinol . . . 46
2.4.1 Description of the data . . . 46
2.4.2 Results . . . 46
2.5 Discussion . . . 51
2.6 Supplements: on the second order parameter □ . . . 51
3 simultaneous over-exposure to many food chemicals 55
3.1 Statistical challenges and objectives . . . 55
3.2 Hypotheses and notations . . . 57
3.2.1 Notations . . . 58
3.2.2 General setting and main hypotheses . . . 58
3.3 A mixture model under multivariate regular variation . . . 60
3.3.1 Exponent and spectral (probability) measures . . . 60
3.3.2 Mixture model of the spectral probability measure . . . 61
3.3.3 Latent variables representation . . . 63
3.4 Statistical inference . . . 65
3.4.1 Preliminaries . . . 66
3.4.2 Dimension reduction and clustering . . . 66
3.4.3 Identifying groups of asymptotically dependent variables . . . 68
3.5 Numerical experiments . . . 70
3.5.1 Settings . . . 70

3.5.2 Results . . . 71
3.6 Application to dietary risk assessment . . . 73
3.6.1 Data and required assumptions . . . 74
3.6.2 Analysis of extreme dependencies . . . 74
3.7 Discussion . . . 77
3.8 Proofs and supplements . . . 79
3.8.1 Intra-class regular variation . . . 79
3.8.2 Face-characterizing functional . . . 83
3.8.3 About $\bar{p}_{j,\vee}(k)$ . . . 86
3.8.4 A little more on PNS and spherical k-means . . . 90
4 a minimum volume set approach to dietary risk-benefit analysis 99
4.1 Theoretical analysis and methods . . . 100
4.1.1 Assessment of dietary exposure to chemicals and nutrients: a MV-set formulation . . . 100
4.1.2 Uniform approximation of generalized U-statistics by their incomplete versions . . . 102
4.1.3 Empirical MV-set estimation based on hypercubes . . . 107
4.1.4 Optimal regions in the consumption space . . . 108
4.2 Proofs and supplements . . . 112
4.2.1 Maximal deviation . . . 112
4.2.2 Maximal deviation in dietary risk analysis . . . 115
4.2.3 Optimal dietary habits . . . 115
5 empirical processes in survey sampling 117
5.1 Background and Preliminaries . . . 118
5.1.1 Survey sampling: some basics . . . 118
5.1.2 Empirical process indexed by classes of functions . . . 124
5.2 Empirical process in survey sampling . . . 127
5.2.1 The Horvitz-Thompson empirical process . . . 127
5.2.2 Alternative estimate in the Poisson sampling case . . . 128
5.3 Asymptotic results . . . 129
5.3.1 Limit of the empirical process for the Poisson survey scheme . . . 130
5.3.2 The case of rejective sampling . . . 133
5.4 Application to non-parametric statistics . . . 135
5.4.1 Hadamard differentiable functionals . . . 136
5.4.2 Fréchet differentiable functionals . . . 137
5.4.3 Simulation-based Gaussian asymptotic confidence regions . . . 138
5.5 Proofs and supplements . . . 143
5.5.1 Limit of the covariance operator . . . 143
5.5.2 FCLT in the Poisson survey case . . . 144
5.5.3 Approximation result . . . 146
5.5.4 FCLT in the rejective survey case . . . 147

5.5.5 CLT for Hadamard differentiable functionals 148
6 tail index estimation based on survey data 149
6.1 Background and Preliminaries 150
6.1.1 Survey sampling 150
6.1.2 Tail index inference - the Hill estimator 152
6.2 The Hill estimator in survey sampling 153
6.2.1 The Horvitz-Thompson variant of the Hill estimator 154
6.2.2 Consistency of $H^{\square}_{K;N}$ 155
6.3 Asymptotic normality of $H^{\square}_{K;N}$ 157
6.3.1 The case of the Poisson survey scheme 159
6.3.2 Extension to rejective sampling schemes 161
6.4 Practical Issues and Illustrative Experiments 163
6.4.1 On the choice of an optimal k 163
6.4.2 Numerical experiments 164
6.5 Discussion 168
6.6 Proofs and supplements 169
6.6.1 Consistency of the Horvitz-Thompson variant of the Hill estimator 169
6.6.2 Limit distribution of $H^{p}_{K;N}$ in the Poisson survey case 172
6.6.3 Limit distribution of $H^{\square}_{K;N}$ in the rejective survey case 181

bibliography 184

ACRONYMS

institutes

Anses	French agency for food, environmental and occupational health safety
Efsa	European food safety authority
FAO	Food and Agriculture Organization
Inra	French national institute for agricultural research
Insee	French national institute of statistics and economic studies
Unep	United Nations Environment Program
WHO	World Health Organization

dietary risk analysis abbreviations

DIL	Dietary intake limit
CIQUAL	Centre d'information sur la qualité des aliments
FFQ	Food frequency questionnaire
FPQ	Food propensity questionnaire
GMO	Genetically Modified Organism
INCA	Etude nationale des consommations alimentaires
TDS	Total dietary survey
24H	24-hour recall

chemical symbols

Ca	Calcium
Cd	Cadmium
Fe	Iron
MeHg	Methylmercury
Na	Sodium
PCB-DL	Dioxin-like polychlorinated biphenyls
Re	Retinol
Zn	Zinc

mathematical abbreviations

cdf	Cumulative distribution function
EVT	Extreme value theory
iid	Independent and identically distributed
rv	Random variable
svf	Slowly varying function

PREFACE

RÉSUMÉ EN FRANÇAIS DES TRAVAUX DE THÈSE

P

RÉGIMES ALIMENTAIRES, EXPOSITION INDIVIDUELLE ET STATISTIQUE

Variées et nombreuses sont les études scientifiques pouvant être caractérisées comme analyses du risque alimentaire. Cette désignation fort générale issue du jargon de la santé publique fait référence à toute entreprise visant à la détection, à la compréhension, et au traitement des dangers liés à l'alimentation. Parmi ces derniers, citons par exemple l'une des problématiques classiques des producteurs : le développement de bactéries au cours des différentes étapes de production, de transport, de distribution et de stockage qui précèdent la consommation (Rigaux et al., 2012). A une toute autre échelle, l'industrie agroalimentaire doit évaluer les impacts à la fois économiques et sanitaires de techniques de production de masse tels l'utilisation de pesticides (http://www.efsa.europa.eu/en/topics/topic/pesticides.htm) ou l'introduction d'organismes génétiquement modifiés dans les plantations (http://www.efsa.europa.eu/en/topics/topic/gmo.htm). Selon leur composition chimique et leur mode d'utilisation, les ustensiles de cuisine peuvent aussi représenter une source de danger, comme en témoigne la récente polémique à propos de la présence de bisphénol A dans les biberons (cf. l'avis de l'Anses y ayant fait suite, Anses, 2013). D'autres problèmes de santé peuvent être engendrés par des troubles du comportement alimentaire (e.g. les personnes souffrant d'anorexie aiguë sont particulièrement sujettes aux maladies provoquées par des carences nutritionnelles) ou par des facteurs biologiques particuliers (e.g. le diabète). A la lumière de ces quelques exemples, il est clair que selon la question examinée, des connaissances spécifiques en chimie, biologie, médecine, sociologie, économie ou même psychologie peuvent être requises, sans oublier la modélisation probabiliste et la gestion informatique des données. L'analyse du risque alimentaire est ainsi un domaine d'étude multidisciplinaire par excellence. Le lecteur curieux d'en apprendre plus est invité à consulter Feinberg et al. (2006), manuel expliquant de manière exhaustive les tenants et aboutissants de l'analyse du risque alimentaire selon un point de vue interdisciplinaire.

Le présent travail est dédié à l'étude d'un type spécifique de risque alimentaire : indépendamment de tout processus de production, de stockage ou de cuisson, ignorant les prédispositions biologiques extraordinaires, nous nous intéressons exclusivement à la très forte (ou très faible) exposition sur le long terme à certains composants alimentaires de la population française dans son ensemble. Après avoir introduit quelques concepts élémentaires du risque alimentaire en section P.1, décrit les données disponibles en section P.2 et brièvement exposé l'état de l'art sur la modélisa-

tion probabiliste des risques alimentaires en section P.3, nous présentons en détail en section P.4 les diverses problématiques traitées au cours de cette thèse avant d'en annoncer les contributions scientifiques principales.

P.1 qu'est-ce que le risque alimentaire?

Une fois qu'un type spécifique de risque alimentaire a été porté à la connaissance de tous, son analyse consiste en trois étapes distinctes, respectivement qualifiées d'évaluation du risque, de gestion du risque et de communication autour du risque. Les divers problèmes soulevés dans chacune de ces phases de recherche sont résumés dans les paragraphes suivants. Afin d'en faciliter la compréhension, nous en illustrons la substance à l'aide de l'exemple concret des régimes riches en sel, dont l'abus récurrent peut favoriser les problèmes cardiaques (se référer par exemple à la page Internet `http://www.anses.fr/en/content/salt` et aux références qui y sont mentionnées).

P.1.1 Evaluation du risque

L'évaluation (ou appréciation) du risque peut elle-même être décomposée en quatre sous-étapes, qui consistent à successivement identifier puis caractériser les dangers potentiels et leur probabilité d'apparition. Dans le cas des régimes riches en sel, le danger n'est autre que celui d'ingérer de manière journalière ou hebdomadaire une trop grande quantité de sodium (Na dans le tableau périodique des éléments), ce qui sur le long terme pourrait engendrer une dégradation du système cardiovasculaire. La connaissance de ces effets nocifs est due à tout un ensemble d'études chimiques et biologiques concernant l'assimilation, l'action et l'élimination du sodium dans le corps humain. Ce processus, appelé identification du danger (Barlow et al., 2002), est suivi de recherches supplémentaires permettant de définir à partir de quel niveau de consommation le composant devient toxique, c'est-à-dire de caractériser le danger. Pour ce faire, des tests dits dose-réponse sont en général réalisés in vitro ou in vivo sur des animaux avant d'en étendre les résultats à l'espèce humaine par le biais de modèles dédiés (Dybing et al., 2002). Une fois le processus de contamination compris dans son intégralité, il devient possible d'en estimer la fréquence d'apparition au sein d'une population choisie. Cette dernière phase requiert en premier lieu la description détaillée des distributions de la consommation et de l'exposition dans la population concernée (Kroes et al., 2002), avant de les comparer à des doses maximales recommandées, déterminées par des experts à l'issue des tests dose-réponse précédemment évoqués (Renwick et al., 2003) et possiblement raffinés par des procédés mathématiques (Edler et al., 2002). En fonction de la nature des données disponibles, les modèles statistiques utilisés à ces fins peuvent inclure une dimension temporelle (Bertail

et al., 2010, 2008; Allais and Tressou, 2009), prendre en compte des caractéristiques comportementales ou biologiques individuelles, ou encore tenter de compenser le manque d'information sur les habitudes alimentaires de long terme (Dodd et al., 2006). Ce dernier problème est discuté en détail en section P.3 et au chapitre 2.

P.1.2 Gestion du risque

Après que le risque a été évalué, il est nécessaire d'en identifier les déterminants, d'en évaluer les impacts relatifs et de définir la stratégie la plus rapide et la plus efficace permettant de le réduire. Ce n'est autre que l'objectif de la gestion du risque. Par exemple, la sur-exposition au sodium peut être attribuée à la consommation répétée de produits à haute teneur en sel comme les plats tout prêts ou les biscuits. Un plan simple d'action dans ce cas serait de prévenir les consommateurs des dangers potentiels et d'imposer en parallèle aux compagnies agroalimentaires de limiter les doses de sel ajoutées à leurs préparations. Dans des situations plus extrêmes, après une évaluation rigoureuse des conséquences économiques d'une telle opération, des produits considérés trop dangereux peuvent même être retirés du marché.

P.1.3 Communication autour du risque

La dernière étape de l'analyse du risque alimentaire est appelée communication autour du risque. Elle peut être mise en place à n'importe quel moment de l'analyse et peut s'adresser aussi bien aux scientifiques et aux gestionnaires du risque qu'à l'industrie agroalimentaire ou à la population. En tant que telle, elle n'est pas cantonnée aux campagnes publicitaires de santé publique du type "évitez de manger trop salé". S'y rapportent aussi tous les rapports scientifiques concernant les procédures considérées appropriées pour détecter et quantifier les risques, la publications d'études cas-témoin, les discussions internationales sur les politiques agroalimentaires, etc. Dans cette thèse nous sommes tout particulièrement intéressés par la production de résultats permettant d'aiguiller les recherches futures et, lorsque possible, de définir des lignes directrices de consommation simples et générales à l'intention de la population.

La grande majorité de nos travaux correspond ainsi à la phase d'évaluation du risque. En quelques mots, nous proposons des méthodes statistiques visant à estimer certaines caractéristiques (comme des quantiles) de la très forte exposition à un ou plusieurs composants alimentaires sur une longue période de temps. A l'exception près du chapitre 4, dans lequel nous élaborons une procédure pour déterminer des paniers de consommation réalisant un compromis entre risque toxicologique et bénéfice nutritionnel, les problématiques relevant de la gestion du risque ne sont pas abordées. Les éléments chimiques que nous prenons en considération sont les nutri-

ments et contaminants dont les effets sanitaires liés à une sur- ou une sous-exposition chronique ont d'ores et déjà été établis. En particulier, les risque aigus sont ignorés, tels ceux impliquant la contamination bactérienne de la nourriture, qui peuvent affecter l'organisme en quelques jours ou même quelques heures seulement. L'analyse statistique est réalisée dans ce contexte à partir de formats standards de bases de données, dont une description rapide est proposée ci-après.

P.2 des sources multiples d'information

Afin de permettre l'analyse statistique de l'exposition chronique d'une population donnée à une collection de composants alimentaires, il est désirable d'observer de tels types d'exposition sur un large échantillon d'individus pendant une longue période de temps. Malheureusement, les quantités de nutriments et de contaminants ingérés durant un repas ne sont pas directement mesurables ; ils ne peuvent qu'être estimés à l'aide de méthodes variées. L'une des procédures classiques consiste à détecter puis quantifier certains marqueurs biologiques, dont la présence est intimement liée au niveau d'exposition (voir par exemple la page Internet de l'Anses http://www.anses.fr/sites/default/files/documents/RSC1205-DossierParticipants.pdf, où sont listées les présentations d'un colloque de mai 2012 dédié à ce sujet). D'un point de vue pratique, cela nécessite de collecter puis d'analyser la composition chimique d'échantillons de fluides corporels, de cheveux ou de peau d'un nombre important de personnes. Au niveau national, le coût de telles enquêtes peut contraindre la quantité de sondés au détriment de l'efficacité statistique. En outre, de nombreux éléments chimiques peuvent être assimilés par l'intermédiaire d'autres substances que les aliments (e.g. l'air), or les marqueurs biologiques ne permettent pas de distinguer les diverses sources d'exposition (Sirot et al., 2009). Pour pouvoir étudier exclusivement l'impact de l'alimentation sur l'exposition individuelle, il est possible d'utiliser une méthode alternative où deux bases de données sont combinées, la première listant les habitudes alimentaires d'un ensemble de consommateurs et la seconde indiquant les teneurs en composants chimiques d'une nomenclature fine de produits. Les principales caractéristiques de ces données, respectivement dites de consommation et de composition, sont décrites dans les paragraphes suivants.

P.2.1 Données de consommation

De nombreux types de bases de données peuvent être utilisées pour évaluer l' exposition chronique de la population française à des nutriments et des contaminants alimentaires, allant de cohortes de ménages dont les dépenses en nourriture sont strictement suivies (Secodip, Nichèle et al., 2008) à des sondages en ligne remplis sur la base du volontariat (Nutrinet, Hercberg et al., 2010) en passant par des enquêtes

institutionnelles dédiées à un niveau national (INCA2, Afssa, 2009). Nous nous intéressons ici à deux grandes familles de données, appelées en anglais 24-hour recalls (24H) et food frequency ou propensity questionnaires (FFQ/ FPQ). La première de ces catégories correspond à des enquêtes de grande envergure (par exemple nationale) où un échantillon de sondés note en détail les quantités d'aliments qu'ils consomment durant 2 à 7 jours, parfois consécutifs, mais la plupart du temps sélectionnés aléatoirement dans l'année. Les prises alimentaires sont indiquées soit de manière exacte, soit relativement à des photographies d'assiettes plus ou moins remplies, qui sont mises à disposition par le sondeur. Bien entendu, la période d'observation étant particulièrement courte, les 24H peuvent sembler inappropriés pour l'étude de l'exposition de long terme (Counil et al., 2006). C'est pourquoi il est souvent demandé aux sondés de remplir en parallèle lesdits FFQ/ FPQ (questionnaires de fréquence ou de propension des prises alimentaires si l'on traduit mot-à-mot). Ils permettent en particulier de distinguer les personnes ne consommant jamais certains produits de celles qui en mangent occasionnellement. Le lecteur curieux d'en apprendre davantage au sujet de ces deux grandes sources d'information est invité à se référer à van Klaveren et al. (2012); EFSA (2006); Dodd et al. (2006). Tout au long de la présente thèse nous utilisons la base INCA2 qui, ainsi que décrit ci-après, peut être assimilée à un mélange de 24H et de FFQ.

P.2.1.1 Les habitudes alimentaires en France : INCA2

INCA2 (second opus de l'enquête nationale sur les habitudes alimentaires individuelles en France) est une enquête d'envergure nationale mise en place par l'Anses (agence française de sécurité sanitaire de l'alimentation, de l'environnement et du travail) en collaboration avec l'Insee (institut national de la statistique et des études économiques) entre décembre 2005 et avril 2007. Elle collecte les informations concernant les comportements alimentaires de 2624 adultes et 1455 enfants sélectionnés aléatoirement dans la population française.

Ú Plan de sondage Les individus composant INCA2 ont été sélectionnés selon un plan de sondage complexe à plusieurs degrés, construit afin de produire un échantillon représentatif de la population française selon des critères géographiques, sociologiques et économiques. Ces mêmes variables auxiliaires ont ensuite été utilisées lors d'une étape de redressement, réalisée dans un second temps afin prendre en compte la possible non-réponse de certains sondés ainsi que les fluctuations d'échantillonnage. Une description détaillée du plan de sondage d'INCA2 est disponible dans le rapport Afssa (2009, Chapitre 2 et Appendice 2).

Ú Données alimentaires Tous les participants ont indiqué la nature et la quantité des aliments qu'ils ont consommés durant les 7 jours consécutifs de l'enquête. Ces produits ont été classifiés selon une nomenclature exhaustive de 1342 aliments, re-

groupés en 123 sous-classes et 45 catégories plus larges (Afssa, 2009, Section 2.2.6.3 et Appendice 1). Afin de les aider à évaluer les quantités mangées, un carnet de photographies représentant des assiettes et des verres progressivement remplis a été mis à la disposition des sondés, servant d'étalon lorsque des mesures exactes ne pouvaient être réalisées. Des informations complémentaires sur les conditions des repas ont aussi été collectées, concernant par exemple le lieu, l'heure et la durée de ces événements. Six catégories de repas sont indiquées dans la base finale, à savoir le petit déjeuner, la collation du matin, le déjeuner, la collation de l'après-midi, le dîner et la collation du soir. Pour prendre en compte les variations saisonnières des habitudes alimentaires, les participants ont été contactés à des moments aléatoires de l'année, durant l'une des 3 vagues successives de collecte des données. Un ensemble de questions concernant la prise de compléments alimentaires a aussi été introduit (Afssa, 2009, Section 2.2.6.4 et Chapitre 8).

Ú Information auxiliaire En complément des régimes alimentaires, un nombre important d'informations au sujet des sondés a été noté, allant des caractéristiques sociologiques (e.g. diplôme, profession, revenu du ménage, nationalité) aux préférences alimentaires, en passant par les activités physiques (e.g. type, fréquence, durée), l'historique médical (e.g. troubles du comportement alimentaire) et d'autres indications générales (e.g. age, poids, sexe).

Ú Consommateurs occasionnels Lorsque l'intérêt est porté sur les habitudes alimentaires et l'exposition de long terme, 7 jours peuvent sembler être une période arbitrairement courte d'observation. En effet, en une semaine seulement, les enquêtés ne peuvent couvrir l'ensemble de leur répertoire alimentaire et de nombreux produits de la nomenclature de référence ne sont pas consommés. Un tel phénomène rend difficile la distinction entre les consommateurs occasionnels et ceux qui ne mangent jamais de certains produits. Afin de remédier à ce problème, il a été demandé aux individus constituant la base INCA2 de décrire leurs régimes usuels et de déclarer clairement à quelle catégorie de consommateurs ils appartiennent. Ces questions additionnelles correspondent aux fameux FFQ/ FPQ précédemment mentionnés qui, combinés aux bases de type 24H, facilitent grandement la modélisation statistique (Dodd et al., 2006 et van Klaveren et al., 2012, Sections 3.6 et 3.8).

Ú Information incomplète Comme nous venons de l'évoquer, n'observer les enquêtés que durant 7 jours peut s'avérer être un inconvénient majeur. Cependant, du point de vue des sondés, une semaine complète peut paraître particulièrement long, et la qualité des données peut en pâtir. En effet, comme indiqué dans le rapport Afssa (2009, Section 2.4.2), plusieurs participants ont ponctuellement omis d'indiquer leurs consommations, tandis que d'autres ont largement sous-estimé les quantités ingérées

durant leurs repas. Dans nos calculs nous avons pris le parti d'ignorer ces lacunes en rapportant les informations disponibles à l'échelle de la semaine.

Ú Dépendance temporelle La seule réelle différence entre INCA2 et les bases de type 24H tient à la nature consécutive des 7 jours d'observation. Ainsi, bien qu'une semaine complète semble clairement plus appropriée que 2 jours d'enquête pour une analyse de long terme, il est alors plus difficile d'ignorer la dépendance temporelle.

Ú Traitement préliminaire d'INCA2 Les travaux réalisés dans cette thèse sont dédiés à l'analyse globale des habitudes alimentaires de long terme des adultes en France. En raison de leurs besoins nutritionnels spécifiques et du caractère temporaire de leur état, nous avons décidé de ne prendre en compte ni les femmes enceintes ni les femmes allaitant dans nos calculs. D'autres individus ont été exclus de nos analyses, notamment ceux ayant omis de renseigner des variables essentielles comme le poids corporel, amenant l'échantillon initial à 2488 unités. Bien entendu, nous aurions pu tenter d'appliquer des méthodes classiques pour remédier au problème des valeurs manquantes. Néanmoins, de telles considérations dépassent le cadre de nos travaux. En particulier, ces techniques statistiques sont en général mises en place pour éviter de dégrader l'estimation de phénomènes moyens. Or notre intérêt est porté sur les événements extrêmes (minimum et maximum). Par ailleurs, la proportion de valeurs manquantes dans l'échantillon étant infime (seuls 91 individus sont concernés, soit 3; 7% des enquêtés), nous avons préféré les ignorer.

P.2.2 Données de composition

Les bases de données dites de composition regroupent en général un ensemble de mesures chimiques réalisées sur des groupes plus ou moins raffinés d'aliments. Elles peuvent provenir d'enquêtes de natures variées. Par exemple, des plans de surveillance sont mis en place pour vérifier la sûreté de produits suspects et induisent la collecte de données à leur sujet. Bien que de nombreux éléments du répertoire alimentaire soient ainsi négligés, le nombre de mesures réalisées dans ce cadre est particulièrement conséquent, avantage non-négligeable pour toute analyse statistique. Néanmoins, en raison de la nature suspecte des produits étudiés, les bases de données résultantes ne peuvent être utilisées pour l'évaluation des risques alimentaires à grande échelle sans introduire un biais non-négligeable. A l'inverse, les études dites de l'alimentation totale tentent de couvrir la quasi-totalité des aliments consommés dans une population d'intérêt. En contrepartie, le nombre de mesures réalisées par produit est bien plus modeste (de 2 à 8 en général). Au vu de nos objectifs, nous préférons ces dernières aux plans de surveillance. Les données de contamination peuvent être combinées aux données de consommation pour ensuite approximer l' exposition à certains composants alimentaires. Malheureusement, les nomenclatures de

produits utilisées dans chacune de ces bases peuvent différer substantiellement, ce qui pose un problème supplémentaire au statisticien (voir par exemple le rapport de l'Efsa concernant la mise en commun de données produites par divers pays européens à l'adresse http://www.efsa.europa.eu/en/search/doc/415e.pdf). Comme expliqué ci-après, les bases de données utilisées dans cette thèse ont été construites spécifiquement pour permettre le croisement avec INCA2, contournant par là-même le problème sus-mentionné.

P.2.2.1 Apports nutritionnels : CIQUAL 2008

La version 2008 de la base de données CIQUAL (Centre d'Information sur la QUalité des ALiments) liste les concentrations moyennes en un large panel de nutriments des 1342 produits alimentaires de la nomenclature d'INCA2. Ces teneurs sont déterminées à partir de multiples sources d'information, allant d'analyses chimiques spécifiquement commanditées pour l'étude à des rapports publiés par des instituts de recherche variés. A chaque source est attribué un poids indiquant sa fiabilité, qui est ensuite pris en compte dans un calcul produisant le résultat final. Plus de détails concernant la construction de cette base de données sont disponibles dans le rapport Anses (2008).

Ú De l'invariance nutritionnelle Bien que relativement exhaustive, CIQUAL 2008 ne donne aucune information concernant la variabilité des concentrations en nutriments à l'intérieur d'un même type d'aliment. Cela suggère par exemple que deux oranges cultivées dans des régions différentes, l'une potentiellement plus ensoleillée que l'autre, ont la même teneur en vitamine A. Si cette hypothèse semble déraisonnable de prime abord, les variations de concentration au sein d'une même famille de produits sont souvent si faibles qu'il est d'usage de considérer les teneurs fixes, à condition que la nomenclature utilisée soit assez détaillée (de Boer et al., 2009, p.1433), contrainte qui se trouve être respectée dans le cas de CIQUAL 2008.

P.2.2.2 Contamination des aliments : EAT2

Contrairement aux plans de surveillance, le second opus de l'étude de l'alimentation totale française (EAT2) a été construite dans le but d'estimer l'exposition de la population française à un large ensemble de contaminants alimentaires. Ainsi, plus de 80% des aliments consommés en France y sont pris en compte (Anses, 2011), ce qui en fait une base de donnée de choix au vu de nos objectifs. Cependant, elle possède malgré tout un certain nombre de défauts inévitables.

Ú Mélange des aliments L'une des principales lacunes d'EAT2 provient de la technique utilisée pour réaliser les mesures de concentration. En effet, les analyses chimiques ont été opérées sur des mélanges d'aliments plutôt que sur des produits bruts : différentes espèces d'une même famille d'aliments ont été mixés avant inspection, et

les plats cuisinés ont été étudiés dans leur ensemble, sans en séparer au préalable les ingrédients. Afin de contourner les problèmes induits par cette procédure, des tables de recettes ont été construites en parallèle, permettant la décomposition des résultats en une nomenclature plus raffinée (Anses, 2011).

Ú Données censurées Comme toute base de donnée de contamination, EAT2 souffre d'un problème intrinsèque de censure lié aux limites des instruments de mesure. En effet, ces derniers ne sont pas d'une précision infinie et peuvent ainsi être incapables de quantifier voire même de détecter des quantités trop faibles d'éléments chimiques dans la nourriture. Cette censure inévitable est usuellement introduite dans les modèles statistiques par le biais de deux paramètres essentiels : la limite de détection (LDD) et la limite de quantification (LDQ). La première représente la quantité minimale de composant qui doit être présente dans la substance analysée pour pouvoir être détectée par l'appareil de mesure, et la seconde le niveau nécessaire pour permettre la quantification. Toutes deux dépendent bien évidemment de l'instrument utilisé. En pratique, il est d'usage d'adopter l'une des trois hypothèses listées ci-après pour contourner ce problème. Notons t_D la quantité réelle de composant lorsqu'elle est en dessous de la LDD, et t_Q lorsqu'elle est située entre la LDD et la LDQ. Les trois scénarii possibles sont alors :

1. hypothèse médiane : t_D = LDD=2 et t_Q = LDQ=2,
2. hypothèse basse : t_D = 0 et t_Q = LDD,
3. hypothèse haute : t_D = LDD et t_Q = LDQ.

L'hypothèse médiane est typiquement choisie lorsque le taux de mesures censurées dans la base de données utilisée ne dépasse pas les 60%. Nous décidons ici d'adopter ce scénario, les considérations à propos de la censure des données de contamination dépassant le cadre de nos travaux.

P.3 estimation des risques alimentaires : petit aperçu de l'état de l'art

Ces 20 dernières années, pléthore de modèles statistiques a été développé afin d'évaluer de multiples types de risques alimentaires. Construits pour répondre à des questions pratiques liées à la nature des bases de données disponibles telles que la censure mentionnée au paragraphe précédent (Tressou, 2006), ils continuent d'évoluer avec les nouvelles méthodes de collecte des données. Dans le contexte des bases de type 24H et FFQ/ FPQ, auxquelles INCA2 peut être comparée, trois grandes problématiques classiques, décrites dans les paragraphes suivants, sont abordées dans cette thèse.

P.3.1 Exposition extrême de long terme

Afin de compenser la courte durée des enquêtes de type 24H, les modèles à correction d'erreur ont récemment gagné en popularité (Tooze et al., 2006; Dodd et al., 2006; van Klaveren et al., 2012; de Boer et al., 2009; Boon et al., 2011). Ils permettent en particulier de prendre en compte deux sources de variabilité de l'exposition observée, à savoir les fluctuations individuelles autour des prises habituelles et l'hétérogénéité au sein de la population. Toutes deux sont estimées à l'aide de modèles paramétriques et la distribution finale de long terme est approchée par simulations de Monte-Carlo en ignorant la variance intra-individuelle. Les hypothèses sur lesquelles reposent les modèles paramétriques évoqués précédemment impliquent en général que la distribution de l'exposition individuelle est de type log-normale. Cependant, pour de nombreux contaminants et nutriments, la queue de distribution de la loi log-normale est souvent trop fine pour rendre compte convenablement de la probabilité d'occurrence des événements extrêmes; il y a fort à parier que la sur-exposition à de tels composants alimentaires est alors sous-estimée. Lorsque les très fortes expositions à des nutriments et contaminants alimentaires sont au cœur des préoccupations, la théorie des valeurs extrêmes (TVE) a déjà fait ses preuves dans le cadre de l'estimation des risques aigus (Tressou et al., 2004b,a; Bertail et al., 2010; Kennedy et al., 2011; Paulo et al., 2006), nous encourageant ainsi à étendre ces procédures à l'évaluation des risques de long terme.

P.3.2 Exposition simultanée à un ensemble de nutriments et contaminants

Au delà de l'opposition entre risques aigus et risques chroniques, un sujet brûlant d'actualité en évaluation du risque alimentaire est celui de l'exposition simultanée à un cocktail de composants chimiques présents dans la nourriture. En effet, si l'ingestion excessive d'un élément toxique peut avoir des effets dévastateurs sur l'organisme, les connaissances actuelles au sujet des potentiels effets synergétiques, combinés, de plusieurs contaminants sont encore faibles (Carpenter et al., 2002). En France, l'Anses a récemment initié le programme Pericles (Crépet et al., 2013), dans le but d'identifier les cocktails de composants alimentaires effectivement consommés dans la population. Les résultats de ces analyses permettront ensuite aux chimistes et biologistes d'établir un ordre de priorités quant aux recherches à effectuer sur les effets sanitaires d'une telle consommation. Pour le moment, ces travaux sont centrés autour des phénomènes moyens (Béchaux et al., 2013), au détriment des extrêmes. Faisant à nouveau appel à la théorie des valeurs extrêmes, la partie multivariée de cette branche de la statistique devrait permettre l'analyse des fortes expositions simultanées. Lorsque seul un petit nombre de composants est considéré, quelques travaux dans cet esprit

ont d'ores et déjà été réalisés (Paulo et al., 2006). Cependant, en particulier dans le cas de certains types de contaminants comme les pesticides ou les polychlorobiphényles, qui possèdent des centaines de congénères, il serait désirable d'être à même de gérer les grandes dimensions. Malheureusement, d'un point de vue théorique, la TVE ne permet pas encore de traiter des problèmes de dimension plus grande que 5 or 6. Ce problème concret issu de l'analyse du risque alimentaire fait ainsi naître une problématique tout aussi théorique qu'appliquée.

P.3.3 Données de sondage

Dans toutes les méthodes sus-mentionnées, il est de coutume de supposer que les données disponibles sont indépendantes et identiquement distribuées (iid) selon une certaine mesure de probabilité. Or les bases de données de type 24H comme INCA2 sont en général construites à l'aide d'un plan de sondage élaboré, dont l'objectif est de produire un échantillon représentatif de la population d'intérêt. L'hypothèse que les données résultantes sont iid n'est alors pas respectée, et les individus se voient chacun attribué un poids de sondage correspondant à l'inverse de leur probabilité d'être sélectionné dans l'échantillon (Droesbeke et al., 1987; Tillé, 1999). Ignorer cette étape de construction de la base de données et négliger les poids de sondage peut produire des estimateurs biaisés (Bonnery, 2011). Dans le cadre de la présente étude, la probabilité d'occurrence de la très forte exposition à un ou plusieurs composants alimentaires a de fortes chances d'être sur- ou sous-estimée. Bien que la littérature sur les sondages soit déjà très riche (Gourieroux, 1981; Droesbeke et al., 1987; Deville, 1987; Cochran, 1977; Tillé, 2006), ce n'est que très récemment que des résultats fonctionnels, qui permettent par exemple l'estimation de l'intégralité d'une fonction de répartition, ont commencé de se développer (Breslow and Wellner, 2008, 2007; Saegusa and Wellner, 2011). Quant à l'analyse des extrêmes, elle reste à notre connaissance encore inexplorée.

Essayant d'apporter des réponses à ces problématiques, nous proposons dans cette thèse un ensemble de méthodes heuristiques et de résultats théoriques développés spécifiquement pour les bases de données de type 24H. Comme expliqué plus tard en détail, plusieurs de ces procédures mathématiques s'avèrent applicables à bien d'autres domaines que l'analyse du risque alimentaire.

P.4 objectifs et contributions principales

Au cours du présent travail, nous essayons d'apporter quelques réponses aux problématiques statistiques évoquées à la section précédente. Le mémoire est structuré comme suit : au chapitre 2, nous montrons comment la théorie des valeurs extrêmes peut être utilisée pour la modélisation de l'exposition chronique à un composant ali-

mentaire lorsque la distribution sous-jacente est à queue épaisse. Nous tournant ensuite vers le cadre multivarié, nous nous intéressons à la forte exposition simultanée à plusieurs nutriments et contaminants. Nous proposons ainsi au chapitre 3 une nouvelle méthode, mélangeant algorithmes issus de l'apprentissage statistique et analyse de la mesure spectrale, qui permet d'identifier des groupes de variables dépendantes dans les extrêmes, réduisant par là-même la dimension initiale du problème. Toujours dans une optique multivariée, nous quittons au chapitre 4 les extrêmes en faveur des phénomènes moyens. Nous y présentons une approche en terme d'ensembles de volume minimum pour l'estimation non-paramétrique de la distribution de l'exposition à un cocktail d'éléments chimiques présents dans la nourriture. Ces résultats sont ensuite adaptés à l'identification de paniers de consommation réalisant un compromis entre risque toxicologique et bénéfice nutritionnel. Enfin, nous nous attelons dans les deux derniers chapitres au traitement des données issues d'un plan de sondage. Après avoir étendu quelques résultats fonctionnels usuels en analyse des processus empiriques au cadre des plans de sondages de type Poisson dans le chapitre 5, nous introduisons au chapitre 6 un nouvel estimateur de l'indice de valeurs extrêmes adapté aux échantillons issus de tels plans. Bien qu'encore trop restrictifs pour être directement appliqués à l'analyse du risque alimentaire, ces résultats préliminaires constituent un premier pas dans la direction de futurs développements qui, nous l'espérons, permettront bientôt la construction de modèles plus généraux, adaptés aux bases de données comme INCA2.

P.4.1 Estimation de l'exposition extrême de long terme à un unique élément chimique

Nous commençons dans ce chapitre par une réflexion sur les modèles à adopter lors du calcul de l'exposition chronique à un unique nutriment ou contaminant alimentaire. Portant un intérêt tout particulier aux phénomènes extrêmes, dans la lignée des travaux de Tressou et al. (2004a), nous remettons au goût du jour une méthode non-paramétrique généralement ignorée en faveur des modèles à correction d'erreur brièvement décrits à la section précédente. Lorsque les données sont issues d'une distribution à queue épaisse, nous montrons qu'elle comporte plusieurs avantages. La première étape de cette technique consiste à moyenner les données temporelles pour les ramener à l'échelle de la journée. Nous évitons ainsi la modélisation statistique des variances intra- et inter-individuelles qui affectent la distribution de l'exposition de long terme, encouragée par Tooze et al. (2010); van Klaveren et al. (2012); de Boer et al. (2009). Les modèles classiques de la théorie des valeurs extrêmes pour l'étude des lois à queue épaisse sont ensuite appliqués pour estimer quantiles extrêmes et faibles probabilités de dépasser un très haut seuil de recommandation. Si l'exposition est en effet à queue plus lourde qu'une loi log-normale, nous montrons qu'en procé-

dant de la sorte, à l'inverse des modèles traditionnels, nous évitons de sous-estimer ces quantités. Comme en pratique le statisticien doit choisir le modèle le plus adapté aux données, nous recommandons l'utilisation préliminaire de quelques tests statistiques connus détectant le cas échéant la présence d'une queue lourde. L'intégralité de notre méthodologie est enfin appliquée à des données réelles pour en montrer les avantages et les inconvénients.

P.4.2 Evaluation de la sur-exposition simultanée à un cocktail de composants alimentaires

Nous étendons ensuite notre approche au cadre multivarié et considérons l'analyse de la forte exposition chronique à plusieurs nutriments et contaminants. Dans l'esprit du programme Pericles (Crépet et al., 2013), nous développons une nouvelle méthode permettant d'identifier des cocktails de composants chimiques consommés simultanément en très grandes quantités dans une population d'intérêt. Nous nous inspirons pour cela à la fois d'algorithmes d'apprentissage statistique et de concepts de la théorie des valeurs extrêmes multivariée. Nous étudions en particulier la mesure spectrale, objet mathématique qui caractérise la dépendance extrême. Elle peut être définie sur l'orthant positif de la sphère unité, i.e. le simplexe, lui-même décomposable en faces ouvertes (sommets et arêtes), chacune indiquant un groupe de variables dépendantes dans les extrêmes. L'objectif est alors d'identifier les faces du simplexe sur lesquelles la mesure spectrale est définie. Pour cela, nous commençons par mettre en œuvre le récent algorithme de Jung et al. (2012), appelé Principal Nested Spheres, qui réduit la dimension des données en les projetant successivement sur des sphères de dimension de plus en plus faible. A la manière d'une analyse en composantes principales, cette première étape permet de faciliter la suite de l'analyse en réduisant le bruit des observations. Ensuite, nous exhibons un modèle de mélange de la mesure spectrale sur chacune des faces du simplexe et définissons une variable latente caractérisant les composantes du mélange. Cette dernière est estimée de manière non-paramétrique en utilisant d'abord un algorithme de classification appelé spherical k-means (Dhillon et al., 2002), puis en utilisant un critère heuristique construit pour déterminer les faces auxquelles les classes obtenues font référence. Ce critère permet de même de choisir le nombre d'observations extrêmes qui peuvent être considérées comme représentatives de la queue de la distribution multivariée. Nous justifions notre approche à l'aide d'une étude par simulations, dont les résultats sont fort encourageants, puis l'appliquons enfin à nos bases de données réelles, INCA2, CIQUAL 2008 et EAT2. Les groupes résultants de contaminants et nutriments supposés consommés simultanément et en de très grandes quantités dans la population française font parfaitement sens et sont cohérents avec une analyse paire par paire, ce qui vient renforcer notre confiance en notre approche.

P.4.3 Ensembles de volume minimum et approche risque-bénéfice

Allant plus loin encore dans l'utilisation de l'apprentissage statistique pour l'analyse du risque alimentaire, nous nous concentrons dans ce chapitre sur les phénomènes moyens, non plus extrêmes. Nous y introduisons pour la première fois la variabilité de la contamination des aliments, les teneurs ayant jusque lors été considérées fixes au sein d'une même famille de produits. Dans la lignée des travaux de Bertail and Tressou (2006), nous proposons une extension de l'estimation d'ensembles de volume minimum de Scott and Nowak (2006) au cas où le volume est inconnu et approché à l'aide d'une U-statistique. Les résultats théoriques sont ensuite appliqués à la construction non-paramétrique des ensembles de niveau de la distribution multivariée de l'exposition à un ensemble de composants alimentaires. Dans un second temps, nous montrons comment cette procédure peut être généralisée à l'identification de paniers de consommation qui réalisent un compromis entre risque toxicologique et bénéfice nutritionnel. Encore en cours de programmation, nous espérons pouvoir appliquer cette dernière technique aux bases de données INCA2, CIQUAL 2008 et EAT2 afin de définir des recommandations simples dans le même esprit que "mangez cinq fruits et légumes par jour".

P.4.4 Processus empiriques dans le cadre des sondages

Nous nous attelons dans les deux derniers chapitres de cette thèse au traitement des données issues d'un plan de sondage, comme le sont souvent les données de consommation. Il est d'un intérêt majeur pour les instituts de santé publique comme l'Anses d'être à même d'estimer convenablement la distribution de l'exposition de long terme à un ou plusieurs éléments chimiques. Or, omettre le plan de sondage dans le processus d'estimation peut induire un biais non-négligeable (Bonnery, 2011). Tentant de contribuer à l'élaboration d'une théorie générale sur les sondages garantissant la normalité asymptotique d'une large classe d'estimateurs, comprenant notamment des estimateurs fonctionnels de la fonction de répartition de l'exposition, nous commençons par étudier les plans de type Poisson. Des extensions à des plans plus complexes comme celui utilisé pour la formation d'INCA2 seront envisagées dans un futur proche. Notre approche est directement inspirée des travaux séminaux de Hàjek (1964) et Berger (1998). Nous définissons en premier lieu le processus empirique de type Horvitz-Thompson, où les observations sont pondérées par l'inverse de leur probabilité d'inclusion, dans le cadre du plan de Poisson. Sous un ensemble d'hypothèses classiques portant à la fois sur les probabilités d'inclusion et le modèle de surpopulation considéré, que nous espérons relâcher dans des travaux à venir, nous en montrons la convergence vers un processus Gaussien dont nous exhibons la covariance. Dans un second temps, nous généralisons ces résultats aux

processus empiriques à la Horvitz-Thompson impliquant des échantillons sélectionnées selon un plan de sondage à forte entropie comme le plan réjectif. Pour ce faire, nous utilisons des résultats connus exhibant la proximité entre les plans de sondage à forte entropie et le plan simple de Poisson. Les théorèmes fonctionnels de la limite centrale ainsi obtenus sont ensuite utilisés pour montrer la normalité asymptotique d'estimateurs pouvant s'écrire comme certaines fonctionnelles (Hadamard ou Fréchet différentiables) du processus empirique. Enfin, nous étudions le cas spécifique de l'estimateur Horvitz-Thompson de la fonction de répartition et, illustrations à l'appui, montrons comment utiliser nos résultats pour en construire des bandes de confiance uniformes. Au delà de l'analyse du risque alimentaire, ces résultats semblent tout à fait appropriés à la gestion de bases de données de taille gigantesque (les fameuses "big data"), dont la taille augmente sans cesse, comme les données financières, et ne peuvent être exploitées sur un unique ordinateur. Dans un tel contexte, l'échantillonnage semble une solution toute naturelle aux problèmes de mémoire informatique. Les plans de sondage que nous avons étudiés se révèlent alors d'un intérêt tout particulier, permettant d'obtenir simplement des estimateurs non biaisés et d'une efficacité optimale (il suffit pour cela de calibrer correctement les probabilités d'inclusion) sur des échantillons de taille raisonnable.

P.4.5 Estimation de l'indice de valeurs extrêmes à partir de données de sondage

Nous concluons cette thèse en revenant au problème initial de l'estimation des phénomènes extrêmes, prenant cette fois-ci en compte le plan de sondage selon lequel les données ont été collectées. A notre connaissance, la théorie des valeurs extrêmes n'a pour le moment pas été étendue au cadre des sondages. Nous commençons donc modestement par l'adaptation dans ce contexte de l'un des estimateurs les plus classiques de la TVE, à savoir l'estimateur de Hill de l'indice de valeurs extrêmes (Hill, 1975). Dans le même esprit que les travaux du chapitre précédent, nous en construisons une version Horvitz-Thompson dont nous montrons la consistance et la normalité asymptotique pour les plans de type Poisson. La vitesse de convergence de ce nouvel estimateur s'avère être la même que si la population entière avait été accessible, et la variance n'est dégradée que d'un paramètre multiplicatif dépendant du choix des probabilités d'inclusion, qu'il est théoriquement possible de minimiser. A l'aide d'expériences numériques, nous montrons que les hypothèses restrictives que nous exigeons pour établir ces résultats mathématiques sont loin d'être nécessaires et pourraient être relâchées dans de futurs travaux. Rappelant enfin les problématiques inhérentes aux "big data", nous encourageons l'extension de ces travaux préliminaires à d'autres familles d'estimateurs issus de la TVE.

THESIS

MULTIVARIATE STATISTICS FOR DIETARY RISK ANALYSIS

1

INTRODUCTION: OF DIETARY HABITS AND EXPOSURE

Dietary risk analysis is a generic public health term that embraces as many scientific problems as there are ways for consumers to get sick by eating (or not eating). For instance, food industries are concerned with the development of bacteria during the consecutive fabrication, transportation, distribution and stocking processes that precede consumption (Rigaux et al., 2012). On another level, agribusinesses may want to evaluate the impact of economically profitable mass production methods (e.g. genetically modified organisms, pesticides) on the human organism (cf. http://www.efsa.europa.eu/en/topics/topic/gmo.htm on GMO and http://www.efsa.europa.eu/en/topics/topic/pesticides.htm on pesticides). Cooking utensils can also represent a source of danger, depending on their chemical composition and the way they are used, as was recalled by the recent polemic about the presence of bisphenol A in baby bottles (see the ensuing Anses avis, Anses, 2013). Other health issues may be caused by behavioral phenomena such as eating disorders (e.g. people suffering from severe anorexia are likely to develop diseases due to nutritional deficiency), or by biological determinants (e.g. diabetes). As suggested by this non-exhaustive list of examples, depending on the examined question, specific knowledge in chemistry, medicine, sociology, economy, biology or even psychology may be required, not to mention probabilistic modeling and computer data management, which makes dietary risk analysis a multidisciplinary field par excellence. We invite the interested reader to consult Feinberg et al. (2006) and the references therein for a comprehensive introduction to the ins and outs of dietary risk analysis from an interdisciplinary point of view.

The present work is dedicated to a very specific type of dietary risk: independently from any production, stocking, cooking, or biological predisposition phenomena, we are concerned with the very high (or low) long-term exposure of the French population as a whole to some food components. After having introduced some basic notions about dietary risk analysis in Section 1.1, described the available data in Section 1.2 and quickly reviewed the literature on probabilistic modeling of dietary risks in Section 1.3, we thoroughly present the various problems tackled in this thesis and succinctly report its main scientific contributions in Section 1.4.

1.1 getting a taste of dietary risk analysis

Once a specific type of dietary risk has been brought into focus, its analysis involves three distinct stages, usually referred to as risk evaluation, risk management and risk communication. In the next paragraphs, we give a concise overview of the various issues addressed in these successive and complementary steps. To help better understand the challenges at stake, we illustrate each introduced methodological concept with the specific example of salty diets, the recurrent abuse of which can favor cardiovascular issues (see for instance http://www.anses.fr/en/content/salt and the references therein).

1.1.1 Risk evaluation

Risk evaluation (or assessment) can be schematically decomposed into four substages, which consist in successively identifying and characterizing both potential dangers and their probability of occurrence. In the case of salty diets, the danger would be to ingest too much sodium (Na in the periodic table) on a daily or weekly basis, since in the long-run it would be likely to damage the cardiovascular system. These noxious effects were spotted by means of chemical and biological studies about the assimilation, action and elimination of Na in the human organism. This process is called hazard identification (Barlow et al., 2002). It is followed by further research designed to understand at what point the component becomes noxious, i.e. characterize the danger. In general, danger characterization involves testing dose-response effects in vitro or in vivo on animals before extending the subsequent results to humans by means of dedicated models (Dybing et al., 2002). Once the contamination process is fully comprehended, it becomes possible to assess its probability of occurrence in a given population. This requires first a thorough description of the distributions of consumption and exposure in the population (Kroes et al., 2002), then a comparison with some maximal intake limit determined by experts from the aforementioned dose-response trials (Renwick et al., 2003), possibly enhanced by mathematical designs (Edler et al., 2002). Depending especially on the data at hand, statistical models may include a temporal dimension (Bertail et al., 2010, 2008; Allais and Tressou, 2009), take into account individual biological or behavioral characteristics, or try to make up for the limited amount of available information (Dodd et al., 2006). Considerations on this matter are discussed in detail in Section 1.3 and Chapter 2.

1.1.2 Risk management

Once dietary risks have been evaluated, it is necessary to identify their determinants, evaluate their relative impact and define the best strategy to rapidly and ef-

ficiently reduce risks. This is the exact purpose of risk management. For instance, over-exposure to sodium can sometimes be traced back to the repeated consumption of some salt-saturated products such as precooked dishes or biscuits. A simple line of action there would be to simultaneously warn consumers of the potential dangers and impose food companies to limit the amount of added salt in their preparations. In more extreme cases, after a thorough evaluation of the economic impact of such measures, some foodstuffs may even be taken off the market.

1.1.3 Risk communication

The last stage of dietary risk analysis is called risk communication. It can take place at any moment of the analysis and be intended for scientists and risk managers involved in the process as well as food industries or consumers. As such, it is not restricted to public health campaigns such as "avoid eating products that contain too much salt" in our example. It also encompasses scientific reports about the methods considered appropriate to detect and quantify risks, publications of case study results, international discussions about food policies, etc. In the sequel, we are mainly interested in producing results that would help orient further research and, when possible, set general, easily understandable and applicable dietary guidelines to the population.

With this objective in mind, most of our work falls into the evaluation phase of dietary risk analysis: in a few words, we design statistical methods to assess some characteristics (e.g. quantiles) of the very high exposure to some food chemicals over a long period of time. Except in Chapter 4 where we elaborate a statistical methodology that ascertains balanced food baskets with regard to toxicological risk and nutritional benefit, risk management is not the main concern here. The components of interest are nutrients and contaminants to which the chronic over- or under-exposure has known detrimental sanitary effects. In particular, we disregard acute risks such as those involving bacterial contamination of the food, which can impact the organism in only a few days or even a few hours. Statistical analysis is based in this context on some standard types of databases, which are described in detail in the next subsection.

1.2 collecting information: multiple sources, multiple difficulties

To statistically analyze the chronic dietary exposure of a given population to a collection of food components, we need to observe such types of exposure on a large sample of individuals over a long period of time. Unfortunately, the amounts of

nutrients and contaminants ingested during a meal are not directly measurable and can only be assessed by means of various methods. This can be achieved for instance by detecting and quantifying some specific biomarkers, the presence of which is ultimately linked to the level of exposure (refer for instance to the Anses website http://www.anses.fr/sites/default/files/documents/RSC1205-DossierParticipants.pdf giving the list of presentations in a dedicated workshop held in May 2012). From a practical point of view, this necessitates collecting then chemically analyzing samples of body fluids, hair or skin of a relatively large array of people. At a national level, the cost of such procedures can constrain the sample size at the expense of statistical efficiency. Moreover, many chemicals can be assimilated via other elements than food (e.g. air) and biomarkers cannot distinguish between the various sources of exposure (Sirot et al., 2009). So as to study the sole impact of nourishment on individual exposure, one may use in an alternative manner a combination of two types of databases, one listing the dietary intakes of a sample of consumers and the other the levels of components in a fine nomenclature of products. The next paragraphs provide an overview of the main characteristics of these so-called consumption and composition data.

1.2.1 Consumption data

There are many types of data that can be used to assess chronic exposure to nutrients and contaminants in France, ranging from panel cohorts on alimentary expenditures of households (Secodip, Nichèle et al., 2008) to Internet repositories filled up on a voluntary basis (Nutrinet, Hercberg et al., 2010) and dedicated nationwide institutional surveys (INCA2, Afssa, 2009). Focus is here on two major categories, namely 24-hour recalls (abbreviated 24H) and food frequency or propensity questionnaires (FFQ/ FPQ). The first category corresponds to large surveys (possibly national) where a sample of selected individuals report in detail their dietary habits during 2 to 7 days, sometimes consecutive but most of the time randomly picked within the year. Food intakes are given exactly or relative to some pictures of more or less filled plates provided by the pollster. Obviously, the short duration of these surveys can impede the estimation of long-term dietary habits (Counil et al., 2006). To help deal with this issue, food frequency (or propensity) questionnaires are proposed to the consumers, in which they declare what type of food they most commonly eat and which products they avoid or never consume. We refer the interested reader to van Klaveren et al. (2012); EFSA (2006); Dodd et al. (2006) for an account of the assets and liabilities of these complementary approaches. In this thesis, we used the INCA2 database described herein-after, which can be assimilated to a mixture of 24H and FFQ.

1.2.1.1 Dietary habits in the French population : INCA2

INCA2 (second opus of the national survey on individual dietary habits in France) is a nationwide survey conducted by Anses (French agency for food, environmental and occupational health safety) in collaboration with Insee (French national institute of statistics and economic studies) between December 2005 and April 2007. It collects information about the dietary habits of 2624 adults and 1455 children taken at random in the French population.

Ú Survey scheme Individuals in INCA2 were selected according to a complex multistage survey scheme that was designed to produce a representative sample relative to geographical, sociological and economic criteria. Post-calibration methods were applied in a second phase with respect to the same set of auxiliary variables to provide corrected survey weights relative to both non-response and sampling fluctuation. We refer to Afssa (2009, Chapter 2 and Appendix 2) for a thorough description of the survey plan.

Ú Dietary information All participants reported both the nature and the amount of food that they ate during 7 consecutive days. These products were classified according to an exhaustive nomenclature of 1342 foods grouped into 123 sub-classes and 45 wider categories (Afssa, 2009, Section 2.2.6.3 and Appendix 1). To help them assess the quantities they ate, inquired people were given a notebook displaying pictures of progressively filled plates and glasses. It served as a referential in case exact measurement was not possible. A precise description of each meal was also provided, indicating for instance when, where and with whom they occurred. In the final database, they were classified into six types of meals, namely breakfast, morning snack, lunch, afternoon snack, dinner and evening snack. So that the seasonal variation of dietary habits may be controlled, individuals were contacted randomly at different periods of the year during 3 distinct phases of data collection. A specific line of questions was also established to assess food supplement consumption in the French population (Afssa, 2009, Section 2.2.6.4 and Chapter 8).

Ú Auxiliary information In addition to dietary habits, socio-professional characteristics (e.g. diploma, profession, household income, nationality), dietary preferences, physical activities (e.g. type, frequency, duration), medical history (e.g. eating disorders), and other auxiliary information about participants (e.g. age, weight, sex) were listed.

Ú Episodic or non-consumers When interested in long-term dietary habits and food chemical exposure, 7 days of observation appear to be an arbitrarily short period of time. In particular, in only one week, individuals cannot cover their entire food repertoire and many items in the reference nomenclature are not consumed. Thus, it be-

comes difficult to discriminate between real non-consumers and episodic consumers. So as to make up for this drawback, individuals in INCA2 were asked to depict their dietary habits, and clearly declare to which category they belonged. These additional queries corresponds to the so-called food frequency/ propensity questionnaires previously mentioned, which, when combined with 24-hour recall data, facilitate the statistical modeling of long-term food or nutritional intakes (Dodd et al., 2006 and van Klaveren et al., 2012, Sections 3.6 and 3.8).

Ú Incomplete information Regarding episodic versus non-consumers, collecting data during only 7 days is clearly a liability, and a wider period of observation might be preferred. Conversely, when concerned with the quality of answers, an entire week can be considered too long to provide reliable information. Indeed, as mentioned in Afssa (2009, Section 2.4.2), some participants did not report their consumption every day and other under-estimated their intakes. In our calculations, we decided to ignore these faults and simply scaled the observed dietary habits to the entire week.

Ú Time dependence The only difference between INCA2 and classical 24-hour recall databases is that the 7 days of observation were consecutive. Consequently, although an entire week is clearly better than the usual 2 days for long-term analysis, it makes temporal dependence harder to ignore.

Ú Preliminary processing of INCA2 The present work is dedicated to the global analysis of the long-term dietary habits of French adults. Because of their very specific nutritional needs and the temporary character of their condition, we decided not to take into account pregnant or lactating women. Other individuals were excluded, namely those for whom important variables were missing (e.g. body weight or consumed food amount), thereby restricting the initial sample to 2488 units. Obviously, we could have tried to apply standard techniques to infer on missing values, but this went beyond the scope of our work. In particular, such methods are usually designed to avoid degrading the estimation of average phenomena. Since we are more interested in extreme (maximum or minimum) events and incomplete data only concerned a very small proportion of the sample, we chose to ignore them instead. In the final considered database, 91 individuals (3.7%) ceased filling the questionnaire after only a few days.

1.2.2 Composition data

Composition databases usually collect an array of chemical measurements realized on more or less refined groups of foodstuffs. They can originate from surveys of very different natures. For instance, surveillance plans are designed to punctually check the safety of some suspect products. Hence, by nature, they do not cover the entire

food repertoire, but are particularly thorough in the sense that the collected samples are of important size. In addition, since the inspected foodstuffs are suspected to be abnormally contaminated, they cannot be used to estimate the toxicological exposure of a large population without introducing a non-negligible bias. Alternative types of data are the total dietary surveys. Based on samples of food of very small sizes (2 to 8 items in general), they however encompass a large array of products. As such, they are more suitable for our purpose. The main difficulty when crossing them with consumption data is that the corresponding nomenclatures may substantially differ, thereby necessitating an additional step, potentially difficult, of association of the various sources. This issue is particularly pregnant with meaning for international institutes such as Efsa, which have to combine data from various countries (see the recent report on the matter at http://www.efsa.europa.eu/en/search/doc/415e.pdf). Here, we work with databases that were constructed specially to fit INCA2. They are briefly introduced below.

1.2.2.1 Nutrient supply : CIQUAL 2008

The 2008 version of the CIQUAL database (Centre d'Information sur la QUalité des ALiments in French) lists the average concentrations with regards to a large set of nutrients for each of the 1342 dietary products of the INCA2 nomenclature. These levels were determined using multiple sources of information, ranging from specifically ordered laboratory analyses to the published reports of various research institutes. Each source was attributed a weight representing its reliability before a final synthesizing calculus. We refer to Anses (2008) for more information concerning the construction of this database.

Ú About nutritional invariability Although already quite exhaustive, CIQUAL 2008 does not provide any information about the variability of concentration within a type of food. This suggests for instance that two oranges grown in different regions, one possibly sunnier than the other, contain the same level of vitamin A. Even if this assumption seems questionable at first sight, nutrient contents usually have such a small variance that it is customary to consider them fixed, provided the associated food nomenclature is detailed enough (de Boer et al., 2009, p.1433). The list of products considered in CIQUAL 2008 was designed to respect this constraint.

1.2.2.2 Food contamination : TDS2

Contrary to surveillance plans, the second opus of the French Total Dietary Survey (TDS2) is designed to assess the global exposure of the French population to a whole set of food contaminants and thereto covers more than 80% of the current food repertoire in France (Anses, 2011). Though particularly adapted to our needs, it still possesses some unavoidable drawbacks.

Ú Pooling foodstuffs One of the major limitations of TDS2 is that the analyzed products are not raw but pooled, i.e. different species of a same food item were mixed together before chemical inspection. Dishes were also treated as such, meaning that instead of separating the elements of a recipe they were considered as a whole. To overcome this issue, recipe tables were constructed in parallel, to enable the decomposition of the results of INCA2 into a more refined nomenclature (Anses, 2011).

Ú Censored data Like all contamination databases, TDS2 suffers from an intrinsic censorship due to the limitations of measurement devices. Indeed, machines cannot have an infinite precision, and as such can fail quantifying, or worse detecting chemicals in the food. This inconvenient censorship is usually introduced in statistical models using two major parameters, namely the Limit Of Detection (LOD) and the Limit Of Quantification (LOQ). The former represents the minimal amount of chemical that has to be present for the instrument to detect its presence and the latter that for which quantification is possible. They clearly depend on the measuring instrument used to perform the chemical analyzes. In practice, the most common method to balance this restriction is to adopt one of the three following hypotheses. Denote by t_D the real amount of component when below the LOD and by t_Q when in between the LOD and the LOQ. Then, the three possible scenarios are:

1. median hypothesis: t_D = LOD=2 and t_Q = LOQ=2,
2. lower hypothesis: t_D = 0 and t_Q = LOD,
3. upper hypothesis: t_D = LOD and t_Q = LOQ.

The median hypothesis is typically chosen when the censored measures do not represent more than 60% of the entire database. In this thesis, we chose to adopt this specific approach, considerations about censorship going beyond the scope of our work.

1.3 statistical assessment of dietary risks: a review of the literature

In the last 20 years, a plethora of dedicated statistical methods have been developed to assess dietary risks of various natures. Designed to answer practical issues linked to the type of available data such as the censorship in the measurement of levels of contents (Tressou, 2006), they keep on evolving with the new methods of data collection. In the context of 24-hour recalls and food frequency questionnaires, to which INCA2 is very similar, we can distinguish three major issues, detailed in the next paragraphs, that are to be tackled in the present thesis.

1.3.1 Extreme long-term exposure

To cope with the limited duration of 24-hour recalls, statistical models of the measurement error type have recently gained popularity (Tooze et al., 2006; Dodd et al., 2006; van Klaveren et al., 2012; de Boer et al., 2009; Boon et al., 2011). They account for the presence of two sources of variance in the observed exposure, namely individual fluctuations around usual habits and populational heterogeneity. Both are estimated by means of parametric modeling and the final long-term distribution is approached with Monte-Carlo simulations that disregard the intra-individual variations. The aforementioned parametric assumptions usually implicitly require that the usual intakes have log-normal distribution. However, for many nutrients and contaminants, the tails of such probability laws are sometimes too thin to accurately account for the occurrence of extreme events; under-estimation is thus a non-negligible risk. When interested in the very high exposure to nutrients and contaminants, methods issued from extreme value theory (EVT) have already proven useful in acute risk estimation (Tressou et al., 2004b,a; Bertail et al., 2010; Kennedy et al., 2011; Paulo et al., 2006), suggesting that an extension to the long-term setting would be worth considering.

1.3.2 Simultaneous exposure to multiple food components

Beyond the issue of chronic versus acute risks, a particularly hot topic in dietary risk assessment is the analysis of simultaneous exposure to multiple chemicals. Indeed, if the excessive ingestion of toxicants can have a detrimental impact on health, knowledge about synergistic, combined effects is still poor (Carpenter et al., 2002). In France, the Anses institute has most recently launched the Pericles program (Crépet et al., 2013) in order to identify cocktails of components that are observed to be consumed in the population. The ensuing results should provide guidelines to chemists and biologists who will then be in charge of inspecting the corresponding sanitary effects on the human organism. For now, most of this work is focused on average phenomena (Béchaux et al., 2013), at the detriment of extremes. Going back to extreme value theory, using the multivariate branch of this field should provide ways of dealing with high joint exposure. When only a few dimensions are involved, attempts in that direction have already been published (Paulo et al., 2006). However, especially for some types of contaminants such as pesticides or Polychlorinated biphenyls, which possess hundreds of congeners, managing high dimensions would be desirable. Unfortunately, from a theoretical point of view, EVT does not handle dimensions higher than 5 or 6 yet, making this branch of dietary risk analysis both a practical and a methodological challenge.

1.3.3 Survey data

In all the aforementioned methods, it is usually assumed that the data at hand is independent, identically distributed (iid) according to some probability measure. However, 24-hour recalls and databases like INCA2 are typically constructed from some elaborate survey scheme, designed to produce samples that are representative of the population of interest. There, the iid hypothesis is no longer satisfied, and individuals are attributed survey weights based on their probability of being included in the sample (Droesbeke et al., 1987; Tillé, 1999). Thus ignoring the underlying design is known to produce biased estimates (Bonnery, 2011), in our case the probability of occurrence of very high exposure is likely to be either over- or under-estimated. Though the literature on survey sampling is quite rich (Gourieroux, 1981; Droesbeke et al., 1987; Deville, 1987; Cochran, 1977; Tillé, 2006), functional results that would enable the estimation of the entire distribution function are just starting to flourish (Breslow and Wellner, 2008, 2007; Saegusa and Wellner, 2011) and extreme analysis is, to our knowledge, still unexplored.

In an attempt to answer these issues, we propose in the present work a collection of heuristic methods and theoretical results that are specially built for 24-hour recall types of databases. As shall be seen later on, some of these mathematical findings have possible applications that go beyond dietary risk analysis.

1.4 objectives and main contributions

In this thesis, we tried to bring some answers to the statistical problems mentioned in the previous section. It is structured as follows: in Chapter 2, we tackle the issue of heavy-tailed long-term exposure to one specific chemical and show how extreme value theory can be of help. Moving then to the multivariate setting, the simultaneous exposure to multiple food components is considered. A new method mixing machine learning algorithms and spectral measure estimation to reduce the dimension in the analysis of multivariate extreme values is introduced in Chapter 3. In Chapter 4, while still remaining in a multidimensional optic, we change focus and go back to average phenomena. There, statistical learning methods borrowed from the minimum volume set literature are adapted to the non-parametric estimation of the distribution of exposure to a collection of nutrients and contaminants. A natural extension to the construction of dietary habits that realize a compromise between toxicological risk and nutritional benefit is also proposed. Finally, Chapter 5 and Chapter 6 are dedicated to the treatment of survey data. After extending classical functional results on empirical processes to the analysis of observations issued from a Poisson-like sampling scheme in the former, we introduce in the latter a novel estimator of the extreme value index for survey data. Although they are not general

enough to enable direct application to dietary risk analysis, these preliminary results constitute the basis of future developments that, we hope, will soon lead to more comprehensive models that can manage databases like INCA2.

1.4.1 Extreme chronic exposure to one chemical

Before even considering multivariate types of exposure to food chemicals, we start by discussing the calculation of the usual intakes of a unique nutrient or contaminant. Particularly interested in extreme phenomena, following in the footsteps of Tressou et al. (2004a), we bring back into fashion a non-parametric method that is usually disregarded in favor of mixed-effects models. It simply relies on the preliminary averaging of temporal observations on a daily scale, thereby avoiding the statistical modeling of the between and within variances that play a role in the distribution of the long-term exposure advocated by Tooze et al. (2010); van Klaveren et al. (2012); de Boer et al. (2009). When dealing with heavy-tailed distributions, thus proceeding is shown to avoid the under-estimation of tail characteristics that is bound to occur with the classical log-normal parametric modeling of usual intakes. Supporting our arguments with a real-data analysis, we propose a systematic procedure that consists in testing first the presence of a fat tail before choosing a specific statistical procedure. For this purpose, techniques directly borrowed from the extreme value theory literature are depicted and the suitability of this field for dietary risk analysis is brought into focus.

1.4.2 Simultaneous over-exposure to many food chemicals

Extending next our approach to the multivariate level, we consider the analysis of extreme types of exposure to many nutrients and contaminants. In the same spirit as the Pericles program (Crépet et al., 2013), we propose a new method that identifies cocktails of chemicals to which individuals are simultaneously highly exposed. Inspired by techniques of both the statistical learning and the extreme value analysis fields, it consists in assessing in a non-parametric manner the elements of the support of the spectral measure, a mathematical object that characterizes extreme dependencies. It is defined on the positive orthant of the unit sphere, which is quite naturally decomposable into open faces (edges and vertices) that happen to point out the variables that are linked together. In order to detect the faces on which the spectral measure is positive (thus the groups of variables that exhibit extreme dependence), we use a novel algorithm called Principal Nested Spheres (Jung et al., 2012) that achieves a sort of Principal Components Analysis on the unit sphere. It reduces the dimension of the data by systematically projecting the cloud of points on sub-spheres of lower dimension, thereby facilitating further analyzes. Tackling this

issue from a latent variable point of view justified by a mixture model of the spectral measure, the projected data is then clustered into groups that supposedly represent the different faces forming its support. Their identification is finally handled with a new heuristic that seems to perform well on simulations. When applied to the multivariate exposure obtained with the INCA2, CIQUAL 2008 and TDS2 databases, it produces comprehensible outcomes that support this promising approach.

1.4.3 A minimum volume set approach to dietary risk-benefit analysis

Exploiting further the assets of learning procedures for dietary risk analysis, we focus this time on average phenomena. In this chapter, we introduce for the first time the variability in the contamination process of the food. Following in the footsteps of Bertail and Tressou (2006), we extend the minimum volume set approach of Scott and Nowak (2006) to the case where the volume is unknown and estimated by a U-statistic. This enables to construct in a non-parametric manner the level sets of the multivariate distribution of types of exposure to multiple chemicals. We then demonstrate how this procedure can be modified to recover in the consumption space the dietary habits that balance toxicological risk and nutritional benefit. In the near future, these theoretical results are destined to be applied to the INCA2, CIQUAL 2008 and TDS2 databases in order to provide general dietary guidelines in the spirit of "eat five fruits and vegetables a day".

1.4.4 Empirical processes in survey sampling

Our next challenge concerns the nature of the consumption data, which often results from a complex survey scheme. Since the distribution of the long-term exposure is of particular interest for public health institutes like Anses, providing functional results about such weighted data would help avoid the bias induced when ignoring the survey scheme. Though unable to develop results for the specific design employed in INCA2 yet, we make our contribution to the elaboration of a more comprehensive theory by studying the asymptotic properties of empirical processes in the context of Poisson-like survey plans. Our approach is directly inspired by the seminal papers of Hàjek (1964) and Berger (1998) and exploits the proximity of large entropy designs to the simple Poisson scheme. Under some assumptions on the inclusion probabilities and a superpopulation framework, which are bound to be relaxed in the near future, we establish a functional central-limit theorem and show its implications for the asymptotic analysis of a large array of estimators. With illustrations based on simulations we present in particular how it can be applied to the construction of uniform confidence bands of the distribution function. Beyond dietary risk analysis, such results appear to be of particular interest for the management of huge databases,

the sizes of which increase in permanence like financial data, and therefore cannot be fully accessed. With complete control over the sampling procedure, the Poisson and rejective plans are revealed as especially convenient for the unbiased statistical analysis of such databases.

1.4.5 Tail index estimation based on survey data

Going back to extreme phenomenons, we propose to adapt the widely celebrated Hill estimator of the extreme value index (Hill, 1975) to the survey sampling framework. Hoping again to extend this results to the complex survey scheme of INCA2, we start by establishing its consistency and asymptotic normality for sampling plans of the Poisson type. The rate of convergence of this novel estimator is found to be the same as if the entire population was available, and the variance is only depreciated by a multiplicative term that depends on the way the inclusion probabilities were chosen. With numerical experiments we show that the restrictive hypotheses that were required in our theorems could actually be relaxed. Recalling the issue of big data, we finally encourage the extension of those preliminary results to many other branches of the extreme value field.

2 EXTREME CHRONIC EXPOSURE TO ONE CHEMICAL

2.1 introduction

When combined with food composition databases, dietary surveys prove useful to assess statistical distributions within a given population of intakes of nutrients as well as different types of chemical substances such as environmental contaminants, food additives or pesticide residues present in the food. Typically, they are built out of short-term follow-ups of representative sub-populations, commonly called 24h-recalls, chosen according to some appropriate survey scheme, and therefore do not enable direct estimation of chronic (long-term) risks. Indeed, participants in cross-sectional representative dietary surveys are usually not solicited more than one week, and the current trend is even to shorten the survey duration to only two or three days (see for instance the reports of the EFCOVAL project on http://www.efcoval.eu/ and Crispim et al., 2011; De Boer et al., 2011). As a result, chronic risk evaluation may not be achieved from such data without any statistical modeling.

In answer to these practical limitations, statistical models have been developed in the last 20 years to estimate usual long-term intakes from short-term measurements (van Klaveren et al., 2012; Dodd et al., 2006). Originally constructed to assess the unbiased prevalence in some population of interest of inadequate and insufficient intakes of nutrients, they are also used now to estimate high percentiles of intake in a food safety perspective. In nutrition, there is for instance a need to verify that upper levels of intakes for vitamins and minerals are not exceeded, considering food fortification or supplements intake. Most of these models use an analysis of variance to separate between- and within-individuals variabilities of usual intakes. An assumption of normality of intakes after a Box-Cox transformation is done (Tooze et al., 2010). The present study tries to propose an adaptation of these methods to better consider high nutrient intakes, which are often not normally distributed, in order not to underestimate the risk of exceeding upper levels. This method could also be applied to chronic risk assessment to chemicals in food (Boon et al., 2011). This work corresponds to a paper currently being written in collaboration with J.L. Volatier (Anses, France).

The chapter is structured as follows. We start off in Section 2.2 by setting notations and presenting the statistical background on which classical models are based. After

reviewing in further detail two popular methods and underlining their limits relative to the estimation of high quantiles and rare events, we thoroughly depict our alternative procedure in Section 2.3.1. As means of illustration, a case-study is conducted in Section 2.4. The different models previously introduced are subsequently compared on real data, which finally leads to the discussion of their respective assets and liabilities in Section 2.5. Supplementary details regarding tail estimation are provided in Section 2.6.

2.2 mixed-effects models to assess usual intakes

The setting under which usual intakes are analyzed is always the same. We observe food consumption of $n \geq 1$ individuals, indexed by i, taken randomly within a given population, during $J \geq 2$ days. Let X_{ij} be the true nutrient (or contaminant) intake of individual i on day j, and $X_i = E[X_{ij}|i]$ their usual intake. Here, $E(.|i)$ represents the expectation conditional on being individual i. Because it is not possible to observe X_{ij} directly in a chosen population, we use a proxy, denoted by W_{ij}, obtained by crossing consumption with composition data. Specifically, let C_{ij}^h be the amount of food h ingested by individual i on day j, and Q^h the average level of nutrient (or contaminant) contained in h. We observe $\mathcal{C}_{ij}^h$, an approximation of C_{ij}^h reported by individual i during a survey, and $\mathcal{Q}^h$ an estimated version of Q^h obtained via multiple measurements on various samples of food h. It is assumed that $E[\mathcal{C}_{ij}^h|i,j] = C_{ij}^h$ and $E[\mathcal{Q}^h] = Q^h$, i.e. data collection methods produce unbiased estimates of the quantities of interest. Moreover, individuals are supposed to choose and report their food independently from their contents, which implies in particular that $E[\mathcal{C}_{ij}^h \mathcal{Q}^h] = E[\mathcal{C}_{ij}^h] E[\mathcal{Q}^h]$. Then, for a given nomenclature of H foods, $W_{ij} := \sum_{h=1}^{H} \mathcal{C}_{ij}^h \mathcal{Q}^h$ is an unbiased estimate of X_{ij}, and one can write the following measurement error model, for all $i \in \{1,\dots,n\}$, $j \in \{1,\dots,J\}$:

$$W_{ij} = X_i + \varepsilon_{ij}, \tag{2.1}$$

where ε_{ij} is a within-individual error with null expectation. A few additional constraints are assumed to hold:

- (C_1) $X_1,\dots,X_n$ are independent, identically distributed (iid) with cumulative distribution function (cdf) F, independent from noise terms ε_{ij},
- (C_2) within-person errors are individually (over $i = 1,\dots,n$) and temporally (over $j = 1,\dots,J$) independent from each other, with common cdf F_ε.

Notice that in practice, (C_1) requires in particular that two people of the same household are not both included in the sample and that (C_2) excludes surveys like INCA2, where the days of observation are consecutive. Many different techniques were developed from this general background (van Klaveren et al., 2012; Dodd et al., 2006).

Though all based on the same master model, they can substantially differ on many points. Depending on which family of distributions both F and $F_{\cdot}$ are assumed to belong to, confidence intervals may or may not be computable, episodic consumption taken into account, auxiliary information exploited, etc. The final objective also plays a role in the choice of method, conservatism being sometimes preferred to realism (we would rather have a systematic over-estimation of the quantities of interest than taking the risk of under-estimating them). These techniques are regularly evaluated and compared by scientific committees mandated by public health institutes, in order to provide methodological guidelines to practitioners (see for instance Sections 2, 3 and 4 of the latest EFSA report van Klaveren et al., 2012, or the older recommendations of U.S. Environmental Protection Agency, 1999; WHO, 2000; EFSA, 2006). Here we present two major approaches advocated by EFSA, namely the BetaBinomial-Normal (BBN) and LogisticNormal-Normal (LNN) models, respectively introduced by de Boer et al. (2009) and Tooze et al. (2006, 2010).

2.2.1 The BetaBinomial-Normal model

Referring to de Boer et al. (2009), the BetaBinomial-Normal model, abbreviated BBN, is particularly appreciated because it enables to deal with episodic consumption. It relies on the simple decomposition:

$$F(x) = P(X_i = 0) + P(X_i \leq x \mid X_i > 0)\, P(X_i > 0).$$

Specifically, it considers that the probability of consuming (or intake frequency), namely $p_0 := P(X_i > 0)$, has Betabinomial distribution, while positive intake amounts, denoted by X_i^+, are normally distributed. Since it can happen that the distribution of X_i^+ is asymmetric, a preliminary logarithmic or power transformation of the data is usually recommended (Box and Cox, 1964). The BBN model further requires that p_0 and X_i^+ are independent. This prohibits for instance situations where the occasional intakes of episodic consumers are systematically small. In practice, estimation is achieved in two separate steps, one dealing with p_0 and the other with X_i^+, as detailed herein-after.

Step 1 Using both null and non-null observations, the parameters of the Betabinomial distribution of p_0 are assessed by means of maximum likelihood estimation.

Step 2 Relying solely on the positive observations, denoted by $x_{i;j}^+$, a preliminary power transformation is achieved:

$$g_\lambda(x_{i;j}^+) := \begin{cases} \log(x_{i;j}^+) & \text{if } \lambda = 0; \\ (x_{i;j}^+)^\lambda & \text{if } \lambda > 0. \end{cases}$$

To choose an appropriate λ, a grid of candidate values is explored and the retained value $\hat{\lambda}$ is that minimizing the sum of squared residuals ensuing from

a regression of normal Blom scores on the transformed intakes (de Boer et al., 2009, p.1436). Then, the parameters of the following model are fitted with the maximum likelihood procedure:

$$g_{\square\square}\left(\square^{+}_{i;j}\right) = \square + c_i + u_{i;j};$$

where c_i is the between-person effect with Normal distribution $N(0; \square^2_B)$ and $u_{i;j}$ the within-person effect with Normal distribution $N(0; \square^2_W)$.

Once all model parameters assessed, the distribution F of the usual intakes is obtained by Monte-Carlo approximation. First, realizations of p_0 are simulated according to its estimated distribution. When they are equal to 1 (i.e. consumption is supposed to occur), the amount is drawn from a normal distribution $N(\hat{p}; \hat{p}^2_B)$, where $\hat{p}$ and $\hat{p}^2_B$ are the estimated versions of $\square$ and $\square^2_B$ respectfully. Concretely, this comes to annihilating the intra-variation effect encapsulated by $u_{i;j}$, which represents the individual fluctuations around the corresponding long-term dietary habits. Finally, a back-transformation is performed to go back to the initial scale (before applying the function $g_{\square\square}$ to the raw data). Notice that it is also possible to add auxiliary information to the procedure by introducing covariates in the models of p_0 and $\square^{+}_{i;j}$ (de Boer et al., 2009, p.1436).

2.2.2 The LogisticNormal-Normal model

The LogisticNormal-Normal model of Tooze et al. (2010), first introduced in Tooze et al. (2006), is very similar to the BBN method depicted above. The main differences are the following.

– The transformation function is of the Box-Cox form (Box and Cox, 1964)

$$g_{\square}\left(\square^{+}_{i;j}\right) := \begin{cases} \log\left(\square^{+}_{i;j}\right) & \text{if } \square = 0; \\ \frac{1}{\square}\left(\square^{+}_{i;j}\right)^{\square} - 1 & \text{if } \square \neq 0; \end{cases}$$

and the choice of an optimal $\square$ is directly handled in the maximum likelihood estimation of the parameters of the distribution of $\square^{+}_{i;j}$.

– Correlation between p_0 and $\square^{+}_{i;j}$ can be introduced.

– The parametric model of p_0 is no longer Betabinomial, but Logistic-Normal.

In both methods (BBN and LNN), statistics of interest such as percentiles can be estimated using the simulated usual intakes. Unfortunately, none of these approaches provides associated confidence intervals, since it would require controlling at the same time the variance of the estimates of the various model parameters, that of the Monte-Carlo approach and the error due to the back-transformation.

Obviously, many more methods than these two seminal contributions have been proposed in the literature. We refer to van Klaveren et al. (2012, Sections 2, 3 and 4)

and Dodd et al. (2006) for a more detailed account of the available models for usual intakes.

2.3 heavy-tail modeling of usual intakes

All the methods that were just introduced share a common characteristic: up to a transformation, both usual intakes and the noise terms are supposed normally distributed. If fitting a Gaussian distribution can convey acceptable approximations of average phenomena, it rarely is the case when focus is on extreme objects such as the 95th and 99th percentiles, or the probability of getting over some maximum intake limit. This issue is clearly mentioned in de Boer et al. (2009, p.1438): "departures from normality may give biased estimation of the model parameters and, hence, may give wrong inference about the usual intake distribution". In the next paragraph, we review the basics of extreme value theory that are of help to understand the validity of this statement. Given this background, we then propose an alternative model dedicated to the assessment of the tail of the distribution of positive usual intakes (we are not interested in episodic consumption).

2.3.1 Extreme value theory and heavy-tailed distributions

In a univariate context, extreme value theory is dedicated to the analysis of the tail of some distribution of interest, corresponding here to $\bar{F}(x) := 1 - F(x)$ for all large intakes $x \in \mathbb{R}_+$. Notice that it is intricately linked to the tail quantile function, defined for all $x \geq 1$ as

$$U(x) := \inf\left\{ y \in \mathbb{R}_+ : \bar{F}(y) \leq \frac{1}{x} \right\};$$

by setting $p := 1/x$ for some $p \in (0,1)$, it is easy to see that $U(1/p)$ is no other than the quantile of order $1 - p$ of the studied usual intakes, also denoted by

$$Q(p) := \inf\{ y \in \mathbb{R}_+ : F(y) \geq p \}.$$

Given a collection of $n \geq 1$ iid observations $X_1, \ldots, X_n$ with cdf F, extreme value theory (EVT) stipulates that observations in the tail of the distribution are approximately governed by the law of the maximum. Let $X_{1,n} \leq \cdots \leq X_{n,n}$ denote the order statistics related to the sample, with $X_{n,n}$ its maximum. Then, in the same spirit as the Central Limit Theorem for the mean, we are led to look for some real sequences of constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that, for all $n \geq 1$,

$$\lim_{n \to \infty} P\left[\frac{X_{n,n} - b_n}{a_n} \leq x \right] = G(x), \tag{2.2}$$

for any $x \in \mathbb{R}$ at which the limiting distribution function G is continuous. If such normalizing constants exist, then G is called an extreme value distribution and F is said

to be in the maximum domain of attraction of G (abbreviated FPMDA(G)). Going back to the tail of the distribution and referring to De Haan and Ferreira (2006, Theorem 1.1.2), Equation (2.2) is equivalent to

$$\lim_{t \to \infty} t \bar{F}\left(a_{\lfloor tu \rfloor} x + b_{\lfloor tu \rfloor}\right) = -\log G(x); \tag{2.3}$$

with $\lfloor tu \rfloor$ the integer part of t. Fortunately, (univariate) extreme value distributions can be written in a parametric form, introduced first by Fisher and Tippett (Fisher and Tippett, 1928) then by Gnedenko (Gnedenko, 1943):

$$G(x) = G_\gamma(x) := \exp\left(-(1+\gamma x)^{-1/\gamma}\right); \text{ for } 1+\gamma x > 0; \tag{2.4}$$

with, by convention, $G_0(x) = \exp(-e^{-x})$ and $\gamma \in \mathbb{R}$. Called the extreme value index (EVI), γ defines three types of maximum domains of attraction. Specifically, if $F \in MDA(G_\gamma)$ with $\gamma > 0$, then F is said to be in the Fréchet maximum domain of attraction, which characterizes heavy-tailed distributions like Pareto laws. Conversely, thin-tailed distributions such as the Normal or Lognormal laws belong to the Gumbel maximum domain of attraction, with $\gamma = 0$. The last maximum domain of attraction, called Weibull, verifies $\gamma < 0$ and encompasses distributions with bounded upper tails. The most classical example of the latter would be the Uniform distribution on some closed interval. Of course, it can so happen that some distributions belong to neither of these three maximum domains of attraction. Log-Pareto distributions with cdf $F(x) = 1 - (1+\gamma(\log x - u)/\sigma)^{-1/\gamma}$; $\gamma > 0$, $u \in \mathbb{R}$ and $\sigma > 0$, marked as super-heavy-tailed, are part of these exceptions (Cormann and Reiss, 2009). Further notice that only some types of thin-tailed (resp. bounded) distributions belong to the Gumbel (resp. Weibull) maximum domain of attraction (Embrechts et al., 2011, Sections 3.3.2 and 3.3.3). In fact, F can have a bounded upper tail and still verify $F \in MDA(G_0)$. Conversely, light-tailed distributions with bounded upper tail can belong to the Weibull maximum domain of attraction. The interested reader may refer to Embrechts et al. (2011, Chapter 3) for a very thorough introduction to the fluctuations of maxima, enlivened by many detailed examples. In particular, Tables 3.4.2, 3.4.3 and 3.4.4 from p.153 to p.157 in this manual display a list of characteristics of the most common distributions in each maximum domain of attraction, including some possible normalizing constants.

In the present study, we are interested in heavy-tailed distributions that respect $F \in MDA(G_\gamma)$, $\gamma > 0$. They are characterized as follows (Embrechts et al., 2011, Section 3.3.1):

$$1 - F(x) = x^{-1/\gamma} L(x); \tag{2.5}$$

where $\gamma > 0$ is the EVI and $L(x)$ a slowly varying function (svf), i.e. for all $t > 0$, $L(tx)/L(x) \to 1$ as $x \to +\infty$. Distributions of that form are said to be regularly varying with index $-1/\gamma$ (Resnick, 2007); the set of such functions is denoted by $R_{-1/\gamma}$. They enjoy quite a few useful properties, e.g. their tail quantile function is

also regularly varying with index γ. When concerned with the estimation of the EVI, Equation (2.5) is a typical semi-parametric model, where the svf L plays the role of the perturbation. Actually, to guarantee its accuracy, statistical inference often requires this disruptive function to fulfill some extra regularity criteria. Called Von Mises or second order conditions, they are usually written as follows (cf. Goldie and Smith, 1987 and De Haan and Ferreira, 2006, Chapter 2).

Assumption 2.1 The regularly varying tail quantile function $U \in RV_\gamma$ with $\gamma > 0$ is such that there is a real parameter $\rho \leq 0$, referred to as the second order parameter, and a positive or negative function A with $\lim_{x \to +\infty} A(x) = 0$ such that for any $t > 0$,

$$\frac{1}{A(x)} \left(\frac{U(tx)}{U(x)} - t^\gamma \right) \xrightarrow[x \to \infty]{} t^\gamma \frac{t^\rho - 1}{\rho};$$

or equivalently

$$\frac{1}{A\left(\frac{1}{\bar{F}(x)}\right)} \left(\frac{\bar{F}(tx)}{\bar{F}(x)} - t^{-1/\gamma} \right) \xrightarrow[x \to \infty]{} t^{-1/\gamma} \frac{t^{\rho/\gamma} - 1}{\rho/\gamma}.$$

This ensures that L is almost of the form of a Hall slowly varying function $1 + cx^{\rho\gamma}$, which makes its influence over the Pareto form $x^{-1/\gamma}$ of $\bar{F}$ controllable. There, the parameter ρ controls the speed of convergence in extremes and, as we shall see later on, plays a crucial role in the asymptotic analysis of most estimators.

In terms of nutrient or contaminant intakes, assuming X_i has cdf F as in Equation (2.5) means that the larger γ, the less the probability that X_i reaches high levels is negligible. On the contrary, constraining F to be in the Gumbel maximum of attraction comes to considering that extreme events can be rare enough to be disregarded. Actually, even if normality is not attributed to the raw data but to its Box-Cox-transformed counterpart, usual intakes will still be implicitly assumed thin-tailed (Teugels and Vanroelen, 2004; Wadsworth et al., 2010). Depending on which of these two hypotheses is privileged, corresponding large percentiles and probabilities of exceeding some maximum intake limit may differ significantly. In particular, dietary risks are bound to be drastically under-estimated if heavy tails are ignored. So as to decide which model to apply to our data, one may use statistical testing procedures borrowed from the literature in EVT such as those depicted in the next section.

2.3.2 Testing the heavy-tail assumption

Because constraining F to be thin-tailed when it is in fact heavy-tailed may induce serious bias in tail estimation, we would like to test the null hypothesis

$$H_0 : F \in MDA(G_\gamma),\ \gamma > 0 \quad \text{versus} \quad H_1 : F \in MDA(G_\gamma),\ \gamma \leq 0.$$

Since we do not observe X_i directly but rather a collection of $(Y_{i,j})_{1 \leq j \leq J}$ as in Equation (2.1), we cannot apply testing procedures directly. An extra assumption has to be made to enable further analysis and we will consider the following conditions.

Assumption 2.2 The cumulative distribution functions F and F_ε are such that

$$\exists \ell \in [0, +\infty] : \frac{1 - F_\varepsilon(x)}{1 - F(x)} \xrightarrow[x \to -1]{} \ell \quad \text{and} \quad \forall t \in (0,1), \limsup_{x \to +\infty} \frac{1 - F(tx)}{1 - F(x)} < +\infty .$$

Assumption 2.3 The cumulative distribution functions F and F_ε are such that

$$\frac{1 - F_\varepsilon(x)}{1 - F(x)} \xrightarrow[x \to -1]{} 0 \quad \text{and} \quad \exists x_0 < +\infty : \forall x > x_0, F_\varepsilon(x) = 1.$$

If either Assumption 2.2 or Assumption 2.3 holds, then under the setting introduced at the beginning of Section 2.2 the cdf of $Y_{i,j}$, denoted by $F_{i,j}$, is in the Fréchet maximum domain of attraction with EVI $\gamma > 0$ if and only if $F \in MDA(G_\gamma)$, as shown in Maddipatla et al. (2011, Theorem 3.3 and Remark 3.4). The first mathematical constraint in Assumption 2.2 simply means that the tail of F_ε should either be equivalently thick or thinner than that of F. Without this assumption, we cannot tell anything about F based solely on $Y_{i,j}$. Though it may seem quite restrictive and cannot be tested, such a requirement encompasses a large class of distributions for both F and F_ε. In particular, F_ε can still be heavy-tailed, as long as its EVI does not exceed that of F. Notice that if the former is indeed in the Fréchet maximum domain of attraction with EVI γ_ε, then to allow the existence of its expectation we need to have $\gamma_\varepsilon < 1$. Similarly, for $\varepsilon_{i,j}$ to have a finite variance, it is required that $\gamma_\varepsilon < 1/2$. We refer to Embrechts et al. (2011, Section A3), De Haan and Ferreira (2006, Chapter 1) or Beirlant et al. (2004, Chapter 2) for more details on these properties. Hence, desirable regularity conditions on individual errors limit the thickness of the tail of their distribution, thereby advocating the reasonableness of our required assumption. At this point, we still cannot use classical testing techniques, because our sample of observations $\{Y_{i,j} ; 1 \leq i \leq n ; 1 \leq j \leq J\}$ is not iid. However, it suffices to average the data over time to get into an appropriate setting. According to the model depicted in Section 2.2, we have

$$\overline{Y}_i := \frac{1}{J} \sum_{j=1}^{J} Y_{i,j} = X_i + \frac{1}{J} \sum_{j=1}^{J} \varepsilon_{i,j} =: X_i + \overline{\varepsilon}_i ; \tag{2.6}$$

where all $\overline{Y}_i$, $1 \leq i \leq n$, are iid with cdf $F_{\overline{Y}}$. Further denote by $F_{\overline{\varepsilon}}$ the cdf of $J^{-1} \sum_{j=1}^{J} \varepsilon_{i,j}$, then if $F_{\overline{\varepsilon}}$ fulfills Assumption 2.2 or Assumption 2.3 instead of F_ε, we still have the equivalence $(F \in MDA(G_\gamma);\ \gamma > 0) \Leftrightarrow (F_{\overline{Y}} \in MDA(G_\gamma))$. This requirement is met for a large set of distributions. In particular, if F_ε has a right upper bound, then so does $F_{\overline{\varepsilon}}$ and Assumption 2.3 is verified. Other examples of accurate

distributions can be found, for instance, in Embrechts et al. (2011, Sections 1.3 and A.3.2).

From now on, we will consider the model in Equation (2.6) and assume that $F_{\leftarrow}$ and F fulfill either Assumption 2.2 or Assumption 2.3. To test whether F (or equivalently $F_{\overline{X}}$) is in the Fréchet maximum domain of attraction, we first have to check beforehand that it does indeed belong to a maximum domain of attraction. For this purpose, one can use the statistical test introduced by Dietrich et al. (2002) and studied at length in Hüsler and Li (2006). It is designed to check H_0: $F \in MDA(G_\gamma)$ for some $\gamma \in \mathbb{R}$, and relies on the statistic

$$E_{n;k} := k \int_0^1 \left(\frac{\log \overline{X}_{n-\lfloor tk \rfloor, n} - \log \overline{X}_{n-k,n}}{p_+} - \frac{t^{-p_-} - 1}{p_-}(1-p_-) \right)^2 t^{\gamma} dt,$$

where $\overline{X}_{1,n} \leq \ldots \leq \overline{X}_{n,n}$ denotes the order statistics relative to the sample of average intakes, $\gamma \neq 0$ is a parameter to be chosen, p_+ and p_- are the moment estimators of Dekkers et al. (1989), and k is a fixed number of upper values of $(\overline{X}_1, \ldots, \overline{X}_n)$ that are considered to be representative of the tail of their distribution. Under H_0 and for some appropriate choice of k, the quantiles of the asymptotic distribution of $E_{n;k}$ can be computed and compared to its actual value. For more details on this procedure, we refer to Hüsler and Li (2006), who also provide a R package implementing this test. In practice, we compute $E_{n;k}$ for various choices of k (from 10 to $0.3 \times n$ for instance), and accept H_0 if the resulting function almost always remains below the corresponding asymptotic quantiles.

Once it has been established that F is in some maximum domain of attraction, we may focus on testing H_0: $F \in MDA(G_\gamma)$, $\gamma \neq 0$ versus H_1: $F \in MDA(G_\gamma)$, $\gamma \leq 0$. Many procedures have been developed in the literature, as reviewed in Neves and Alves (2008). We propose to use the statistic introduced in Beirlant et al. (2006); it is based on a modified version of the Jackson statistic (Jackson, 1967) that tests the "exponentiality" in the tail of the log-transformed data. Under the second order conditions in Assumption 2.1 and provided that the parameter γ is consistently assessed (e.g. with an estimator of the class presented in Alves et al. (2003)), it converges as $n \to +\infty$ to a Gaussian distribution. Because it has a complicated form, for clarity purposes we do not provide the explicit formula of this statistic, but rather refer to the original paper (Beirlant et al., 2006). However, we precise that just like the aforementioned $E_{n;k}$, it depends on the choice of some number k of largest values. Again, we compute this statistic for k ranging from 10 to $0.3 \times n$, and accept H_0 if the resulting function generally stays in the acceptance interval.

2.3.3 Assessing extreme quantities

In this section we deal with the case where both tests depicted herein-before were accepted, i.e. where $F_{\overline{X}}$, thus F, was found to be heavy-tailed, in the Fréchet maximum

domain of attraction. We provide examples of widely-used estimators to assess the corresponding EVI γ, extreme percentiles and small probabilities of exceeding some maximum intake limit. Of course, there is a plethora of such estimators, that are eluded here in favor of the general advantage of our probabilistic model for dietary risk assessment. In particular, it is well-known that the statistical tools introduced herein-after can perform quite poorly in a number of situations (Embrechts et al., 2011, Remarks p.337-338), and alternative estimators may be preferred. However, the latter are often quite complicated and based on intricate mathematical results that go beyond the scope of the present chapter. For the sake of clarity, and because we are only interested here in the methodological aspect of the modeling of usual intakes, we deliberately choose to introduce simple, well-known, techniques as means of illustration. The interested reader may refer to comprehensive textbooks such as Embrechts et al. (2011); De Haan and Ferreira (2006); Beirlant et al. (2004); Resnick (2007); Reiss and Thomas (1997) or to Caeiro and Gomes (2008); Beirlant et al. (2012); Scarrott and MacDonald (2012) to get a wider overview of the state of the art on this topic.

Ú The EVI One of the most famous estimators of the EVI for heavy-tailed distributions is the Hill estimator (Hill, 1975), computed on a number k of largest values in the sample, supposedly representative of the tail of the underlying distribution:

$$H_{k;n} := \frac{1}{k} \sum_{i=1}^{k} \log \overline{X}_{n-i+1;n} - \log \overline{X}_{n-k;n}. \tag{2.7}$$

It corresponds to the maximum likelihood estimator of γ when the tail of the distribution is assumed to be exactly Pareto, i.e. $\bar{F}(x) = cx^{-1/\gamma}$ for some $c > 0$ and $x > x_0 > 0$ such that $\int_{x_0}^{1} -c/\gamma\, x^{-1/\gamma-1}\, dx = 1$. Under Assumption 2.1 and assuming $k = k(n) \to \infty$, $k/n \to 0$, $\sqrt{k}\, A(n/k) \to \lambda < \infty$ as $n \to \infty$, then

$$\sqrt{k}\,(H_{k;n} - \gamma) \xrightarrow[n\to\infty]{} \mathcal{N}\left(\frac{\lambda}{1-\rho}; \gamma^2\right) \text{ in distribution,}$$

see for instance Embrechts et al. (2011, Example 4.1.12 and Section 6.4.2, Method 2), Beirlant et al. (2004, Section 4.4) or De Haan and Ferreira (2006, Section 3.2). Obviously, one needs to select some acceptable value k on which to calculate $H_{k;n}$ while ensuring its asymptotic normality. This is a typical bias-variance dilemma: if k is small, the selected observations are likely to belong to the tail of the distribution, producing slightly biased estimates. However, they should be in small number, thereby implying a wide variance. On the contrary, as k increases more observations are used and the variance reduces, but the bias may grow, because observations that are not located in the tail of the distribution, and do not correspond to the stipulated model, may contaminate the estimation. To overcome this issue in practice, we propose to use the double-bootstrap algorithm of Danielsson et al. (2001), recommended in Gomes and Oliveira (2001).

Assessing γ usually represents a preliminary step to the estimation of some extreme quantities, such as high quantiles or small probabilities of exceeding a large threshold. We now review some classical estimators of both objects, based on $H_{k,n}$.

Extreme quantiles and the probability of exceeding maximum intake limits From the estimation of γ, by referring to Equation (2.4) one may easily build estimators to assess extreme quantiles or the probability of getting over some high threshold. Indeed, combining it with Equation (2.3) yields the following approximation, for some large enough $x \in \mathbb{R}_+$ and well-chosen $t > 0$:

$$\mathbb{P}\left(\bigcap_{i=1}^{n} X_i > x\right) = \mathbb{P}(X_i > x)^n \simeq \left(\frac{1}{t}\right)^n \left(1 + \gamma \frac{x - b(t)}{a(t)}\right)^{-n/\gamma}. \tag{2.8}$$

Notice that by setting $a(t) = \gamma U(t)$, $b(t) = U(t)$, $\gamma > 0$ and replacing x by $(x-1)/\gamma$ in Equation (2.3), we recover Equation (2.5). Moreover, if we set $t = n/k$ for some number k of largest observations, then the empirical counterpart of $U(n/k)$ is simply

$$U_n(n/k) := \inf\left\{ y \in \mathbb{R}_+ : \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i > y\} \leq \frac{k}{n} \right\} = X_{n-k,n}.$$

Coupling these remarks to the foregoing approximation in Equation (2.8) conveys quite natural estimators of both $Q(p) := U(1/(1-p))$, $p > 0.95$, the quantile of order p, and $\mathbb{P}(X_i > \ell_\gamma)$ for some maximum intake limit ℓ_γ (Embrechts et al., 2011, Section 6.4.2, Method 2):

$$\widehat{\mathbb{P}}(X_i > \ell_\gamma) := \frac{k^\gamma}{n}\left(\frac{\ell_\gamma}{X_{n-k,n}}\right)^{-1/H_{k,n}}; \tag{2.9}$$

$$\widehat{Q}(p) := \inf\left\{ y \in \mathbb{R}_+ : \widehat{\mathbb{P}}(X_i > y) \leq 1-p \right\} = X_{n-k,n}\left(\frac{n}{k^\gamma}(1-p)\right)^{-H_{k,n}}; \tag{2.10}$$

where k^γ is the optimal number of largest observations obtained with the double-bootstrap algorithm of Danielsson et al. (2001). Confidence intervals may be computed via the asymptotic result below (Caeiro and Gomes, 2008, Proposition 1.2), assuming again that $k = k(n) \to \infty$, $k/n \to 0$ and $\sqrt{k}\,A(n/k) \to \lambda < \infty$ as $n \to \infty$:

$$\frac{\sqrt{k}}{\log\frac{k}{n(1-p)}}\left(\frac{\widehat{Q}(p)}{Q(p)} - 1\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(\frac{\lambda}{1-\gamma}; \gamma^2\right). \tag{2.11}$$

Equipped with these statistical tools, we now illustrate our approach on a set of nutrients, namely iron, zinc, calcium and retinol. Results are subsequently compared to those obtained with the BBN and LNN methods recommended by EFSA, which were briefly presented in Section 2.2.

2.4 case study: estimating the tail of usual intakes of iron, zinc, calcium and retinol in the french population

2.4.1 Description of the data

To assess tail characteristics of the distribution of usual intakes for iron (Fe), zinc (Zn), calcium (Ca) and retinol (arbitrarily abbreviated Re), we crossed the INCA2 consumption database with the levels of nutrients within foods of the CIQUAL base. They are both described at length in Chapter 1. So as to evaluate the impact of the number of reporting days in our model, we built a 24h-recall like data from the weekly information of INCA2. Specifically, among all 7 days considered in the latter, we selected one day of the week-end and one day of the week, with at least 3 days of interval, to mimic the sampling scheme of 24h-recall surveys. These declared amounts of ingested food were then combined with informations on their nutritional composition to produce estimates of individual intakes. We precise that no null intake was observed on the considered data, hence there is no need to model intake frequencies, denoted by p_0 in the presentation of the BBN and LLN methods (Section 2.2).

2.4.2 Results

We start this statistical analysis by computing mean intakes over all dates $J = 2; 7$, as in Equation (2.6). Before applying a heavy-tail model to all 4 types of nutritional intakes, we use the testing procedure of Dietrich et al. (2002) mentioned in Section 2.3.2 to check if their distributions belong to one of the three possible maximum domains of attraction. Results are displayed in Figure 2.1. Given that the corresponding test statistics mostly remain in the acceptance area when k varies, the distributions of intakes of Fe, Zn and Ca may well respect Equation (2.2). However, it is obviously not the case with Re; given the shape of its histograms in Figure 2.2, we suspect it is in fact super-heavy-tailed. Because our model does not encompass such distributions, we remove retinol from further analyses.

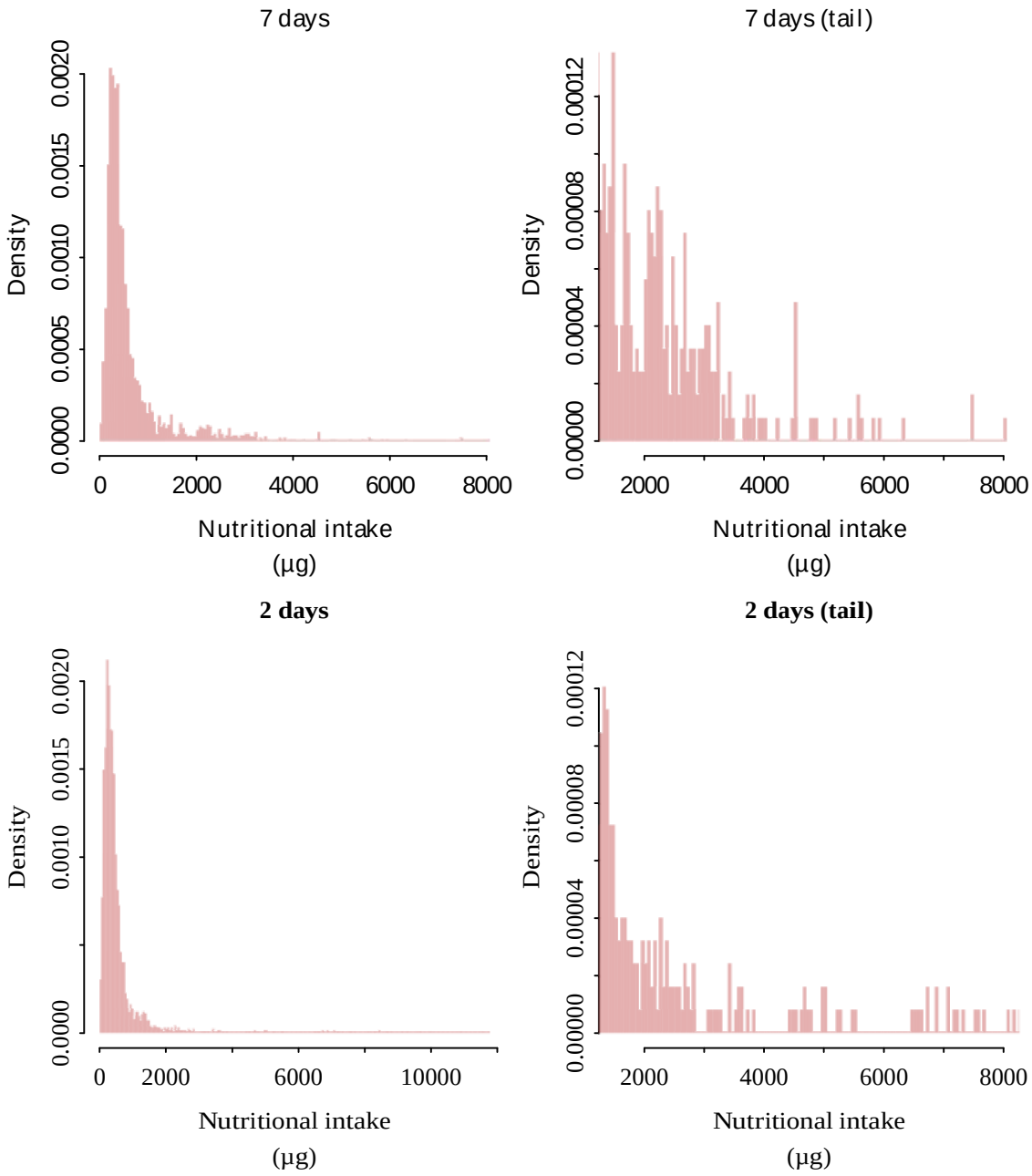


Figure 2.2 – Histograms of intakes of retinol

Nutrients that passed the former test are then subjected to the second statistical test of Beirlant et al. (2006), designed to detect distributions belonging to the Fréchet maximum domain of attraction (heavy-tailed). Results are presented in Figure 2.3. Again, statistics remain in the acceptance region whatever k, especially with Fe, thereby suggesting that the distributions of iron, zinc and calcium are all heavy-tailed.

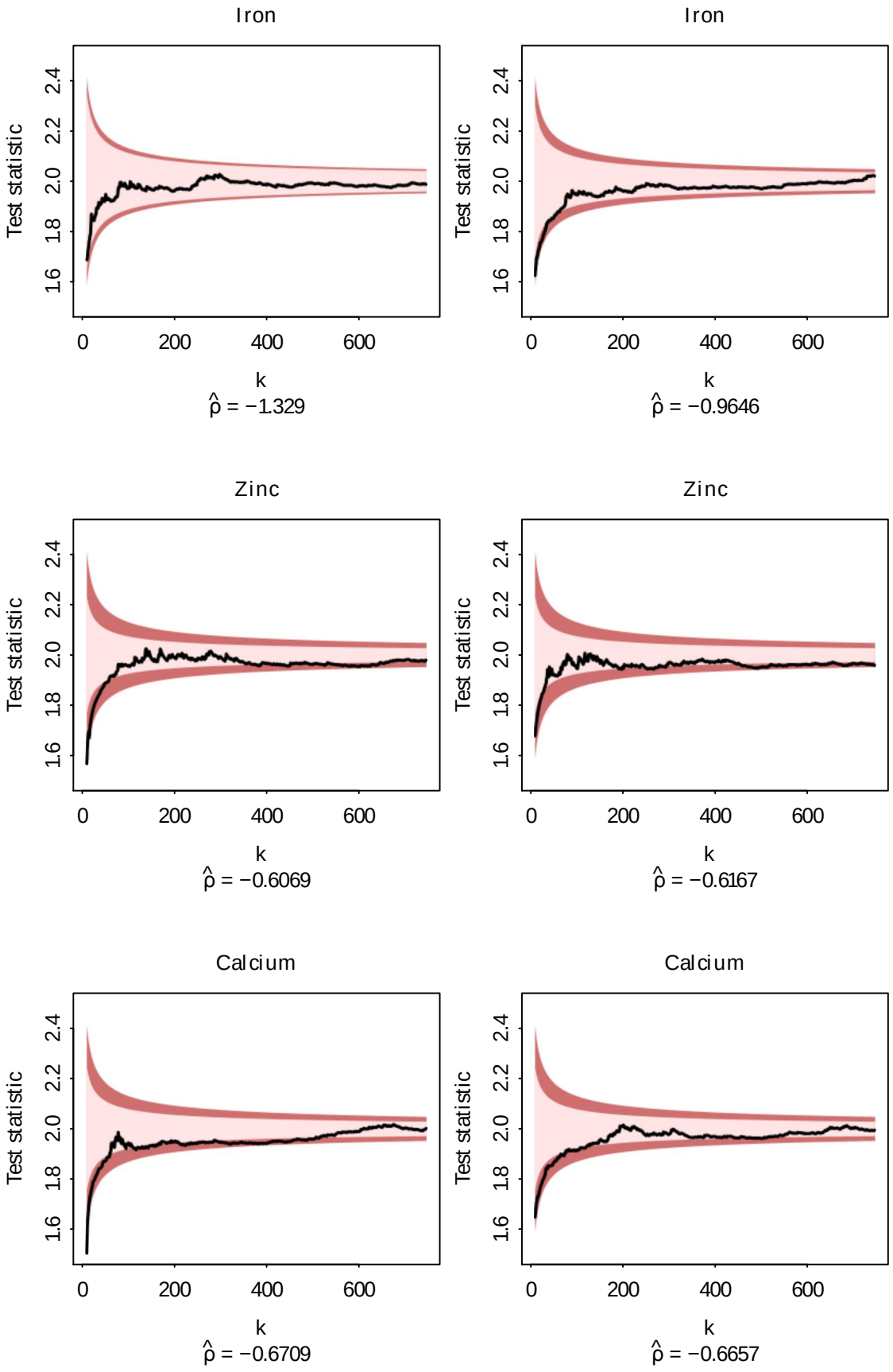


Figure 2.3 – Fluctuations of the test statistic of Beirlant et al. (2006) with k on the 30% largest values of the sample, and the corresponding acceptance regions (light pink area for $\alpha = \hat{\rho}$ and dark pink area for $\alpha = \alpha/2$) for iron, zinc and calcium intakes over 2 days (left hand plots) and 7 days (right hand plots) of observation

It is now possible to estimate the EVI γ for all three distributions, extreme quantiles and probabilities of getting over some maximum intake limit, with the techniques introduced in Section 2.3.3. For comparison purposes, we also assess the last two mentioned quantities with the LLN methods recommended by EFSA, using the source SAS code available at http://riskfactor.cancer.gov/diet/usualintakes/macros.html. Results are displayed in Table 2.1 and Table 2.2.

Table 2.1 – Estimates of the EVI based on the Hill estimator (standard errors in parentheses) and associated optimal numbers of largest values

	k_γ		$H_{k_\gamma;n}$	
Nutrient	2 days	7 days	2 days	7 days
Iron	693	554	0.268 (0.010)	0.218 (0.009)
Zinc	19	6	0.158 (0.036)	0.134 (0.055)
Calcium	336	82	0.192 (0.010)	0.143 (0.016)

Table 2.2 – Estimates of the 95-th and 99-th percentiles using the Hill estimator (standard errors in parentheses) and the LLN method (in mg/day)

	Quantiles					
	Hill (7 days)		Hill (2 days)		LLN (2 days)	
Nutrient	95%	99%	95%	99%	95%	99%
Iron	20.51	29.13	22.36	34.41	19.40	23.53
	(0.014)	(0.029)	(0.017)	(0.034)		
Zinc	14.59	18.09	17.52	22.60	15; 25	17; 97
	(0.165)	(0.078)	(0.068)	(0.010)		
Calcium	1427	1798	1479	2014	1371	1645
	(0.007)	(0.019)	(0.010)	(0.027)		

In view of Table 2.1, if the various tests performed on the data revealed that they were heavy-tailed, their EVI is relatively small, never exceeding 1=3. Provided these results are correct, this suggests that moments of order at least 3 exist for the distributions of iron, zinc and calcium. Moreover, it seems that observing only 2 days of consumption results in a significant increase in the heaviness of the tail. This phenomenon is naturally passed on to the quantile estimates, as can be observed in Table 2.2. This is relatively natural, since the smaller the period of observation, the further away we are from long-term habits by averaging the daily intakes. Even with 7 days, using the Hill estimator makes quantile estimates significantly higher than with the LLN model. Given that these usual intakes were found to be heavy-tailed, it is probable here that the LLN approach under-estimates extreme quantiles. From

a risk assessment point of view, it is thus safer to rely on extreme value theory than on classical methods to assess tails of distributions.

2.5 discussion

We have shown in this chapter how classical methods like the BetaBinomial-Normal and LogisticNormal-Normal presented in Section 2.2 can severely under-estimate the characteristics of the tail of the distribution of heavy-tailed usual intakes. Far from rejecting these approaches, we underlined in Section 2.3.2 that preliminary testing of the presence of a fat tail can be of substantial help when trying to decide which method to apply to the data. The nutrients that were taken as examples in Section 2.4 happened to pass these tests, but there are many other substances for which normality (up to a Box-Cox transformation) is a reasonable assumption. When it is not, we proposed in Section 2.3.3 simple methods to estimate extreme quantities such as high percentiles. Of course, EVT overflows with more refined statistics, which help avoid the classical problems of the Hill estimator (asymptotic bias). In particular, for distributions in the Fréchet domain of attraction with a very small index □, Cai et al. (2011) have recently developed a family of dedicated estimators. Regression models for extreme values also exist and would permit the insertion of covariates such as the age or the sex of individuals (Beirlant et al., 2004, Chapter 7). Such extensions are of major interest, since nutritional recommendations usually depend on auxiliary variables (women or children do not have the same nutritional needs as adult men). Beyond improvement of the estimation, more work remains to be done, in particular to include the modeling of intake frequencies in our approach. This is left for future research.

2.6 supplements: on the second order parameter □

Let us dwell for a moment on the estimation of the second order parameter □. To approximate its value, we used the family of estimators proposed by Alves et al. (2003) and the tuning parameters they recommend:

$$\rho(k;\square) = \rho_{n|T}^{(1;2;3;\square)}(k) = \frac{3\left(T_n^{(1;2;3;\square)}(k)\right)^{\square} - 1}{T_n^{(1;2;3;\square)}(k) - 3};$$

with $\alpha \in \{0, 0.5, 1, 2\}$,

$$T_n^{(1,2,3,\alpha)}(k) = \begin{cases} \dfrac{M_n^{(1)}(k)^{\alpha} \cdot M_n^{(2)}(k)^{2\alpha}}{M_n^{(2)}(k)^{2\alpha} \cdot M_n^{(3)}(k)^{3\alpha}} & \alpha \neq 0, \\ \dfrac{\log M_n^{(1)}(k) \cdot \log M_n^{(2)}(k)^{2}}{\log M_n^{(2)}(k)^{2} \cdot \log M_n^{(3)}(k)^{3}} & \alpha = 0, \end{cases}$$

and

$$M_n^{(m)}(k) = \frac{1}{k} \sum_{i=1}^{k} \left(\log X_{(n-i+1)} - \log X_{(n-k)} \right)^m, \quad m \in \{1, 2, 3\}.$$

For the results can be very volatile with k, they recommend to choose a final $\hat{p}$ where $p(k; \alpha)$ is the most stable, usually for large k. Therefore, we start by selecting the value of $\alpha \in \{0, 0.5, 1, 2\}$ that yields the most stable sample path $\{k, p(k; \alpha)\}$; we consider that a path is stable if the difference between two successive occurrences is small for a relatively long period. Thus, for a fixed α, we calculate

$$D_k^{\alpha} := p(n-k+1; \alpha) - p(n-k; \alpha)$$

for $k \in \{1, \ldots, n-1\}$ and get the cumulative empirical variance:

$$V_k^{\alpha} = \frac{1}{k} \sum_{i=1}^{k} (D_i^{\alpha})^2 - \left(\frac{1}{k} \sum_{i=1}^{k} D_i^{\alpha} \right)^2, \quad k \in \{1, \ldots, n-1\}.$$

The α that minimizes V_k^{α} for a maximum of occurrences of k is then selected and denoted by α_0. In a second step, we identify the values of k for which $V_k^{\alpha_0}$ is small, e.g. below the median of $\{V_k^{\alpha_0}\}_{1 \leq k \leq n-1}$, and calculate our final estimate $\hat{p}$ as the empirical mean of $p(k; \alpha_0)$ on these k. Results on exposures to our 4 nutrients are displayed in Figure 2.4.

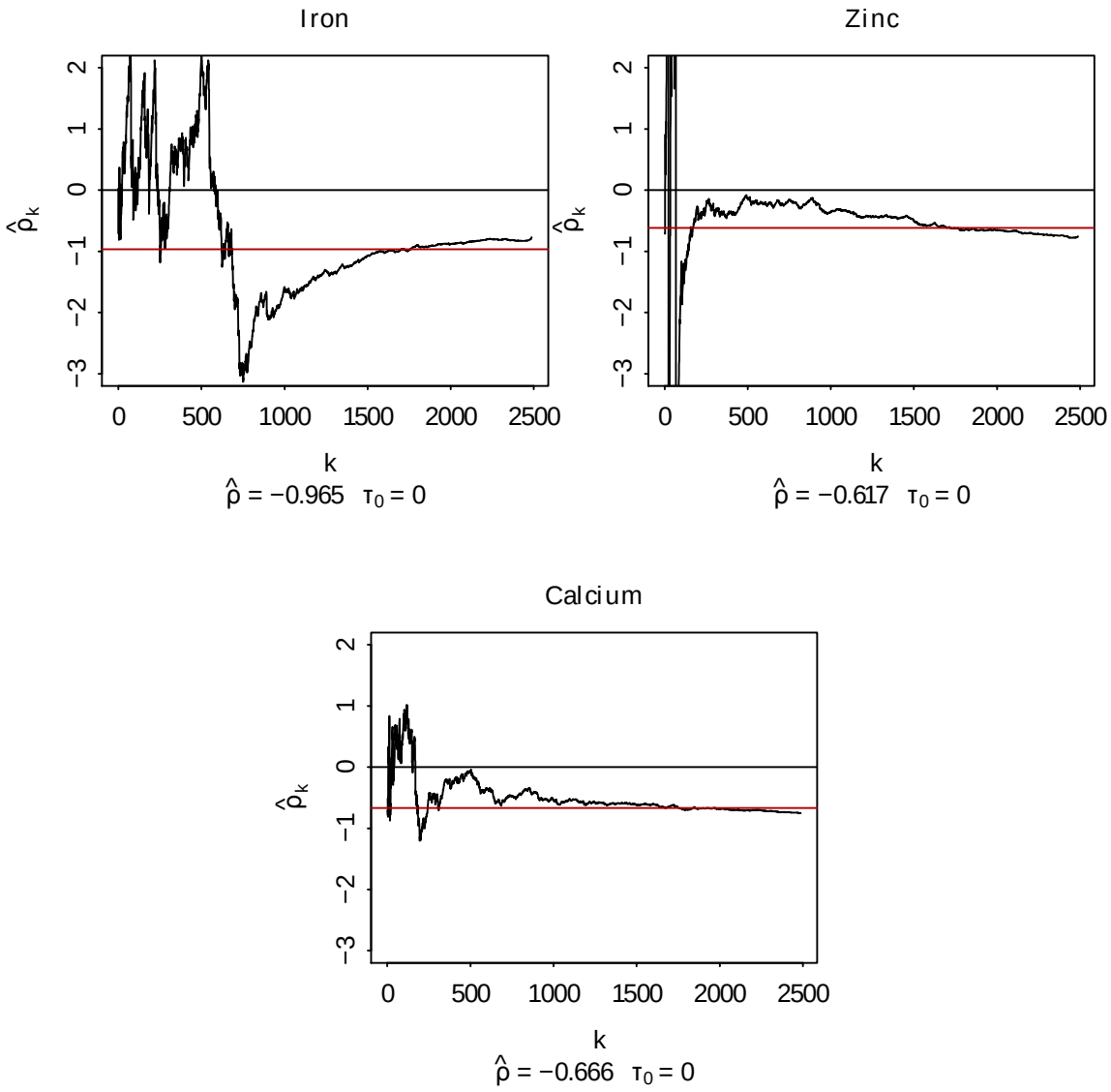


Figure 2.4 – Fluctuations of the estimated second order parameter with k on and the corresponding optimal value (red line) for iron, zinc and calcium intakes over 7 days of observation

3

SIMULTANEOUS OVER-EXPOSURE TO MANY FOOD CHEMICALS

In Chapter 2, we focused for a while on univariate extreme types of exposure to food nutrients or contaminants. The objective was to analyze the tail of the distribution of the chronic exposure to some chemical. It allowed in particular to estimate the small proportion of individuals in a given population who exceed some tolerable doses (the so-called dietary intake limits) and are thus likely to develop serious health problems in the long run. However, by analyzing only one component at a time, we ignored further noxious effects that may be caused by possible interactions between elements that are ingested simultaneously (Carpenter et al., 2002). Evaluating the sanitary impact of cocktails of nutrients and contaminants is a hot and complex topic public health institutes are currently mobilizing efforts on. For instance, Anses recently launched the Pericles research program (PEsticide Residue In vitro Combined Level of Exposure Study), which is dedicated to the identification and quantification of the risk due to the exposure to mixtures of pesticides (Crépet et al., 2013; Crépet et al., 2012; Crépet and Tressou, 2011; Béchaux et al., 2013). The present chapter corresponds to a paper written in collaboration with S. Clémençon (Télécom ParisTech, France) and recently submitted for publication, in which we develop a statistical methodology to find groups of any number of chemicals that are jointly absorbed in high quantity in the population of interest.

3.1 statistical challenges and objectives

High dimension raises important issues in applied multivariate statistics; while sample sizes are finite, the set on which probability measures are defined can be so large that extrapolation is intricate. Referred to as the curse of dimensionality (Donoho, 2000), this phenomenon makes the variance of classical estimators explode, thereby impeding inference. In extreme value analysis, the quality of estimation is all the more degraded as it is not carried out on the entire sample, but on some relatively small number of largest observations that are considered representative of the tail of the distribution. Whereas a plethora of techniques has been developed in the field of statistical learning to overcome this issue (Friedman et al., 2001), multivariate extremes in dimension larger that 2 are still handled with difficulty. It is the main purpose of the present chapter to address this issue, by developing a

non-parametric technique for identifying groups of variables exhibiting asymptotic dependence. Beyond a possible overall description of the tail dependence structure, when these classes are of small dimension, our method would enable further and more efficient assessment of multivariate tails. It combines novel statistical learning algorithms with multivariate extreme value theory (MEVT). From a practical perspective, it should be also pointed out that it includes a heuristic criterion to help select the sub-sample of extreme observations on which inference should be performed.

From a theoretical perspective, non-parametric assessment of multivariate extreme dependencies is already well documented. It relies on the necessary but quite mild assumption that there exists a tail dependence distribution, or equivalently that once marginal distributions have been transformed into standard Pareto, the cumulative distribution function of the resulting random vector is multivariate regularly varying. Then, its limit measure characterizes the extreme behavior of the original variables and possesses useful properties that facilitate its investigation. In particular, when switching to a pseudo-polar representation of the data, it can be expressed as a tensor product of two measures, one related to the radius, the other to the angles. The limit measure of the angles, termed spectral or angular measure, exhaustively embodies extreme dependencies. In the bivariate setting, the classical estimators introduced in the literature of this angular measure may vary depending on how marginals are standardized, which radius norm is picked and how measures are assessed. For instance, Einmahl et al. (2001); Resnick (2007); Beirlant et al. (2004) use the rank transform for standardization, then alternatively use the L_1, L_2 and L_1 norms and ground estimation on the basic empirical measure. However, the spectral measure is required to be Lebesgue-dominated, thereby failing to encompass situations where it is degenerate on some points. Breaking the barrier, Einmahl and Segers (2009) introduced a maximum empirical likelihood statistic, while extending theoretical results to the full set of L_p-norms, p • 1. Bayesian models have also flourished (Boldi and Davison, 2007; Guillotte et al., 2011). In the same vein, Sabourin and Naveau (2012) recently proposed a novel algorithm that handles moderate dimensions. Though it can be viewed as a subsequent improvement in multivariate extremes analysis, their technique is only efficient when all variables considered are asymptotically dependent; higher-complexity spectral measures may unfortunately not be studied by their method. Hence the need to first identify groups of dependent variables in regard to their extreme behavior: once this preliminary analysis carried out, the aforementioned estimators would enable more precise estimation up to dimension 5.

Lately, Haug et al. (2010) have ingeniously adapted one of the most celebrated dimension reduction method, namely Principal Components Analysis (PCA), to multivariate extremes analysis. Under an elliptical copula assumption, they recover the set of straight lines summarizing best the extreme covariance function, thereby leading to a clustering of variables based on extreme dependence. Following in their footsteps, we propose to borrow concepts from statistical learning to achieve dimen-

sion reduction, without making any parametric assumption in contrast. We rather base our analysis on a mixture model of the spectral measure exploiting its specific geometry, tackled from a latent variable point of view, under which useful properties arise and enable identification and interpretation of hopefully small groups of asymptotically dependent variables. Inference mimics classical non-parametric spectral measure estimation and focuses on the cloud of observation angles related to the L_2-norm. Since they belong to the positive orthant of the unit hypersphere, also called simplex, their structure is explored through a recent algorithm fitting PCA to Riemannian manifolds (Jung et al., 2012). Not only does this procedure respect the intrinsic combinatorial geometry of the simplex, but it also enriches the set of eligible summarizing sub-manifolds compared to standard PCA. Identification of the variables exhibiting asymptotic dependence is subsequently achieved using an appropriate clustering technique (Dhillon et al., 2002) on the obtained sub-space. We also provide a heuristic to help select the number of upper values most representative of extremes, thereby circumventing a traditionally intricate issue in MEVT.

To illustrate the assets and liabilities of our method, we perform numerical experiments and conduct a real case study for long-term dietary risk assessment. Extreme value theory (EVT) has already proven useful in studying high exposures to a single toxicant (Tressou et al., 2004a; Paulo et al., 2006), but to our knowledge the question of simultaneous extreme exposures to multiple chemical elements has never been addressed from a statistical point of view.

The chapter is organized as follows: we start off in Section 3.2 with the introduction of a few notations and hypotheses, subsequently used throughout the methodological part of our work in Section 3.3. There, after recalling a few basic notions in spectral measure analysis, we introduce a mixture model for the spectral probability measure and emphasize the ensuing fruitful properties it enjoys, when viewed as a latent variable model. Then we turn to the practical aspects of the approach we promote, and thoroughly depict our strategy for statistical inference under the assumed model, based on dimension reduction algorithms, in Section 3.4. It is supported by numerical experiments carried through in Section 3.5, and subsequently applied for illustration purposes to dietary risk assessment in Section 3.6. In view of both simulation and case study results, assets, liabilities, natural extensions and required improvements of our method are finally listed and discussed in Section 3.7.

3.2 hypotheses and notations

We start by introducing a few essential notations used throughout the chapter, followed by a short listing of the main hypotheses involved in the subsequent analysis.

3.2.1 Notations

Ú Vectors We shall deal with both random and non-random mathematical objects. To distinguish between them, we use uppercase to refer to random variables and lowercase otherwise. In both situations, vectors are denoted in bold and are written in columns, so that for any vector x with d ° 1 components,

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \text{ and its transpose is denoted by } \mathbf{x}^1 = (x_1; \ldots; x_d).$$

In particular, we write $\mathbf{0} = (0; \ldots; 0)^1$ to mean the null vector in $\mathbb{R}^d$, and $\mathbf{e}_1; \ldots; \mathbf{e}_d$ the vectors of the canonical basis of $\mathbb{R}^d$. Operations between vectors should be interpreted matricially, e.g. $\mathbf{x}^1\mathbf{x} = \sum_{j=1}^{d} x_j^2$.

Ú Sets As usual, braces refer to sets, and we denote by #I the cardinal of any set I. Its power set is denoted by P(I), its complementary by I^c and we further define $\mathcal{P}^*(I) := \mathcal{P}(I) \setminus \{\emptyset\}$. When I is included in a topological space, $\mathring{I}$ denotes the reunion of all open subsets of I and $\bar{I}$ its closure (the smallest closed set containing I).

Ú Norms When working on $\mathbb{R}^d$, recall that all norms are equivalent. For any collection of norms $\|.\|_{(1)}; \|.\|_{(2)}; \ldots$ in $\mathbb{R}^d$, we denote their corresponding unit spheres by $S^{d-1}_{(1)}; S^{d-1}_{(2)}; \ldots$ respectively, thereby emphasizing the topological dimension of these objects.

Ú Random variables For any sample $\mathbf{Z}_1; \ldots; \mathbf{Z}_n$ of n ° 1 independent and identically distributed (iid) random vectors on a space product of d ordered vector spaces $E_1 \times \cdots \times E_d$ with multivariate cumulative distribution function (cdf) F, dimensions are indexed by j P t1; ...; du and observations by i P t1; ...; nu. Order statistics are denoted $Z_{(1;j)} \leq \cdots \leq Z_{(n;j)}$, for all j P t1; ...; du. This notion is intrinsically linked to that of ranks; we define the rank function

$$\text{Rank} : \begin{cases} E_j \to \mathbb{N}^* \\ Z_{i;j} \mapsto \sum_{m=1}^{n} \mathbb{I}\{Z_{m;j} \leq Z_{i;j}\} \end{cases};$$

where for any condition A, $\mathbb{I}\{A\} = 1$ if A is true and 0 otherwise. Then we have $\text{Rank}(Z_{(i;j)}) = i$.

3.2.2 General setting and main hypotheses

Throughout this article, we consider a d-dimensional random vector

$$\mathbf{X} := (X_1; \ldots; X_d)^1;$$

$d \geq 2$, with Lebesgue-dominated probability distribution P on the positive orthant $C^d := [0,1]^d$ and cumulative distribution function (cdf) F, the tail structure of which we wish to assess. For $1 \leq j \leq d$, we denote by P_j the j-th 1-dimensional marginal distribution of P, i.e. the probability distribution of X_j, with corresponding continuous cdf $F_j(x) := P_j([0,x])$, $x \geq 0$. Statistical inference on the extreme behavior of F will be based on the observation of a sample $X_1, \ldots, X_n$, $n \geq 1$ (we shall write $X_i = (X_{i,1}, \ldots, X_{i,d})'$ for $1 \leq i \leq n$), supposedly drawn independently from P. We do not assume that F is characterized by its marginals, as would be the case in a situation where the X_j's are independent, or when considering copulas for modeling the dependence structure. Additionally, neither F nor any F_j, $1 \leq j \leq d$, are supposed to belong to the maximum domain of attraction of an extreme value distribution. The only imposed regularity constraint, apart from continuity of marginal cdfs, is the existence of a Radon measure Φ, not identically zero and not degenerate at a point, concentrated on the blunt convex cone $C^d_\Phi := [0,\infty]^d \setminus \{0\}$ such that

$$t\, P\left[\frac{1}{t}\,\frac{1}{(1-F_1(X_1))}, \ldots, \frac{1}{(1-F_d(X_d))}\right] \xrightarrow[t\to\infty]{v} \Phi(.). \tag{3.1}$$

Here, the notation "$\xrightarrow{v}$" stands for the vague convergence of measures in C^d_Φ: for all continuous functions $f : C^d_\Phi \to \mathbb{R}_+$ with compact support,

$$t\, E\left[f\left(\frac{1}{t}\,\frac{1}{(1-F_1(X_1))}, \ldots, \frac{1}{(1-F_d(X_d))}\right)\right] \xrightarrow[t\to\infty]{} \int_{C^d_\Phi} f\, d\Phi.$$

In words, Equation (3.1) simply states that there is, indeed, an extreme dependence structure between the random variables $X_1, \ldots, X_d$, exhaustively described by measure Φ, see for instance Section 8.2.3 in Beirlant et al. (2004) or Section 6.5.6 in Resnick (2007). Alternatively, consider the random vector $Z := (Z_1, \ldots, Z_d)'$ of standardized components

$$Z_j := \frac{1}{(1-F_j(X_j))},\ j \in \{1, \ldots, d\}, \tag{3.2}$$

and the corresponding transformed sample $Z_1, \ldots, Z_n$. Written this way, all Z_j with $j \in \{1, \ldots, d\}$, are standard Pareto distributed, i.e. $\forall x \geq 1$, $P[Z_j > x] = x^{-1}$. Let $R_{-\alpha}$ refer to the set of regularly varying functions with index $-\alpha$, then by definition $P[Z_j > x] \in R_{-1}$, with tail quantile function $\inf\{x \in \mathbb{R}_+ : P[Z_j \leq x] \geq 1 - 1/u\} = t$. For more details on univariate regular variation, we refer to Chapter 2 in Resnick (2007). Then, Equation (3.1) defines the distribution of Z as regularly varying in the multivariate sense. A larger overview of multivariate variation can be found for instance in Resnick (2007, Chapters 3 and 6).

Because we are interested in summarizing the dependence structure of the random variables $X_1, \ldots, X_d$, focus is here on the analysis of measure Φ in Equation (3.1). Theoretical properties of this mathematical object are detailed in Section 3.3, as a preliminary to the proposed inference technique, subsequently introduced in Section 3.4.

3.3 a mixture model under multivariate regular variation

In the present section, after reviewing a few known properties of the limit measure Λ in Equation (3.1) and defining spectral (probability) measures, we introduce a useful mixture model of the latter on which inference is next based. For simplicity, we shall work exclusively with the vector Z of standardized random variables defined in the previous section.

3.3.1 Exponent and spectral (probability) measures

Under the regular variation hypotheses listed herein-before, the limit measure Λ in Equation (3.1), called exponent measure, exhibits some convenient properties that we shall exploit later on. Specifically, it is homogeneous, i.e.

$$\begin{aligned} &\text{for all } 0 < s < 1 \text{ and Borel subset } B \text{ of } C_{\infty}^{d}; \\ &\Lambda(sB) = s^{-1}\Lambda(B); \end{aligned} \tag{3.3}$$

and fulfills d marginal constraints expressing the nature of the marginal survival functions, namely

$$\begin{aligned} &\text{for all } j = 1,\ldots,d \text{ and } 0 < z < 1; \\ &\Lambda\left\{x \in C_{\infty}^{d} : x_j > z\right\} = z^{-1}; \end{aligned} \tag{3.4}$$

see for instance Section 8.2.2 in Beirlant et al. (2004) and Section 6.1.4 in Resnick (2007). Consequently, Λ can be expressed as a tensor product of two measures when switching to pseudo-polar coordinates. Indeed, choose two norms $\|.\|_{(1)}$ and $\|.\|_{(2)}$ on $\mathbb{R}^d$ and define the following mapping:

$$T : \begin{array}{ccc} C_{\infty}^{d} & \longrightarrow & (0,1] \times S_{(2)}^{d-1} \\ x & \longmapsto & (r, w) = \left(\|x\|_{(1)}, x/\|x\|_{(2)}\right) \end{array};$$

with $T^{-1}(r, w) = rw = rw/\|w\|_{(1)} = x$. Typical choices of norms include the L_p-norm or the sup-norm L_∞. Then, the homogeneity property stated in Equation (3.3) implies that

$$\Lambda \circ T^{-1} := \Lambda_{-1} \otimes S, \tag{3.5}$$

where the radius measure Λ_{-1}, defined on $(0,1]$, is such that for all $x > 0$, we have $\Lambda_{-1}((x,1]) = x^{-1}$, and the angle measure S, referred to as the spectral measure, has support on $\Omega^{d-1} := S_{(2)}^{d-1} \cap C_{\infty}^{d}$ and satisfies

$$S(B) = \Lambda\left\{tx \in C_{\infty}^{d} : \|x\|_{(1)} > 1,\ x/\|x\|_{(2)} \in B\right\} \tag{3.6}$$

for all Borel subsets B of $\mathbb{R}^{d-1}$. A simple normalization of S yields the so-termed spectral probability measure Q on $\mathbb{R}^{d-1}$,

$$Q := S/S(\mathbb{R}^{d-1}). \tag{3.7}$$

Let us set $\omega = Z/\|Z\|_{(2)}$ and $\Theta = \|Z\|_{(1)}$, then Equations (3.1), (3.6) and (3.7) imply

$$t\,P(\omega \in P.;\ \Theta > t) \sim \frac{v}{t^{\alpha}}\, S(.); \tag{3.8}$$

$$P\left(\omega \in P.\ \middle|\ \Theta > t\right) \sim \frac{D}{t^{\alpha}}\, Q(.). \tag{3.9}$$

In words, the latter expression stipulates that Q is the limit distribution of the angles when the radius gets infinitely large. It thereby encapsulates the extreme (or asymptotic) dependence structure between the d variables in dimension $d > 1$. Observe that Equation (3.4) can be expressed in terms of moment constraints for S and Q respectively. Namely, for all $j \in \{1,\dots,d\}$, we have:

$$\int_{a\mathbb{R}^{d-1}} \frac{\omega_j}{\|\omega\|_{(1)}}\, S(d\omega) = 1; \tag{3.10}$$

$$\int_{\mathbb{R}^{d-1}} \frac{\omega_j}{\|\omega\|_{(1)}}\, Q(d\omega) = 1/S(\mathbb{R}^{d-1}). \tag{3.11}$$

3.3.2 Mixture model of the spectral probability measure

The extreme dependence structure between the variables $X_1,\dots,X_d$ can be expressed in terms of the geometry of the support of Q (or S), which we denote by supp(Q). Indeed, recall that supp(Q) is included in $\mathbb{R}^{d-1}$, the positive orthant of the unit hypersphere $S^{d-1}_{(2)}$, or the simplex associated with $\|.\|_{(2)}$. The latter can be partitioned into $2^d - 1$ non-empty and disjoint open faces with dimensions ranging from 0 up to $d-1$. They are identified by the collections of indexes $I_1,\dots,I_{2^d-1}$ forming $\mathcal{P}(\{1,\dots,d\})$: for any $h \in \{1,\dots,2^d-1\}$ such that $I_h \in \mathcal{P}(\{1,\dots,d\})$, the open face generated by $\{e_j, j \in I_h\}$ is

$$\mathbb{R}^{d-1}_h := \mathbb{R}^{d-1}(I_h) := \mathbb{R}^{d-1} \cap \mathrm{Vect}(\{e_j, j \in I_h\});$$

with dimension $m_h - 1$, where $m_h := \#I_h$. The star of vertex $\{e_j\}$, i.e. the reunion of all open faces the closure of which contains $\{e_j\}$, is denoted by $\mathrm{star}(\{e_j\})$ with corresponding set of indexes $S(j) := \{h \in \{1,\dots,2^d-1\} : \{e_j\} \subset \overline{\mathbb{R}^{d-1}_h}\}$. Notice that we also have $S(j) = \{h \in \{1,\dots,2^d-1\} : j \in I_h\}$. See Figure 3.1 for an illustration in dimension 3. By extension, we denote by $I_0 := \{0\}$ the set referring to the empty face $\mathbb{R}^{d-1}_0 := \emptyset$, with $m_0 = 0$.

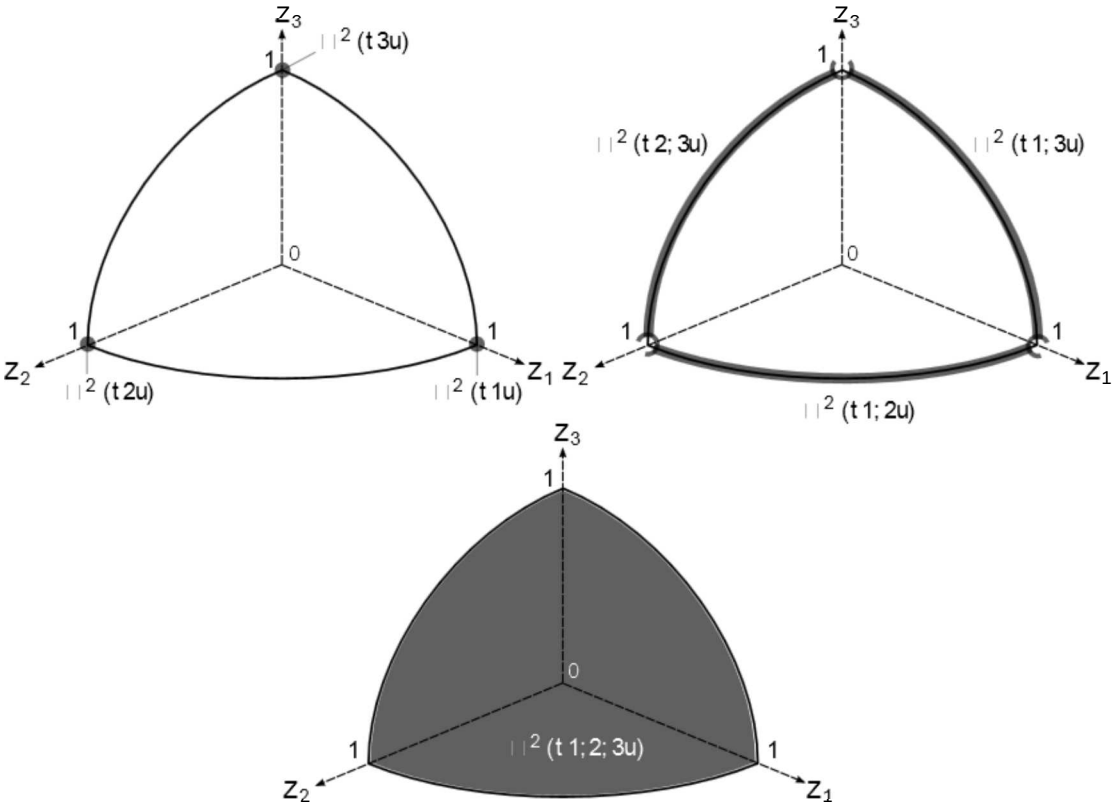


Figure 3.1 – The 7 nonempty open faces in the L_2-norm simplex Δ^2: the 3 vertices (left), the 3 edges (right), and the interior (bottom). The set $star(te_1u)$ corresponds to the reunion of the 4 open faces $\Delta^2(t1u)$, $\Delta^2(t1;2u)$, $\Delta^2(t1;3u)$ and $\Delta^2(t1;2;3u)$.

Given this decomposition, for any $h \in \{1;\ldots;2^d - 1\}$, $supp(Q) \cap \Delta_h^{d-1} \in H$ means that all X_j such that $j \in I_h$ exhibit asymptotic dependence (see Beirlant et al., 2004, Section 8.2.3). Consequently, recovering the set of faces intersecting the support of Q suffices to identify the sets of variables which are dependent in the extremes and those which are not. This motivates the following mixture model:

$$Q(.) = \sum_{h=1}^{2^d-1} \omega_h \, Q_h(.); \tag{3.12}$$

where for all $h \in \{1;\ldots;2^d - 1\}$,

$$Q_h(.) := \begin{cases} Q(. \cap \Delta_h^{d-1}) / Q(\Delta_h^{d-1}) & \text{if } Q(\Delta_h^{d-1}) > 0; \\ 0 & \text{otherwise,} \end{cases}$$

$$\omega_h := Q(\Delta_h^{d-1}).$$

In addition, denote by $\mathcal{H}$ the set made of all open faces intersecting $supp(Q)$, i.e.

$$\mathcal{H} := \{h \in \{1;\ldots;2^d - 1\} : supp(Q) \cap \Delta_h^{d-1} \in H\};$$

then for all $h \in \mathcal{H}$, Q_h is by definition a probability distribution on Δ_h^{d-1}. Obviously, $\omega_h \in [0;1]$, $\sum_{h=1}^{2^d-1} \omega_h = \sum_{h \in \mathcal{H}} \omega_h = 1$, and by extension $\omega_0 = 0$.

Now that the theoretical framework has been set out, following in the footsteps of standard mixture model analysis (see for instance McLachlan and Peel, 2000), we shall exploit the properties deriving from Equation (3.12) when reasoning in terms of latent variables and intrinsic clustering, in order to identify all Ω_h^{d-1} with $h \in H$.

3.3.3 Latent variables representation

We consider now the iid copies $Z_1, \ldots, Z_n$ of the standardized random vector Z defined in Section 3.2. Going back to the model in Equation (3.12) and setting $H := \#H$, one would expect the empirical distribution of these observations to reflect the decomposition of Q as a mixture of H probability measures on $\bigcup_{h \in H} \Omega_h^{d-1}$. Hence, there should be an intrinsic (unknown) clustering of the data into H classes leading to an identification of these open faces. Formally, define n unobserved random vectors $\Gamma_i := (\Gamma_{i,0}, \Gamma_{i,1}, \ldots, \Gamma_{i,2^d-1})$ with standard multinomial distribution such that $P(\Gamma_{i,h} = 1) =: p_h \in [0,1]$, where $\sum_{h=0}^{2^d-1} p_h = \sum_{h=0}^{2^d-1} \Gamma_{i,h} = 1$. Consider each $\Gamma_{i,h}$ as an indicator of whether observation Z_i is drawn from a distribution with spectral probability measure Q_h or not, i.e. when $\Gamma_{i,h} = 1$, individual i may reach extreme values on the m_h variables identified by I_h alone. Conversely, for any other $h \in H$, the event $\Gamma_{i,h} = 1$ indicates that each of the d coordinates of Z_i should generally have small to moderate values. Mathematically, for all $h \in \{0, \ldots, 2^d - 1\}$, $i \in \{1, \ldots, n\}$, we define the latent vector Γ_i as satisfying

$$P\left(\Theta_i \in \Theta, R_i > t \mid \Gamma_{i,h} = 1\right) \xrightarrow[t \to \infty]{} \begin{cases} Q_h(\Theta) & \text{if } h \in H; \\ 0 & \text{otherwise;} \end{cases} \tag{3.13}$$

where Θ_i and R_i are respectively the angle and radius of individual i. In fact, for the sake of interpretation, as soon as all $\Gamma_{i,h}$, $h \in H$, are null, we set $\Gamma_i = (1, 0, \ldots, 0)$. In other words, we impose the following equivalence:

$$\forall h \in \{1, \ldots, 2^d - 1\}, \ (\Gamma_h = 0) \Leftrightarrow (p_h = 0). \tag{3.14}$$

Some useful properties can be established in such a setting. We display here two results which are subsequently exploited for inference, as shall be seen in the next section. Proofs and technical details are deferred to Section 3.8. The proposition below exhibits the asymptotic behavior of conditional marginals under the latent variable model depicted herein-before.

Proposition 3.1 – Intra-class regular variation. We place ourselves in the framework of Section 3.3 and denote by $H(j)$ the set $\{h \in \mathcal{H} : j \in I_h\}$, i.e. the intersection between $\mathcal{H}$ and $S(j)$.
Then, for all $j \in \{1, \dots, d\}$, $h \in \{0, \dots, 2^d - 1\}$, $x \geq 1$,

$$\mathbb{P}\left(Z_{i,j} > x \,\middle|\, \Theta_{i,h} = 1\right) \underset{t \to \infty}{\sim} c_{j,h}\, x^{-1}, \tag{3.15}$$

where $c_{j,h} \in [0, 1/p_h]$ is non-null if and only if $h \in H(j)$, and $\sum_{h=0}^{2^d-1} p_h\, c_{j,h} = 1$.

In words, if $\Box_h^{d-1}$ is an open face where Q is null or if its closure does not contain vertex $\{e_j\}$, then it cannot project any mass on the j-th dimension. On the contrary, as soon as $h \in H(j)$, the marginal distribution of $Z_{i,j}$ given that it was drawn from a distribution with spectral probability measure Q_h is tail equivalent with the non-conditional distribution of $Z_{i,j}$. Hence, nonempty open faces intersecting with $\mathrm{supp}(Q)$ are identifiable by remaining only in the univariate level. In particular, the following result reveals that one can build a discriminative function of conditional marginal distributions the asymptotic behavior of which enables the characterization of $\{H(j);\ 1 \leq j \leq d\}$, and by extension of $\{I_h;\ h \in \mathcal{H}\}$.

Proposition 3.2 – Face-characterizing functional. We place ourselves in the framework of Section 3.3 and consider $H(j)$ as in Proposition 3.1. For all j in $\{1, \dots, d\}$, $h \in \{1, \dots, 2^d - 1\}$, $x \geq 1$, define the functional

$$\chi_{j,h}(t) := \int_1^{a_1} t\, \mathbb{P}\left(\Theta_i \geq t\right) \mathbb{P}\left(Z_{i,j} > x t \,\middle|\, \Theta_i \geq t,\ \Theta_{i,h} = 1\right) dx,$$

and assume that there exist some constants $\tau \in (0, 1)$, $c_\tau > 0$ and $t_\tau > 1$, such that for all $j \in \{1, \dots, d\}$, $h \in H(j)$,

$$\forall x > 1, \quad \forall t > t_\tau, \quad \left| \frac{\mathbb{P}\left(Z_{i,j} > x t,\ \Theta_{i,h} = 1\right)}{\mathbb{P}\left(Z_{i,j} > t,\ \Theta_{i,h} = 1\right)} \leq c_\tau x^{-1/\tau} \right|.$$

Then

$$\chi_{j,h}(t) \underset{t \to \infty}{\sim} \begin{cases} +1 & \text{if } h \in H(j) \\ 0 & \text{if } h \notin H(j) \cup \{0\} \end{cases} \quad \text{and} \quad \forall t > t_\tau,\ \chi_{j,0}(t) < 1. \tag{3.16}$$

As a consequence, for fixed dimension $j \in \{1, \dots, d\}$, the set $H(j)$ consists of all indexes h such that $\chi_{j,h}(t)$ diverges as t tends to infinity, instead of converging towards a finite constant, possibly zero. Since for all $h \in \mathcal{H}$, $I_h = \{j : h \in H(j)\}$, we have

$$\Box_h^{d-1} := \Box^{d-1}(I_h) = \Box^{d-1}\left\{j : \chi_{j,h}(t) \underset{t \to \infty}{\to} +1\right\}, \quad h \in \mathcal{H}, \tag{3.17}$$

and H is the set of all $h \subset \{1,\dots,2^d - 1\}$ such that there exists at least one dimension $j \in \{1,\dots,d\}$ for which $\Omega_{j,h}(t) \to +\infty$ as $t \to \infty$.

Remark 3.3 The assumption in Proposition 3.2 simply requires that the extreme dependence structure is reached at a reasonably fast rate. It can be directly linked to the concept of hidden regular variation introduced in Heffernan and Resnick (2005); Resnick (2002, 2007, 2008); Das and Resnick (2011); Das et al. (2013). Roughly speaking, if the distribution of Z had hidden regular variation, there would be an angular measure on $\mathcal{S}_h^{d-1}$ when making the radius increase with some regularly varying function $b(t) = o(t)$ with index $1/\alpha > 1$ instead of t. In that case, our assumption guarantees that $1/\alpha > 1 - \beta$, which is a rational condition for hidden regular variation not to be mistaken for multivariate regular variation in practice.

The proposed approach to statistical inference is based on Proposition 3.2 and Equation (3.17), as explained at length in the next section. Numerical experiments illustrating the relevance of the method we promote here are subsequently presented in Section 3.5.

3.4 statistical inference

Relying on the probabilistic framework detailed in Section 3.3, we now review the various steps of the proposed methodology to assess the dependence structure governing the extreme values of $X_1, \dots, X_d$. In short, it combines techniques borrowed from multivariate extreme value theory with clustering algorithms. Its declared purpose is to try to circumvent the classical curse of dimensionality that gravely deteriorates estimator variances (Massart, 1989). In particular, under some sparsity-like hypothesis, there is real hope of improvement in the estimation of the spectral measure: if supp(Q) is condensed on small manifolds of $\mathcal{S}_h^{d-1}$, recovering its geometrical structure should be manageable and would ultimately enable inference in lower, well-identified dimensions. Though high dimension is a classical issue in multivariate statistics (Friedman et al., 2001; Donoho, 2000), it is even more pregnant with meaning in extreme value analysis, where statistical inference only relies on a small sub-sample of most extreme observations. This justifies our approach, detailed step by step in the next three subsections: after a brief review of standard preliminaries, we depict our clustering algorithm given that the number H of open faces intersecting with supp(Q) is known. Finally, we propose some heuristic tools to choose both the aforementioned H and the number of upper values on which to base statistical analysis appropriately.

3.4.1 Preliminaries

Just as in classical spectral measure assessment, we consider that for some high enough threshold t, asymptotic relations such as in Equations (3.8), (3.9), (3.15) and 3.8.2 are sufficiently well approached to enable estimation. We use $t = n/k$ in practice, where k represents a fixed number of upper radii, that has to be chosen carefully as shall be seen later on. Other essential elements of the aforementioned asymptotics are not known a priori and need to be estimated beforehand, namely the marginals $F_1, \ldots, F_d$, used to standardize the original variables in Equation (3.2). To avoid restrictive hypotheses, we privilege here a non-parametric procedure, usually referred to as the rank transform: for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, d\}$, set

$$\hat{F}_j(X_{i,j}) = \frac{1}{n} \mathrm{Rank}(X_{i,j}), \tag{3.18}$$

and pursue the analysis with $\hat{X}_{i,j} = 1/(1 - \hat{F}_j(X_{i,j}))$, $1 \leq i \leq n$, $1 \leq j \leq d$ (Einmahl et al., 2001; Beirlant et al., 2004; Resnick, 2007; Einmahl and Segers, 2009). The generic random vector associated with this sample is written $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_d)$. Angles and radii are subsequently denoted by $\hat{\omega}_i$ and $\hat{r}_i$ respectively. For geometrical reasons explained in the next subsection, we set $\|.\|_{(2)}$ as the L_2-norm. In addition, we use the L_1-norm for $\|.\|_{(1)}$ because of its natural adequacy with marginal analysis. Observe that whereas it is unimportant regarding the angle, selecting a specific norm for the radius can have major implications. This is due to the selection process of "tail observations", defined as those with radius larger than n/k; clearly, different norms are bound to produce different sub-samples (Einmahl and Segers, 2009). However, such issues go beyond the scope of our analysis and are not discussed further here.

3.4.2 Dimension reduction and clustering

Equipped with the objects and notations introduced herein-before, we now proceed with the analysis of the spectral measure. Assume for the moment that H is known, k is appropriate, and that our only task is to identify the set of open faces Ω_h^{d-1}, $h \in H$, or equivalently the corresponding collections of indexes I_h. For this, we propose to mimic a classical approach in statistical learning, namely Principal Components Analysis (PCA, Friedman et al., 2001). Actually, Haug et al. (2010) have already extended PCA to extreme dependence analysis, but they assume an elliptical copula to describe the dependence structure in extremes. Because we would like to avoid any parametric restriction, we propose to work on the angles instead of the raw data and carry out PCA directly on the simplex. Then, choosing the L_2-norm enables the use of algorithms that respect the intrinsic distance of the unit hypersphere $S_{(2)}^{d-1}$, now identified with the usual unit hypersphere S^{d-1} of $\mathbb{R}^d$. We shall refer to the geodesic distance introduced hereafter.

Definition 3.4 The geodesic distance between two points x and y of $\mathbb{R}^d$ located on $\mathbb{S}^{d-1}$ is written

$$d_G(x,y) := \arccos\left(x^{\top}y\right).$$

Among all algorithms that were proposed in the literature (Jung et al., 2011; Huckemann and Ziezold, 2006; Fletcher et al., 2004), we consider the most general one, namely the Principal Nested Spheres (PNS) technique developed by Jung et al. (2012). In short, it is an iterative procedure that projects the data on smaller and smaller hyperspheres ($\mathbb{S}^{d-2}$, $\mathbb{S}^{d-3}$, ...), until the unit circle $\mathbb{S}^1$ is reached. These so-called PNS enrich the set of summarizing sub-manifolds compared to classical PCA. In Figure 3.2, we provide an illustration of the better adequacy of PNS compared to PCA when studying spectral measures. Among all $d-1$ resulting PNS, we may pick one of reasonably small dimension, which can be considered as representative of the geometry of supp(Q). In practice, this is achieved with a simple rule of thumb argument: we iteratively calculate the marginal level of geodesic variance encapsulated in $\mathbb{S}^1, \mathbb{S}^2, \ldots$ as in Jung et al. (2012), and select the biggest PNS still providing a gain in variance that exceeds a given threshold (e.g. 10% of the total variance). As was pointed out by Jung et al. (2012), $\mathbb{S}^1$ and $\mathbb{S}^2$ usually capture a high enough level of variance to ignore higher dimensions.

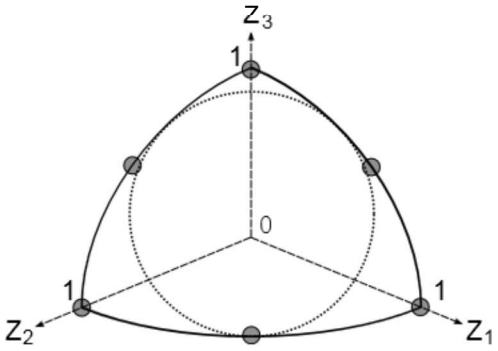


Figure 3.2 – A PNS on the L_2-norm simplex Δ^2: assume data points are uniformly concentrated around the greyed areas, then there is no straight line that can achieve a better separation than the represented dotted circle.

Once a sub-sphere has been selected, we still need to analyze the structure of the projected points to identify $\{\Delta_h^{d-1}; h \in H\}$. For this, we propose to use an accurate clustering procedure such as spherical k-means (Dhillon et al., 2002; Maitra and Ramler, 2010), based on the geodesic distance again. Though it may not seem necessary, the first stage of PNS makes the subsequent clustering more robust by getting rid of any misleading noise while best preserving the geometry of the original cloud of points. In fine, because of the structure of Δ^{d-1}, we can expect an adequacy between the obtained set of clusters and the underlying mixture model stated in Equation (3.12). Specifically, individuals reaching extremes on all $\{Z_j; j \in I_h\}$ should

be concentrated near $\binom{d}{h}^{-1}$ and projected onto a different zone of the PNS than observations taking high values on $\{Z_{j,1}, j \neq 1\} \cup \{h=1\}$, and be affected to different classes. Hence, identifying the set of open faces intersecting supp(Q) comes down to finding out on which dimensions individuals of each class reach extreme values. Technicalities about this last process are detailed in the next subsection. Obviously, many other natural techniques in mixture model analysis could have been adopted here (McLachlan and Peel, 2000); our preference for geometrical methods is based on a strong belief that Riemannian geometry is a key concept for understanding the structure of the spectral (probability) measure, as suggested by the encouraging results of the numerical experiments conducted in Section 3.5.

3.4.3 Identifying groups of asymptotically dependent variables

Let us begin by still assuming that H is known and k is well chosen. In view of the procedure explained above, we have at our disposal a clustering into H groups of the set of most extreme observations $\mathcal{A}_k := \{i \in \{1,\dots,n\} : \hat{\Gamma}_i \bullet \Theta \geq n/k\}$. These correspond to estimates of H coordinates of the unobserved vectors Γ_i introduced in Section 3.3.3, for all observations $i \in \mathcal{A}_k$. We denote them by $p_{i,\ell}$, $\ell \in \{1,\dots,H\}$, $i \in \mathcal{A}_k$, where $p_{i,\ell} = 1$ if i is in group ℓ, and a fixed ℓ corresponds to some unknown $h \in \{1,\dots,2^d - 1\}$. Observe that the case $h = 0$ is neglected, since Proposition 3.1 suggests that $P(\Gamma_i \bullet \Theta \geq n/k) \to 0$ for large enough n/k. Unfortunately, we are not able to comprehend to which open faces the events $p_{i,\ell} = 1$ are referring yet. In order to recover them, we propose to take advantage of the marginal properties stated in Proposition 3.2, and start by assessing $H(j)$ with the empirical counterpart of $\Gamma_{j,h}(t)$ in Equation (3.17). Formally, define

$$p_{j,\ell}(k) := \int_1^a \frac{1}{k}\frac{n_k}{n_\ell} \prod_{i \in \mathcal{A}_k} \mathbb{1}\left(Z_{i,j} > x\frac{n}{k}, p_{i,\ell} = 1\right) dx, \quad 1 \leq \ell \leq H, \ 1 \leq j \leq d,$$

where $n_k := \#\mathcal{A}_k$ is the number of observations the radius of which exceeds n/k, and $n_\ell := \sum_{i \in \mathcal{A}_k} \mathbb{1}\{p_{i,\ell} = 1\}$ the size of class ℓ. A more explicit version of this statistic and a short discussion regarding its accuracy are available in Section 3.8.3. Then, we set

$$\hat{H}(j) := \#\{\ell \in \{1,\dots,H\} : p_{j,\ell}(k) > 0\}.$$

In practice, to decide which $\ell \in \{1,\dots,H\}$ provide a large enough $p_{j,\ell}(k)$ to be selected, we perform a scree test-like analysis (Cattell, 1966). Specifically, we compute $p_{j,\ell}(k)$ on all $\ell \in \{1,\dots,H\}$, and $j \in \{1,\dots,d\}$. The $V := H \times d$ resulting values, denoted for instance by $p^1,\dots,p^V$, are subsequently ordered so that $p^{(1)} \geq \cdots \geq p^{(V)}$. In fine, we say $p^v > 0$, $1 \leq v \leq V$, if

$$V - v + 1 \leq \operatorname*{argmax}_{1 \leq w \leq V} p^{(V-w+1)} - p^{(V-w)}.$$

Because to each result index v corresponds a couple (j; `), we obtain in this way a collection of such couples from which all I(j), 1 § j § d, are determined. Then, identification of the corresponding open faces is straightforward: for all ` P t1; : : : ; Hu, a natural estimator of I` is

$$\hat{I}_\ell := \{j : \ell \in \hat{I}(j)\},$$

subsequently characterizing the desired spaces $\mathbb{S}^{d-1}(\hat{I}_\ell)$, 1 § ` § H.

Unfortunately, in practice H is usually unknown and k has to be picked at hand. To overcome this issue, we develop a heuristic criterion to measure the quality of a clustering, given a couple (k; I(k)), where I(k) is a number of clusters fixed a priori. Consider I(j) as before, but replace H with I(k) in its original definition, and set

$$\Delta(k, I(k)) = \sum_{j=1}^{d} \sum_{\ell=1}^{I(k)} (-1)^{1+\mathbb{1}_{I(k)}(j)} \hat{p}_{j,\ell}(k).$$

This statistic is simply built from Section 3.8.2: after having computed the empirical counterpart of $p_{j,\ell}(k)$ on all ` P t1; : : : ; I(k)u, we add up all quantities corresponding to ` P I(j) (which should be large) and substract the others (supposedly close to zero). When (k; I(k)) provides an accurate clustering of our data, (k; I(k)) should reach high values. To avoid possible practical errors, we further refine this criterion with some additional constraints. Specifically, classes should contain more that 1 individual, groups should each identify a different open face and no set I(j), 1 § j § d, should be empty. Observe that while the first two conditions are just common sense, the last one is necessary to respect the theoretical properties of Q: were there any empty H(j), marginal distributions could not be standard Pareto, and finding extra meaningless classes would come down to stating that the chosen threshold was not high enough to get rid of the empty face. Finally, we retain the partition inherited from $(k, I(k))^\star$, defined below.

$$(k, I(k))^\star = \underset{(k, I(k))}{\operatorname{argmax}} \Delta(k, I(k)) \times \prod_{\ell=1}^{I(k)} \mathbb{1}_{\{n_\ell > 1\}} \times \prod_{j=1}^{d} \mathbb{1}_{\{I(j) \neq H\}} \times \prod_{1 \leq \ell' < \ell \leq I(k)} \mathbb{1}_{\{\hat{I}_\ell \neq \hat{I}_{\ell'}\}}. \tag{3.19}$$

On account of the nice properties of the rank transform (Resnick, 2007; Heffernan and Resnick, 2005; Das and Resnick, 2011), we expect these statistical objects to converge to the true quantities they approximate as n — 1 . Unfortunately, due to the lack of probabilistic results on PNS and spherical k-means, which were originally introduced as geometrical techniques, we cannot provide here a thorough asymptotic analysis of the solution output by the statistical procedure described above. Further developments are the object of an ongoing work. Nonetheless, as shall be seen in the next section, numerical experiments provide strong empirical evidence of the efficiency of the approach we propose.

3.5 numerical experiments

We tested our method through a number of numerical experiments, for various values of n, d, and H. In doing so, we tried to handle various types of extreme dependence structures, to illustrate the impact of the complexity of supp(Q) on our algorithm. In the next two subsections, we first describe the different scenarios analyzed, then present and comment on the simulation results.

3.5.1 Settings

We generated n i.i.d. copies of a d-dimensional random vector $(X_1; \dots; X_d)^1$ with varying degrees of extreme dependence. Observations were drawn using a symmetric multivariate logistic model, with function `rmvevd` in `R` package `evd` (Stephenson, 2003). Asymptotic dependence was controlled via a parameter $r \in (0; 1]$, which indicates the strength thereof. In particular, $r = 1$ gives asymptotic independence, whereas asymptotic perfect dependence occurs when $r \to 0$. We repeated 100 trials of our algorithm under 3 scenarios, listed in Table 3.1.

Table 3.1 – List of scenarios considered in our numerical experiments; open faces intersecting with supp(Q) are filled in, with the corresponding extreme dependence coefficient r

Scenario 1 H = 2 d = 20		Scenario 2 H = 4 d = 20		Scenario 3 H = 4 d = 6	
Open faces	r	Open faces	r	Open faces	r
$\square^{d-1}(t1; 2; 3u)$	0.1	$\square^{d-1}(t1; 2u)$	0.1	$\square^{d-1}(t1; 2u)$	0.1
$\square^{d-1}(t4; \dots; 20u)$	0.1	$\square^{d-1}(t2; 3u)$	0.1	$\square^{d-1}(t2; 3u)$	0.1
		$\square^{d-1}(t4u)$	1	$\square^{d-1}(t4u)$	1
		$\square^{d-1}(t5; \dots; 20u)$	0.2	$\square^{d-1}(t5; 6u)$	0.2

To limit computation time, we tested its performance on 5 different sample sizes, namely $n = 5 \cdot 10^2; 10^3; 5 \cdot 10^3; 10^4$, and 10 thresholds $t = n/k$, with corresponding $k = n \cdot 0.001; n \cdot 0.002; \dots; n \cdot 0.01$. Not all possible number of classes were exploited either, because the constraint $\lvert n_i \rvert \geq 1u$ appearing in the definition of $(k; \mathcal{P}(k))^\square$ in Equation (3.19) restricts the maximum number of clusters to $n_k/2$. In practice, we iteratively compute our criterion for $n_i = 1; 2; \dots$ and stop as soon as the next 5 iterations cease improving it. For the same reasons, we disregarded situations where $d > 20$. However, in multivariate EVT, $d = 6$ and $d = 20$ can already be considered as high dimensions. Our code is based on the PNS algorithm provided by its authors at `http://www.stat.pitt.edu/sungkyu/MiscPage.html`, as well as the spherical k-means version of Dhillon et al. (2002), available in `R` package `skmeans` by setting `method = "pclust"`, and `start = "S"`.

3.5.2 Results

Results are displayed in Table 3.2, Table 3.3 and Table 3.4. The highlighted row reports the number of trials where we managed to exactly recover the set of open faces intersecting with the support of the spectral probability measure. To better understand the assets and liabilities of our algorithm, we also provide a detailed account of all inaccurate results, ordered relatively to their impact on the final interpretation.

Table 3.2 – Results of our numerical experiments in Scenario 1, repeated on 100 trials

		n			
	$\{\emptyset, \{1\}, \ldots, \{1, \ldots, k\}\}$	500	1000	5000	10000
Accurate sets	$\{1, 2, 3\}, \{4\}, \ldots, \{20\}$	90	96	100	100
Other inaccurate sets		10	4	0	0

Table 3.3 – Results of our numerical experiments in Scenario 2, repeated on 100 trials

		n			
	$\{\emptyset, \{1\}, \ldots, \{1, \ldots, k\}\}$	500	1000	5000	10000
Accurate sets	$\{1, 2\}, \{3\}, \{4\}, \{5\}, \ldots, \{20\}$	54	72	97	93
Extra sets with cardinal 1	$\{1, 2\}, \{3\}, \{4\}, \{5\}, \ldots, \{20\}, \{1\}$	2	3	0	0
	$\{1, 2\}, \{3\}, \{4\}, \{5\}, \ldots, \{20\}, \{3\}$	7	2	0	1
	$\{1, 2\}, \{3\}, \{4\}, \{5\}, \ldots, \{20\}, \{1\}, \{3\}$	1	0	0	0
Missing set $\{1, 2\}$ or $\{2, 3\}$	$\{1, 2\}, \{3\}, \{4\}, \{5\}, \ldots, \{20\}$	13	12	0	1
	$\{1\}, \{1, 2\}, \{3\}, \{4\}, \{5\}, \ldots, \{20\}$	1	0	0	0
	$\{1\}, \{2, 3\}, \{3\}, \{4\}, \{5\}, \ldots, \{20\}$	8	5	1	1
	$\{1\}, \{2, 3\}, \{4\}, \{5\}, \ldots, \{20\}$	0	0	0	0
Other inaccurate sets		14	6	2	4

Table 3.4 – Results of our numerical experiments in Scenario 3, repeated on 100 trials

		n			
	$\{\emptyset, \{1\}, \ldots, \{1, \ldots, k\}\}$	500	1000	5000	10000
Accurate sets	$\{1, 2\}, \{2, 3\}, \{4\}, \{5\}, \{6\}$	39	65	85	88
Extra sets with cardinal 1	$\{1, 2\}, \{2, 3\}, \{4\}, \{5\}, \{6\}, \{1\}$	0	1	0	0
	$\{1, 2\}, \{2, 3\}, \{4\}, \{5\}, \{6\}, \{3\}$	1	1	0	0
	$\{1, 2\}, \{2, 3\}, \{4\}, \{5\}, \{6\}, \{1\}, \{3\}$	1	0	0	0
Missing set $\{1, 2\}$ or $\{2, 3\}$	$\{1, 2\}, \{3\}, \{4\}, \{5\}, \{6\}$	16	11	6	2
	$\{1\}, \{1, 2\}, \{3\}, \{4\}, \{5\}, \{6\}$	0	0	0	0
	$\{1\}, \{2, 3\}, \{3\}, \{4\}, \{5\}, \{6\}$	23	14	5	6
	$\{1\}, \{2, 3\}, \{4\}, \{5\}, \{6\}$	0	0	0	1
Other inaccurate sets		20	8	4	3

As expected, in all scenarios, results improve when n increases, and success rates become particularly satisfactory as soon as n • 5000, for they then exceed 85% in all 3 scenarios. The best performance is obtained in scenario 1, where d = 20 and H = 2. Indeed, even with a very small sample (n = 500), only 10 trials out of 100 fail to recover the true decomposition of supp(Q), while in scenario 3, where d = 6 and H = 4, this rate never goes below 12% whatever n. This suggests that rather than the dimension, the complexity of supp(Q) may be one of the principal determinants of the performance of our procedure. Actually, given two spectral probability measures with equivalently complex supports, increasing dimensionality can produce better outcomes. This is the case with scenarios 2 and 3, where supp(Q) is contained on small subsets of 4 open faces, but d = 6 in the former while d = 20 in the latter. These results are not surprising and illustrate a typical phenomenon called the blessing of dimensionality (Donoho, 2000); as d increases, observations occur in relatively small subsets of the original space and are therefore easier to detect and separate. This property is the basis for common techniques in statistical learning, such as the widely celebrated Support Vector Machine (Friedman et al., 2001, Chapter 12), which projects the data onto some space with higher dimension in which they are well divided. In our numerical experiments, switching from Scenario 3 to Scenario 2 significantly reduces the risk of overriding either $\square^{d\square 1}(t1; 2u)$ or $\square^{d\square 1}(t2; 3u)$, which are very close to one another in the unit hypersphere and may be wrongfully confused during the PNS procedure. Observe nonetheless that these simulations were performed for very small values of parameter r, i.e. all dependencies were strong. Since we used the multivariate logistic model, this means that for all h P H, subsets supp(Q) X $\square_h^{d\square 1}$ did not cover the entire open faces $\square_h^{d\square 1}$ but were concentrated around small neighborhoods of one of their points. Had we considered less obvious extreme dependencies, these results would have probably been significantly degraded. This remark can be linked to the influence of the hidden spectral measure on inference (Resnick, 2002), for it controls the rate at which extreme structure is reached and thus dangerously impacts statistical analysis if the chosen threshold n=k is too small.

In fine, these results are quite encouraging, and underline the usefulness of methods from the field of statistical learning for multivariate EVT. Our next step will be to conduct a full theoretical analysis of our approach, which would demonstrate its rate of convergence in terms of sample size, strength of extreme dependencies, and complexity of the underlying spectral probability measure. This would for instance enable the construction of some confidence intervals and may help design a more efficient algorithm in terms of computation time. Indeed, at this early stage of development, our procedure is very long to implement; this forced us to limit the number of trials to 100 and prevented us from exploring the whole range of largest values on which our criterion may be calculated. Hence, results might have improved, had we been able to reach its true maximum.

3.6 application to dietary risk assessment

While eating is the privileged way of providing the necessary nutrients for the human organism, it also conveys toxic elements that, due to various environmental causes, contaminate the food. When consumed over certain tolerable doses, called dietary intake limits (DIL), these toxic elements can have a non-negligible impact on health. Similar phenomena also occur when diets are either too rich or too poor in nutrients. More importantly, further noxious effects may be caused by possible interactions between elements that are ingested simultaneously (Carpenter et al., 2002). For international institutes concerned about public health issues such as the WHO (World Health Organization), FAO (Food and Agriculture Organization), Unep (United Nations Environment Program), Efsa (European Food Safety Authority) or for national agencies such as the Anses (the French agency for food, environmental and occupational health safety), it is then of major interest to identify cocktails of food chemicals to which populations are indeed highly exposed. EVT has already proven useful to assess the probability of getting over a single dietary intake limit, in both univariate (Tressou et al., 2004a) and bivariate settings (Paulo et al., 2006). Here, we propose to apply the method detailed in Section 3.3 and Section 3.3.3 to examine the relationships between high simultaneous long-term exposure to 6 common nutrients and contaminants, namely iron (Fe), calcium (Ca), sodium (Na), methylmercury (MeHg), cadmium (Cd) and dioxins and dioxin-like polychlorinated biphenyls (PCB-DL). Their long-term toxicity is well-known, see for instance Anses (2011) and Carpenter et al. (2002). Methylmercury, cadmium, and PCB-DL are three contaminants found mainly in seafood products. While cadmium was recognized in 2004 as a type 2 carcinogen by the European Union, methylmercury and PCB-DL can attack the nervous system. Sodium, calcium and iron are three minerals principally found in animal products such as meat or dairy products. Long-term over-exposure to these nutrients is also harmful, e.g. consuming too much calcium can provoke urinary and renal calculi and excessive ingestion of sodium favors cardiovascular issues. As for iron, some studies have underlined a probable link between its excessive ingestion and Parkinson disease (Jenner et al., 1992). The current knowledge about possible synergistic effects between these chemicals, which may increase sanitary risks, is still quite poor, due to the complexity of these phenomena. Only methylmercury and PCB-DL have been studied jointly, and their simultaneous consumption was observed to amplify health issues in a number of experimental surveys (Bemis and Seegal, 1999; Carpenter et al., 2002). Henceforth, recovering groups of nutrients or contaminants to which the population is observed to be simultaneously over-exposed can help orient future biological and chemical research, which would in turn provide a better understanding of dietary risks. In terms of statistical analysis, thus reducing the dimension would also enable a more accurate estimation of the complex relationships between these types of exposure. Indeed, even though they are clearly linked

by the type of food (fish or meat) introduced in the diet, there are differences of composition between species — like tuna or salmon — that can imply independence between types of extreme long-term exposure. In particular, exceeding the DIL of more than 3 of these elements is an event never observed in the data. Because of the variety of individual dietary habits and the complexity of the contamination process, simultaneous types of high exposure are not an obvious phenomenon, are rarely observed, and need to be analyzed in detail. In the next paragraphs, after a brief presentation of the data, we apply the procedure introduced in the previous sections to the 6 aforementioned nutrients and contaminants.

3.6.1 Data and required assumptions

Our vectors of 6 types of exposure were calculated on the $n := 2488$ non-pregnant, non-lactating adults of the INCA2 database for which no important variable was missing, as described in Section 1.2. Levels of nutrients within each of the 1342 food items were given in the CIQUAL database and equivalents for contaminants were found in TDS2, both described in Section 1.2.2. In keeping with Chapter 2, the vectors of exposure $X_1; \dots; X_n$ were obtained by multiplying amounts of food with average contents then averaging over the number of reported days. Using similar notations and hypotheses as in the previous chapter, for all components $j \in \{1; \dots; d\}$ and consumers $i \in \{1; \dots; n\}$, we assume that

$$X_{i;j} = X^{\square}_{i;j} + U_{i;j};$$

where $X^{\square}_{i;j}$ is the long-term individual daily exposure to contaminant or nutrient j (also called usual intake) and $U_{i;j}$ is an independent noise with lighter tail than $X^{\square}_{i;j}$. We also suppose that the n vectors X_i are iid and that for any $i \neq i'$ and $j \neq j'$, $U_{i;j}$ is independent from $U_{i';j'}$. Under this setting, it is clear that the extreme dependence between the d components of vector X is determined solely by that between the d components of $X^{\square}$, the vector of interest. Relaxing these quite restrictive hypotheses would require further work; this is left for future research.

3.6.2 Analysis of extreme dependencies

Results of our method on the aforementioned sample are summarized in Figure 3.3. For various choices of threshold $t = n/k$ such that $k \in [10; n/3]$, we selected the partition that maximized our criterion $\Phi(k; \Pi(k))$ over all $(k; \Pi(k))$, and represented schematically the corresponding dependence structure. To get further confidence in this outcome, we summarize in Table 3.5 the strongest relationships that were found over all thresholds. The evolution of our criterion $\Phi(k; \Pi(k))$ with k is displayed on Figure 3.4. Its maximum is reached when $k = 564$, i.e. when calculations are based on the 1591 observations with largest radii.

In fact, the dependence structure represented in Figure 3.3 is found on all 16 largest values of $\Box(k; \hat{R}(k))$. The corresponding number of largest values k can be divided into two groups, one where k is in a neighborhood of 360, and another where k is around 560, as illustrated by the highlighted regions in Figure 3.4. Moreover, Table 3.5 shows that some dependencies are spotted whatever the number of largest values. In particular, methylmercury is almost always associated to PCB-DL, while cadmium and calcium get separated from all other chemicals. Concerning iron and sodium, uncertainty remains quite high, and a complementary bivariate analysis seems necessary to confirm the nature of their relationship. Figure 3.5 shows the estimated bivariate spectral probability measures of joint exposure first to MeHg and PCB-DL, then to Fe and Na. They were obtained using the maximum empirical likelihood (abbreviated MEL) approach of Einmahl and Segers (2009).

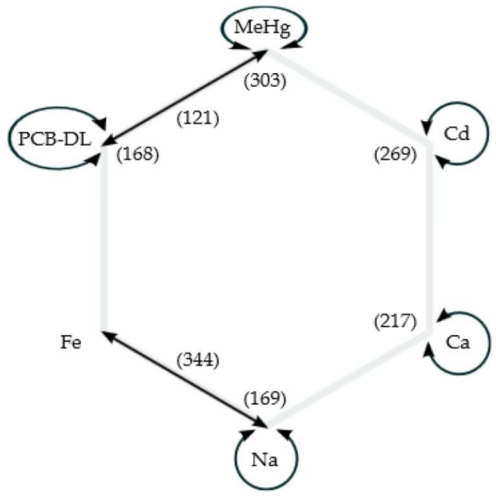


Figure 3.3 – Dependence structure between the 6 nutrients and contaminants of interest, on k = 564 that maximizes our criterion; arrows indicate extreme dependencies and the number of observations within each class is given in parentheses

Table 3.5 – Number of times extreme dependencies occur among all thresholds $t = n/k$, $k \in [10; n \cdot 0.3]$ (in %)

MeHg	MeHg & PCB-DL	PCB-DL
7.60	97.15	45.32

Cd	Ca
95.52	78.83

Fe & Na	Na
50.88	48.85

Clearly, the strong asymptotic dependence between methylmercury and PCB-DL is confirmed, on whatever value of k the estimation may be carried out. The presence of a sub-population reaching extreme exposure to PCB-DL alone is also suggested by the form of $\hat{Q}$, which gets close to vertical height on the extreme left part of the plot for many values of k. However, methylmercury does not exhibit such a behavior, and given that a specific class of independent exposure to MeHg only occurs for 7.60% of the largest values, we decide to disregard it. Actually, in terms of dietary habits, getting two clusters of individuals, one highly exposed to both MeHg and PCB-DL, and another solely to PCB-DL makes perfect sense. Indeed, contrary to PCB-DL, methylmercury is a contaminant found exclusively in seafood products. Hence, it is possible to get over-exposed to PCB-DL without ingesting high amounts of MeHg.

Now, let us turn to iron and sodium. According to the evolution of $\hat{Q}$ with k shown on Figure 3.5, if these two types of exposure exhibit asymptotic dependence, the latter is clearly weak. In fact, we are more inclined to believe in the presence of a mixture of three sub-populations, one ingesting high amounts of both Fe and Na, and the other two getting over-exposed to only one of these nutrients. It is also possible that k = 564 being quite high, the relationship appearing in Figure 3.3 corresponds not to extreme but moderately high levels of exposures. This inconclusive example suggests that extending our approach to the analysis of the hidden spectral measure (Resnick, 2002, 2008) would be of major interest.

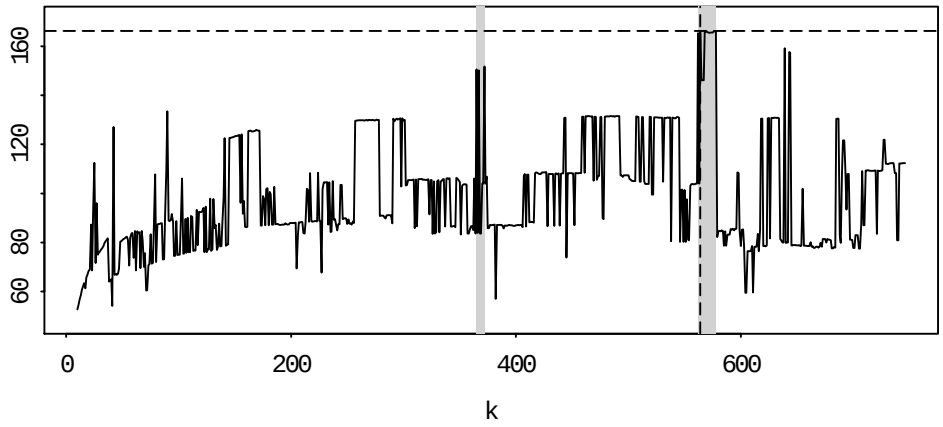


Figure 3.4 – Evolution in log-scale of $\Phi(k; \hat{H}^{\Phi}(k))$ with the number of largest values k, where $\hat{H}^{\Phi}(k)$ is the number of classes maximizing our criterion for some fixed k. The two dashed lines indicate the location of $\Phi(k; \hat{H}(k))$, while the grayed areas highlight regions where the 16 best criteria are obtained

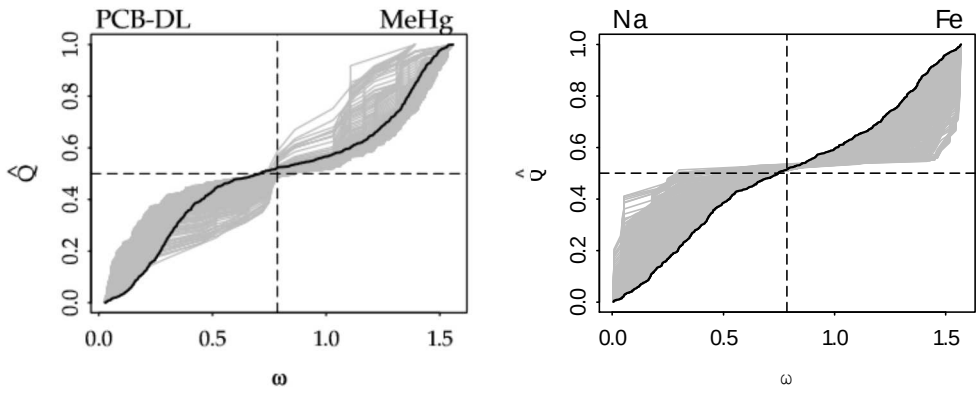


Figure 3.5 – MEL estimator of the bivariate spectral probability measure Q, obtained for various values of k (grey lines) up to k = 564 (black line), the optimal number of largest values selected by our criterion; the horizontal dashed line represents asymptotic independence, and the vertical one perfect asymptotic dependence

3.7 discussion

Non-parametric analysis of extreme dependencies via the spectral measure in high dimension d is still an open issue in multivariate extreme value theory. Though the bivariate setting has already been thoroughly investigated (Beirlant et al., 2004; Resnick, 2007; de Haan, 1985; Einmahl et al., 2001; Einmahl and Segers, 2009; Guillotte et al., 2011), and moderate dimensions are now accessible when all variables are asymptotically dependent (Sabourin and Naveau, 2012), the matter is still unresolved for $d \geq 5$. Following in the footsteps of Haug et al. (2010), who adapted the most celebrated Principal Components Analysis to extreme dependence assessment, we proposed a method combining multivariate extreme value theory with statistical learning and data mining standards so as to identify sub-groups of variables exhibiting asymptotic dependence. Once these clusters are identified, if they each encompass less than 5 variables, it then becomes possible to further estimate the corresponding sub-parts of the spectral measure with any existing method, for instance those that were cited herein-before.

We started in Section 3.3 by developing the theoretical context under which our approach was constructed. First of all, contrary to Haug et al. (2010), we did not make any parametric assumption on the extreme dependence structure. This led us to focus on the spectral measure itself, or more specifically its standardized version called spectral probability measure Q. After recalling that this can be viewed as the limit distribution of observation angles given that their radius is getting infinitely large, we underlined the adequacy between the geometry of its support on the positive orthant of the unit hypersphere and the nature of extreme dependencies. Indeed, if a group of variables, say Z_1 and Z_2, are asymptotically dependent, then Q will have positive mass on the open face generated by the corresponding dimensions, here $\square^{d-1}(t1; 2u)$. Therefore, we proposed a natural model of the spectral probability measure as a mixture of angular distributions with supports on each of the $2^d - 1$ non-empty open faces of the simplex. Tackled from a latent variable point of view, this model provided particularly useful properties, formulated in Proposition 3.1 and Proposition 3.2, that reduced an initially d-dimensional problem to d univariate ones. In particular, we showed that open faces intersecting the support of Q, namely $t \square \binom{d-1}{h}$, $h \in \mathcal{H}u$, could be identified by means of a simple functional $\square_{j;h}(t)$, $1 \leq j \leq d$, $1 \leq h \leq 2^d - 1$, introduced in Proposition 3.2.

Then, we moved to the practical part of our method in Section 3.4. Because we had pinpointed the major role of geometry for analyzing spectral measures in the preceding section, we adopted geometrical techniques suited for Riemannian objects such as the unit hypersphere for statistical inference. Borrowed from the statistical learning field, they consist in first projecting the initial cloud of points on a lower-dimensional space by means of the Principal Nested Spheres algorithm of Jung et al.

(2012), then clustering the obtained data with spherical k-means (Dhillon et al., 2002; Maitra and Ramler, 2010). By first implementing PNS, we reduced potential noise and enabled more efficient classification. Resulting clusters were then considered as representative of the open faces that intersect supp(Q), and analyzed as such. To recover to which Ω_h^{d-1}, $1 \leq h \leq 2^d - 1$, they were referring, we constructed estimators based on the empirical counterpart of the functional $\chi_{j;h}(t)$. The latter was also exploited to build a heuristic statistic that selects both the appropriate numbers of groups of dependent variables and of "extreme" observations. Unfortunately, due mainly to the absence of probabilistic analysis of PNS and spherical k-means in the literature, we were not able to provide asymptotic results about the aforementioned objects (this is the object of an ongoing work). Hence, assets and liabilities of our technique were discussed based solely on numerical experiments.

In Section 3.5, we tested our method on a set of simulated data bases. Three scenarios were considered, which try to encapsulate as many different situations as possible: they differed depending on d, on the number H of open faces containing mass, and on the complexity of supp(Q). In spite of a clearly improvable practical algorithm, the encouraging results we obtained enabled us to define which characteristics of Q have most influence on estimation. In particular, we saw that unlike H, d is of negligible importance to the complexity of supp(Q) and the strength of extreme dependence. The closer $t \in \Omega_h^{d-1}; h \in H\mathbb{u}$ are to one another (e.g. both $\Omega^{d-1}(t1; 2u)$ and $\Omega^{d-1}(t2; 3u)$ intersect supp(Q)), the harder it is to separate and correctly identify each of them. Estimation may also be impeded if H is large in comparison with n, or if asymptotic dependencies are weak. Indeed, to easily spot the desired open faces, the corresponding angular distributions denoted by Q_h should concentrate most of their mass on a small neighborhood of the middle point of Ω_h^{d-1}; the weaker dependencies are, the farther we are from this ideal situation. Though they were not considered in the simulations, we added some comments on rates of convergence to the asymptotic dependence structure that were sensed as a determining factor in assessment efficiency. Specifically, we insisted on the role that the hidden spectral measure may play when selecting an optimal number of largest values and suggested the interest of generalizing our approach to its analysis.

Further insight into our method was provided by a case-study illustration. Applied to real databases about exposures to 6 food contaminants, it produced stable outcomes, thereby giving confidence in the results. We were able to conclude that only two pairs of chemicals are actually linked in extremes, namely methylmercury and PCB-DL on the one hand, and iron and sodium on the other hand. These associations were confirmed by further computing the MEL estimator of Einmahl and Segers (2009) on the two pairs of variables. In addition, our method spotted a configuration usually hard to notice with traditional estimators, but quite natural given the underlying mixture model on which we based the analysis: it underlined the presence of a mixture of populations, some being jointly over-exposed to a couple of

elements, while others ingest high quantities of only one of them (PCB-DL or Na). In terms of public health implications, this means that people who are over-exposed to methylmercury tend to ingest simultaneously high amounts of dioxins and PCB. Knowing that these two toxicants have similar noxious effects on the human organism (Fischer et al., 2008; Weihe et al., 1996), and that when combined, synergistic effects can occur (Bemis and Seegal, 1999; Carpenter et al., 2002), this suggests paying particular attention to the populations that do not respect the corresponding DIL. It also justifies the need for specific research on potential combined effects of these two contaminants, which would help in assessing the sanitary risks brought upon the concerned population.

In view of these results, one advantage of our multivariate approach is that people in the data are dispatched into multiple classes that embody different types of extreme dependencies. In our case-study example, it facilitates the understanding of over-exposure categories by allowing classical discriminant analyses. An interesting alternative would be to model the various Γ_h appearing in the mixture model of the spectral probability measure in function of auxiliary covariates, e.g. some sociologic or economic variables here. More than providing easily interpretable results, this would probably increase the performance of our procedure by helping discriminate between the various clusters. Such generalizations of the present work will be the subject of further investigation in the near future.

3.8 proofs and supplements

3.8.1 Intra-class regular variation

We shall start the proof of Proposition 3.1 by exhibiting two preliminary results. The first one, given in the lemma below, states that Γ_h can be viewed as the limit probability that $\Gamma_{i;h}$ equals 1, $1 \leq i \leq n$, when the radius becomes infinitely large.

Lemma 3.5 Consider the same framework as in Proposition 3.1, then for all $h \in \{0, \ldots, 2^d - 1\}$, $i \in \{1, \ldots, n\}$,

$$P\left(\Gamma_{i;h} = 1 \,\middle|\, \Gamma_i \bullet t\right) \underset{t \to \infty}{\longrightarrow} \Gamma_h. \tag{3.20}$$

Proof First of all, extend Q to the whole sphere by setting $Q\left(S_{(2)}^{d-1} \cap z^{\perp}\right) = 0$, then consider the following neighborhoods of each of the $2^d - 1$ open faces of the simplex: for any $\varepsilon > 0$, $h \in \{1, \ldots, 2^d - 1\}$ and the geodesic distance $d_G(.;.)$ on $\mathcal{S}^{d-1}$, set

$$V_\varepsilon\left(\Gamma_h^{d-1}\right) := \left\{\omega \in S_{(2)}^{d-1} : \inf\left\{d_G(\omega; x);\, x \in \Gamma_h^{d-1}\right\} \leq \varepsilon\right\}.$$

We shall prove that for all $h \in \{1, \ldots, 2^d - 1\}$,

$$\lim_{t \to \infty} P\left(\left|\frac{1}{t} \sum_{i=1}^{t} V_\varepsilon(\square_h^{d-1})(X_i) - Q\left(V_\varepsilon(\square_h^{d-1})\right)\right| > \delta\right) = 0, \tag{3.21}$$

for an arbitrary small δ. This result can be obtained by applying the Portmanteau theorem to Equation (3.9), provided that we find at least a decreasing sequence of positive constants $\varepsilon_1, \varepsilon_2, \ldots$ that tends to 0 such that for any $m \geq 1$ and open face $\square_h^{d-1}$, the frontier of $V_{\varepsilon_m}(\square_h^{d-1})$ has null measure relative to Q. Since Q is a finite measure, its associated cdf admits at most countably many discontinuity sets, hence the requirement is met.

Now we shall prove that for all $h \in \{1, \ldots, 2^d - 1\}$,

$$\lim_{\varepsilon \to 0} Q\left(V_\varepsilon(\square_h^{d-1})\right) = Q\left(\overline{\square_h^{d-1}}\right). \tag{3.22}$$

Observe that $Q\left(V_\varepsilon(\square_h^{d-1})\right) = \int \mathbb{I}\left(x \in V_\varepsilon(\square_h^{d-1})\right) Q(dx)$; and that $V_\varepsilon(\square_h^{d-1})$ tends to $\overline{\square_h^{d-1}}$ as ε tends to 0. Therefore, Equation (3.22) can be deduced from the dominated convergence theorem, using $\mathbb{I}\left(x \in V_\varepsilon(\square_h^{d-1})\right) \leq 1$.

By combining Equation (3.21) and Equation (3.22), we obtain

$$\lim_{\varepsilon \to 0} \lim_{t \to \infty} P\left(\left|\frac{1}{t} \sum_{i=1}^{t} V_\varepsilon(\square_h^{d-1})(X_i) - Q\left(\overline{\square_h^{d-1}}\right)\right| > \delta\right) = 0, \tag{3.23}$$

for all $h \in \{1, \ldots, 2^d - 1\}$.

Now let D_j, $j \in \{0, \ldots, d-1\}$ denote the set of indexes $h \in \{1, \ldots, 2^d - 1\}$ that identify j-dimensional open faces. We shall prove Lemma 3.5 by strong induction. First, observe that for all $h \in D_0$ we have $\overline{\square_h^{d-1}} = \square_h^{d-1}$ and that the events $\{\square_{i,h} = 1;\ h \in \{0, \ldots, 2^d - 1\}\}$ are disjoint by construction. Hence, Equation (3.23) can be rewritten as follows:

$$\begin{aligned}
Q(\square_h^{d-1}) &= \lim_{\varepsilon \to 0} \lim_{t \to \infty} \prod_{\ell=0}^{2^d-1} P\left(\left|\frac{1}{t} \sum_{i=1}^{t} V_\varepsilon(\square_h^{d-1})(X_i) - Q\left(\overline{\square_h^{d-1}}\right)\right| > \delta;\ \square_{i,\ell} = 1\right)^{P(\square_{i,\ell} = 1)} \\
&= \lim_{\varepsilon \to 0} \lim_{t \to \infty} P\left(\left|\frac{1}{t} \sum_{i=1}^{t} V_\varepsilon(\square_h^{d-1})(X_i) - Q\left(\overline{\square_h^{d-1}}\right)\right| > \delta;\ \square_{i,h} = 1\right)^{P(\square_{i,h} = 1)} \\
&\quad + \prod_{\substack{\ell=0 \\ \ell \neq h}}^{2^d-1} P\left(\left|\frac{1}{t} \sum_{i=1}^{t} V_\varepsilon(\square_h^{d-1})(X_i) - Q\left(\overline{\square_h^{d-1}}\right)\right| > \delta;\ \square_{i,\ell} = 1\right)^{P(\square_{i,\ell} = 1)}.
\end{aligned}$$

Since Equation (3.15) ensures that $\lim_{t \to \infty} P\left(\left|\frac{1}{t} \sum_{i=1}^{t} V_\varepsilon(\square_h^{d-1})(X_i) - Q\left(\overline{\square_h^{d-1}}\right)\right| > \delta;\ \square_{i,h} = 1\right)^{1} = 1$ for all $\varepsilon > 0$ and that for all $\ell \neq h$, $\lim_{\varepsilon \to 0} \lim_{t \to \infty} P\left(\left|\frac{1}{t} \sum_{i=1}^{t} V_\varepsilon(\square_h^{d-1})(X_i) - Q\left(\overline{\square_h^{d-1}}\right)\right| > \delta;\ \square_{i,\ell} = 1\right)^{1} = 0$, we can conclude that

$$\lim_{\varepsilon \to 0} \lim_{t \to \infty} P(\square_{i,h} = 1)^{P(\square_{i,h} = 1)} = \lim_{t \to \infty} P(\square_{i,h} = 1)^{P(\square_{i,h} = 1)} = Q(\square_h^{d-1}).$$

Lemma 3.5 is thus true for all $h \in D_0$. Now fix some $J \in \{1, \ldots, d-2\}$ and assume that it holds for all $h \in \bigcup_{j=0}^{J} D_j$. Set

$$F(h) := \left\{ \ell \in \{1, \ldots, 2^d - 1\} : \ell^{d-1} \in \overline{\left\{ \frac{h}{\|h\|}, \ldots, \frac{h^{d-1}}{\|h\|} \right\}} \right\}.$$

Using the same arguments as before, for all $h \in D_{J+1}$ we have:

$$\begin{aligned} Q(\overline{h^{d-1}}) = \lim_{\varepsilon \to 0} \lim_{t \to \infty} & \mathbb{P}\left[\| PV_\varepsilon(h^{d-1}) \| > t, \theta_{i,h} = 1 \right] \mathbb{P}\left[\theta_{i,h} = 1 \right]^{-1} t^{-1} \\ & + \sum_{\ell \in F(h)} \mathbb{P}\left[\| PV_\varepsilon(h^{d-1}) \| > t, \theta_{i,\ell} = 1 \right] \mathbb{P}\left[\theta_{i,\ell} = 1 \right]^{-1} t^{-1} \\ & + \sum_{\substack{\ell \in R F(h) \\ \ell \neq h}} \mathbb{P}\left[\| PV_\varepsilon(h^{d-1}) \| > t, \theta_{i,\ell} = 1 \right] \mathbb{P}\left[\theta_{i,\ell} = 1 \right]^{-1} t^{-1}. \end{aligned}$$

Invoking again Equation (3.15), $\lim_{t \to \infty} \mathbb{P}\left[\| PV_\varepsilon(h^{d-1}) \| > t, \theta_{i,h} = 1 \right] = 1$ for all $\varepsilon > 0$ and for all $\ell \neq h$,

$$\lim_{\varepsilon \to 0} \lim_{t \to \infty} \mathbb{P}\left[\| PV_\varepsilon(h^{d-1}) \| > t, \theta_{i,\ell} = 1 \right] = \begin{cases} 1 & \text{if } \ell \in F(h); \\ 0 & \text{if } \ell \in RF(h). \end{cases}$$

Combined with the induction hypothesis, these equations entail

$$\lim_{t \to \infty} \mathbb{P}\left[\theta_{i,h} = 1 \right] t \, Q(\overline{h^{d-1}}) \prod_{\ell \in F(h)} \ell = \Theta_h.$$

This concludes the proof. □

The second preliminary result in the lemma below states that the distribution of vector Z_i given that $\theta_{i,h} = 1$ is multivariate regularly varying when $h \in H$.

Lemma 3.7 Consider the same framework as in Proposition 3.1, then for all $h \in H$, there is a Radon measure Θ_h, non identically zero and not degenerate at a point, concentrated on the blunt convex cone $C_h^d := \{x \in C_\Theta^d : \|x\|_{(2)} \in h^{d-1}\}$, such that

$$t \, \mathbb{P}\left[\frac{Z_i}{t} \in \cdot \right] \xrightarrow{v} \Theta_h(\cdot). \tag{3.24}$$

Proof By Lemma 3.5, Equation (3.8) and Equation (3.13), we have that

$$t \, \mathbb{P}\left[\| Y_i \| \cdot \Theta_i \in \cdot, \Theta_i \neq 0 \right] \mathbb{P}\left[\theta_{i,h} = 1 \right] \xrightarrow{v} S_h(\cdot) := \begin{cases} S(\cdot \cap h^{d-1}) / p_h & \text{if } h \in H; \\ 0 & \text{otherwise,} \end{cases} \tag{3.25}$$

where by definition,

$$S(\cdot \cap h^{d-1}) = \Big\{ x \in C_\Theta^d : \|x\|_{(1)} > 1, \tfrac{x}{\|x\|_{(2)}} \in \cdot \cap h^{d-1} \Big\}$$

$$= \square \; tx \; P \; C_{\square}^{d} : \}x\}_{(1)} \bullet 1; x=\}x\}_{(2)} \; P \; . \; uX \; tx \; P \; C_{\square}^{d} : x=\}x\}_{(2)} \; P \; \square_{h}^{d \square 1} u^{\square}.$$

Recall that $C_h^d := tx \; P \; C_{\square}^{d} : x=\}x\}_{(2)} \; P \; \square_{h}^{d \square 1} u$, and set

$$\square_h(.) := \begin{cases} \square(. \; X \; C_h^d)=p_h & \text{if } h \; P \; H \\ 0 & \text{otherwise,} \end{cases}$$

then we can rewrite S_h in function of $\square_h$ as below:

$$S_h(.) = \square_h(tx \; P \; C_{\square}^{d} : \}x\}_{(1)} \bullet 1; x=\}x\}_{(2)} \; P \; . \; u).$$

Since C_h^d is a cone, the homogeneity property of $\square$ stated in Equation (3.3) is passed on $\square_h$, $h \; P \; H$. Indeed, for all $0 \dagger s \dagger 1$ and Borel subset B of $C_{\square}^{d}$,

$$\square_h(sB) = \square^{\square}(sB) \; X \; (C_h^d)^{\square} =p_h = \square^{\square} s(B \; X \; C_h^d)^{\square} =p_h = s^{\square 1} \square^{\square} B \; X \; C_h^d)^{\square} =p_h = s^{\square 1} \square_h(B).$$

According to Theorem 6.1 in Resnick (2007), it naturally follows that for all $i \; P \; t1; : : : ; nu$,

$$t \; P \; \frac{Z_i}{t} \; P \; . \; \square_{\square}^{\square} \square_{i;h} = 1^{\square} \square_{t \to \infty}^{v} \square_h(.);$$

where, just like $\square$, $\square_h$ can be written as the product of a measure on the radius with a measure on the angles when switching to pseudo-polar coordinates:

$$\square_h \; \square T^{\square 1} = \square_{\square 1} \square \; S_h.$$

$\square$

We can now tackle the proof of Proposition 3.1, which is recalled below for convenience.

Proposition – Intra-class regular variation. We place ourselves in the framework of Section 3.3 and denote by $H(j)$ the set $th \; P \; H : j \; P \; I_h u$, i.e. the intersection between H and S(j).
Then, for all $j \; P \; t1; : : : ; du$, $h \; P \; t0; : : : ; 2^d \square 1u$, $x \bullet 1$,

$$t \; P \; \square^{\square} Z_{i;j} \circ x \; t \; \square_{\square}^{\square} \square_{i;h} = 1^{\square} \square_{t \to \infty}^{v} c_{j;h} \; x^{\square 1};$$

where $c_{j;h} \; P \; [0; 1=p_h]$ is non-null if and only if $h \; P \; H(j)$, and $\sum_{h=0}^{2^d \square 1} p_h \; c_{j;h} = 1$.

Proof Going back to the marginal level, multivariate regular variation of conditional distributions gives for all $x \bullet 1$, $1 \S i \S n$, $1 \S j \S d$,

$$t \; P \; \frac{Z_{i;j}}{t} \circ x \; \square_{\square}^{\square} \square_{i;h} = 1^{\square} \square_{t \to \infty}^{v} \square_h(tz \; P \; C_{\square}^{d} : z_j \circ xu).$$

Notice that we now have a null limit for all $h \; R \; H(j)$, i.e. the intersection between H and S(j), which identifies the star of vertex $te_j u$. Indeed, if $te_j u$ is not included in the

closure of $\square_h^{d-1}$, then by definition of Q_h, S_h and $\square_h$, after projection no mass is put on the j-th dimension. Furthermore, for all $h \in H(j)$, we have

$$\square_h(t z \in C_\square^d : z_j \leq xu) = \int_{\square^{d-1}} \mathbb{1}\left\{\frac{!_j}{\{!\}_{(1)}} \leq x\right\} \square_{\square 1}(d\square)\, S_h(d!) = x^{\square 1} \underbrace{\int_{\square^{d-1}} \frac{!_j}{\{!\}_{(1)}} S_h(d!)}_{c_{j;h}}.$$

Hence, for all $h \in \{0;\ldots;2^d-1\}$, $i \in \{1;\ldots;n\}$, $j \in \{1;\ldots;d\}$, $x \geq 1$, we can write

$$t \in \square Z_{i;j} \leq xt \square_{i;h} = 1 \square = \sum_{t=1} c_{j;h}\, x^{\square 1};$$

where $c_{j;h} \geq 0$ when $h \in H(j)$, and $c_{j;h} = 0$ otherwise. Based on the marginal constraints on S stated in Equation (3.10) and because $\{\square_h^{d-1}u\}_{0 \leq h \leq 2^d-1}$ forms a partition of $\square^{d-1}$, we have that $c_{j;h} \in [0;1/p_h]$ for all $h \in \{0;\ldots;2^d-1\}$ and $\sum_{h=0}^{2^d-1} p_h\, c_{j;h} = \sum_{h \in H(j)} p_h\, c_{j;h} = 1$. □

3.8.2 Face-characterizing functional

Before tackling its proof, Proposition 3.2 is recalled for convenience.

Proposition – Face-characterizing functional. We place ourselves in the framework of Section 3.3 and consider H(j) as in Proposition 3.1. For all j in $\{1;\ldots;d\}$, $h \in \{1;\ldots;2^d-1\}$, $x \geq 1$, define the functional

$$\square_{j;h}(t) := \int_1^1 t \in (\square_i \geq t) \in \square Z_{i;j} \leq xt \square_i \geq t; \square_{i;h} = 1 \square dx;$$

and assume that there exist some constants $\square \in (0;1)$, $c\square > 0$ and $t\square \geq 1$, such that for all $j \in \{1;\ldots;d\}$, $h \in H(j)$,

$$\forall x \geq 1, \quad (t \geq t\square) \Rightarrow \frac{\in \square Z_{i;j} \leq xt \square \square_{i;h} = 1 \square}{\in \square Z_{i;j} \leq t \square \square_{i;h} = 1 \square} \leq c\square x^{\square 1/\square}.$$

Then

$$\square_{j;h}(t) \sum_{t=1} \square \begin{cases} +1 & \text{if } h \in H(j) \\ 0 & \text{if } h \in H(j) \cap \{0\} \end{cases} \quad \text{and} \quad \forall t \geq t\square, \square_{j;0}(t) < 1.$$

Proof We shall handle the situations where $h \in H(j)$ and $h \in H(j)$ separately. To simplify notations, for all $h \in \{0;\ldots;2^d-1\}$ and $x \geq 0$ we will denote by $\bar{F}_{j;h}(x)$ the conditional probability that $Z_{i;j}$ exceeds x given $\square_{i;h}$ equals 1, for any $i \in \{1;\ldots;n\}$:

$$\bar{F}_{j;h}(x) := \in \square Z_{i;j} \leq x \square \square_{i;h} = 1 \square.$$

Ú h PH(j) : p_h □ 0 and □$_h$ □ 0

From Equation (3.15) in Proposition 3.1, it is straightforward that $\bar{F}_{j;h}$ is regularly varying with index □1, i.e. for any x • 1,

$$\frac{\bar{F}_{j;h}(xt)}{\bar{F}_{j;h}(t)} \underset{t\to\infty}{\to} x^{-1}.$$

Hence, $\bar{F}_{j;h}$ may be written as follows:

$$\bar{F}_{j;h}(x) = x^{-1} L_{j;h}(x);$$

where $L_{j;h}(x)$ is a slowly varying function ($L_{j;h}$ PR$_0$) that converges to $c_{j;h}$ as x — 1 .

Remark 3.11 Define $x^{\star}_{j;h} := \inf\{x \bullet 1 : \bar{F}_{j;h}(x) = 0\}$, the right endpoint of survival function $\bar{F}_{j;h}$ for any j P {1; : : : ; d} and any h P {0; : : : ; 2^d □ 1}. Then for all h PH(j), $x^{\star}_{j;h} = +\infty$, that is @t • 1, $\bar{F}_{j;h}(t) > 0$.

Since Bayes' formula gives

$$\int_1^{\infty} t\,P(\Omega_i \bullet t)\,P\left(Z_{i;j} > xt \,\middle|\, \Omega_i \bullet t;\ \Omega_{i;h} = 1\right) dx = \int_1^{\infty} \frac{t\,P\left(Z_{i;j} > xt;\ \Omega_i \bullet t \,\middle|\, \Omega_{i;h} = 1\right) p_h}{P\left(\Omega_{i;h} = 1 \,\middle|\, \Omega_i \bullet t\right)} dx;$$

and for all x • 1, $P\left(Z_{i;j} > xt;\ \Omega_i \bullet t \,\middle|\, \Omega_{i;h} = 1\right) = \bar{F}_{j;h}(xt)$, we have

$$\int_1^{\infty} t\,P(\Omega_i \bullet t)\,P\left(Z_{i;j} > xt \,\middle|\, \Omega_i \bullet t;\ \Omega_{i;h} = 1\right) dx = \frac{t\,\bar{F}_{j;h}(t)\,p_h}{P\left(\Omega_{i;h} = 1 \,\middle|\, \Omega_i \bullet t\right)} \int_1^{\infty} \frac{\bar{F}_{j;h}(xt)}{\bar{F}_{j;h}(t)} dx = \frac{L_{j;h}(t)\,p_h}{P\left(\Omega_{i;h} = 1 \,\middle|\, \Omega_i \bullet t\right)} \int_1^{\infty} x^{-1} \frac{L_{j;h}(xt)}{L_{j;h}(t)} dx.$$

Fix some □ > 0, small enough to verify $c_{j;h}$ □ □ > 0, and some $t_\Box$ > 0 such that @t • $t_\Box$, we have simultaneously $P\left(\Omega_{i;h} = 1 \,\middle|\, \Omega_i \bullet t\right) > p_h$ † □ (Lemma 3.5) and $L(t) > c_{j;h}$ † □. Obviously, as soon as t • $t_\Box$, we also have $L(xt) > c_{j;h}$ † □ for all x • 1, and

$$0 < \frac{c_{j;h} - \Box}{c_{j;h} + \Box} < \frac{L_{j;h}(xt)}{L_{j;h}(t)}.$$

Hence, @t • $t_\Box$,

$$\int_1^{\infty} t\,P(\Omega_i \bullet t)\,P\left(Z_{i;j} > xt \,\middle|\, \Omega_i \bullet t;\ \Omega_{i;h} = 1\right) dx > \frac{(c_{j;h} - \Box)^2\,p_h}{(p_h + \Box)(c_{j;h} + \Box)} \int_1^{\infty} x^{-1} dx = +\infty;$$

or equivalently,

$$\int_1^{\infty} t\,P(\Omega_i \bullet t)\,P\left(Z_{i;j} > xt \,\middle|\, \Omega_i \bullet t;\ \Omega_{i;h} = 1\right) dx \underset{t\to\infty}{\to} +\infty.$$

Ú h PHzH(j) : p_h □ 0 and □$_h$ □ 0

Contrary to the case where h P H(j), we no longer have $\bar{F}_{j;h}$ P R□1. In particular, the conditional cdf can have either finite or infinite right endpoint. When its support is bounded, relying on the Bayes decomposition exhibited in the previous paragraph, the desired result is straightforward: because there exists some t_0 ° 1 such that for all t ° t_0, $\bar{F}_{j;h}(t)$ = 0, then as t — 1 , the integral also becomes null. If on the contrary, $\bar{F}_{j;h}$ ° 0 for all t • 1, then, as previously, we can rewrite the quantity of interest in the following form:

$$\int_1^{\infty} t\,P(\Omega_i \bullet t)\,P\left[Z_{i;j} \circ x t \,\middle|\, \Omega_i \bullet t;\ \Omega_{i;h} = 1\right] dx = \frac{t\,\bar{F}_{j;h}(t)\,p_h}{P\left[\Omega_{i;h} = 1 \,\middle|\, \Omega_i \bullet t\right]} \int_1^{\infty} \frac{\bar{F}_{j;h}(x t)}{\bar{F}_{j;h}(t)}\,dx.$$

Since as t tends to infinity $t\,\bar{F}_{j;h}(t)$ tends to 0 (Proposition 3.1), $P\left[\Omega_{i;h} \,\middle|\, \Omega_i \bullet t\right]$ tends to □$_h$ ° 0 (Lemma 3.5) and since p_h ° 0, for the integral of interest to converge to 0 it suffices to prove that there exists some t_0 ° 1 such that for all t ° t_0,

$$\int_1^{\infty} \frac{\bar{F}_{j;h}(x t)}{\bar{F}_{j;h}(t)}\,dx \dagger 1 .$$

According to the assumption in Proposition 3.2, there exists some constants □□ in (0; 1), c□ • 0 and t□ ° 1 such that

$$(t \circ t^{\square}) \Rightarrow \frac{\bar{F}_{j;h}(x t)}{\bar{F}_{j;h}(t)} \S c^{\square} x^{\square 1/\square^{\square}}.$$

Hence, for all t ° t□,

$$\int_1^{\infty} \frac{\bar{F}_{j;h}(x t)}{\bar{F}_{j;h}(t)}\,dx \S c^{\square} \int_1^{\infty} x^{\square 1/\square^{\square}}\,dx = \frac{\square c^{\square}}{1 \square 1/\square^{\square}} \dagger 1 ;$$

which produces the desired outcome.

Ú h PH^{c}zt0u: p_h = □$_h$ = 0

By definition, for all h Pt1; : : : ; 2^d □ 1u, the equivalence below holds true:

$$(h \in H^c \setminus \{0\}) \Leftrightarrow (\Omega_h = 0) \Leftrightarrow (p_h = 0).$$

Consequently, when h P H^{c}zt0u, we have $P\left[Z_{i;j} \circ x \,\middle|\, \Omega_i \bullet t;\ \Omega_{i;h} = 1\right] = 0$ for all x • 0, and by extension

$$\int_1^{\infty} t\,P(\Omega_i \bullet t)\,P\left[Z_{i;j} \circ x t \,\middle|\, \Omega_i \bullet t;\ \Omega_{i;h} = 1\right] dx = 0;$$

for all t ° 0. This remains true as t — 1 .

Ú $h = 0 : p_h \neq 0$ and $\ell_h = 0$

Let us start again with the following decomposition :

$$\ell_{j;0}(t) = \frac{t p_0 \bar{F}_{j;0}(t)}{P\left(\bigcup_{i;0} = 1\right) \bullet t} \int_1^a \frac{\bar{F}_{j;0}(x t)}{\bar{F}_{j;0}(t)} dx.$$

Contrary to the case where $h \in H_{zH}(j)$, we cannot guarantee the convergence of $\ell_{j;0}(t)$ to 0 as t grows to infinity, since $P\left(\bigcup_{i;0} = 1\right) \bullet t$ now tends to 0 instead of a positive constant. Nonetheless, it is still possible to prove that it does not diverge to 1 . Indeed, notice that

$$\frac{t p_0 \bar{F}_{j;0}(t)}{P\left(\bigcup_{i;0} = 1\right) \bullet t} = \frac{t P(\bullet t) \bar{F}_{j;0}(t)}{P\left(\bullet t \bigcup_{i;0} = 1\right)};$$

and that $\bar{F}_{j;0}(t) \leq P\left(\bullet t \bigcup_{i;0} = 1\right)$. Hence,

$$\ell_{j;0}(t) \leq t P(\bullet t) \int_1^a \frac{\bar{F}_{j;0}(x t)}{\bar{F}_{j;0}(t)} dx.$$

We have already seen that according to the assumption in Proposition 3.2, there exists some constants $\in P(0;1)$, $c \bullet 0$ and $t \circ 1$ such that for all $t \circ t$,

$$\int_1^a \frac{\bar{F}_{j;0}(x t)}{\bar{F}_{j;0}(t)} dx \leq \frac{c}{1 \Box 1/}.$$

Moreover, by virtue of Equation (3.8), for all $\circ 0$ there exists some $t \circ 0$ such that for all $t \circ t$, $t P(\bullet t) \leq S(d 1) \dagger$. Fix some $\circ 0$ and set $:= \frac{(1 \Box 1/)}{c}$, then for all $t \circ \max(t; t)$, we have

$$\ell_{j;0}(t) \leq \frac{S(d 1) c}{1 \Box 1/} + \dagger 1 .$$

Observe that the smaller , i.e. the faster the limit dependence structure is reached, the smaller the bound of $\ell_{j;0}(t)$. Ideally, when all $\bar{F}_{j;0}$, $1 \leq j \leq d$, are rapidly varying, i.e. $c = 0$, we obtain the same result as in the case where $h \in H_{zH}(j)$. This would correspond in fact to the absence of hidden regular variation, like mentioned in Section 3.5 and Section 3.7 (Resnick, 2002; Heffernan and Resnick, 2005; Resnick, 2008).

3.8.3 About $p_{j;`}(k)$

For the sake of clarity, we give here a more explicit version of the statistic $p_{j;`}(k)$, which was defined as

$$p_{j;`}(k) := \int_1^a \frac{1}{k} \frac{n_k}{n_`} \prod_{i \in \mathcal{P}_k} \mathbb{I}\left(2_{i;j} \circ x \frac{n}{k}; p_{i;`} = 1\right) dx; \quad 1 \leq ` \leq H; \quad 1 \leq j \leq d;$$

where $n_k := \#\mathcal{Q}_k$ is the number of observations the radius of which exceeds n/k, and $n_\ell := \sum_{i \in \mathcal{Q}_k} \mathbb{1}\{\mathcal{C}_{i,\ell} = 1\}$ the size of class ℓ. Recall that ℓ is supposed to refer to some $\{1, \ldots, d\}$, that indexed the open face $\mathcal{F}_h^{d-1}$.

Let us begin by considering that k is fixed, and set

$$f_{j,\ell}(x) = \frac{1}{k} \frac{n_k}{n_\ell} \sum_{i \in \mathcal{Q}_k} \left(\mathbb{1}\{\mathcal{R}_{i,j} \leq x \frac{n}{k}\} \mid \mathcal{C}_{i,\ell} = 1 \right), \quad 1 \leq \ell \leq H, \; 1 \leq j \leq d.$$

Our statistic of interest, $\mathcal{C}_{j,\ell}(k)$ is none other than the integral over $x \geq 1$ of $f_{j,\ell}(x)$. Actually, because it relies on a finite set of $n \geq 1$ observations, $f_{j,\ell}(x)$ is a step function with support on the interval $\left[\min_{1 \leq i \leq n} \mathcal{R}_{i,j} k/n, \max_{1 \leq i \leq n} \mathcal{R}_{i,j} k/n \right]$. As $f_{j,\ell}$ only takes into account observations verifying $\mathcal{C}_{i,\ell} = 1$, we denote by $(\mathcal{R}_{1,j}^\ell, \ldots, \mathcal{R}_{n_\ell,j}^\ell)$ the sub-sample of n_ℓ observations within $(\mathcal{R}_{1,j}, \ldots, \mathcal{R}_{n,j})$, for any $j \in \{1, \ldots, d\}$. Corresponding ordered statistics are denoted by $\mathcal{R}_{(1,j)}^\ell \leq \ldots \leq \mathcal{R}_{(n_\ell,j)}^\ell$. Further consider

$$\mathcal{R}_{(n_\ell - u,j)}^\ell = \inf\{\mathcal{R}_{i,j}^\ell, 1 \leq i \leq n_\ell : \mathcal{R}_{i,j}^\ell \geq \frac{n}{k} u\},$$

the smallest observation $\mathcal{R}_{i,j}^\ell$ that exceeds n/k, and arbitrarily set $\mathcal{R}_{(n_\ell - u,1)}^\ell = 1$, then $f_{j,\ell}$ can be expressed as follows:

$$f_{j,\ell}(x) = \frac{1}{k} \frac{n_k}{n_\ell} \sum_{u=1}^{u_\ell - 1} u \left(x \in \left[\mathcal{R}_{(n_\ell - u,j)}^\ell \frac{k}{n}, \mathcal{R}_{(n_\ell - u+1,j)}^\ell \frac{k}{n} \right[\right).$$

In particular, when $x \geq \mathcal{R}_{(n_\ell,j)}^\ell k/n$, there is no $\mathcal{R}_{i,j}^\ell$, $1 \leq i \leq n_\ell$, such that $\mathcal{R}_{i,j}^\ell k/n \geq x$, and conversely, when $x \in [1, \mathcal{R}_{(n_\ell - u,j)}^\ell k/n[$, there are exactly $u_\ell + 1$ observations $\mathcal{R}_{i,j}^\ell$ in the sub-sample defined by $\mathcal{C}_{i,\ell} = 1$ that exceed $x n/k$. Therefore, the integral of $f_{j,\ell}(x)$ over all $x \geq 1$ verifies

$$\int_1^{\max_{1 \leq i \leq n} \mathcal{R}_{i,j}^\ell k/n} f_{j,\ell}(x)\, dx = \mathcal{C}_{j,\ell}(k) = \frac{n_k}{n} \frac{1}{n_\ell} \sum_{u=1}^{u_\ell - 1} u \left(\mathcal{R}_{(n_\ell - u+1,j)}^\ell - \mathcal{R}_{(n_\ell - u,j)}^\ell \right).$$

Let us dwell for a moment on this expression. The part

$$\frac{1}{n_\ell} \sum_{u=1}^{u_\ell - 1} u \left(\mathcal{R}_{(n_\ell - u+1,j)}^\ell - \mathcal{R}_{(n_\ell - u,j)}^\ell \right)$$

represents the integral under the empirical survival function of variable $\mathcal{R}_{i,j}$ conditional on i being in cluster ℓ and $\{\mathcal{R}_i\}_{(1)} \geq n/k$. When $\ell \in H(j)$, there should be a lot of extreme observations $\mathcal{R}_{i,j}$ in cluster ℓ, and this quantity should be very large. Conversely, in all clusters $\ell' \notin H(j)$, there should be very few to no extreme values on the j-th dimension, and the corresponding integral should be very small. Figure 3.6 and Figure 3.7 give an illustration of this phenomenon on exposures to the 6 nutrients and contaminants investigated in Section 3.6. Notice that dividing by n_ℓ

enables comparison between classes and avoids systematically selecting poor classifications. In a similar way, the term $n_k=n$ penalizes small values of k, which would otherwise always be preferred to higher ones and provide non-explicable groups, in the sense that they would contain too few observations to be interpreted. In terms of bias-variance compromise, intuitively it would generate overly wide variances for the final estimates of $t\,\square\,{}_{h}^{d\,\square\,1}$; h PHu to be reliable.

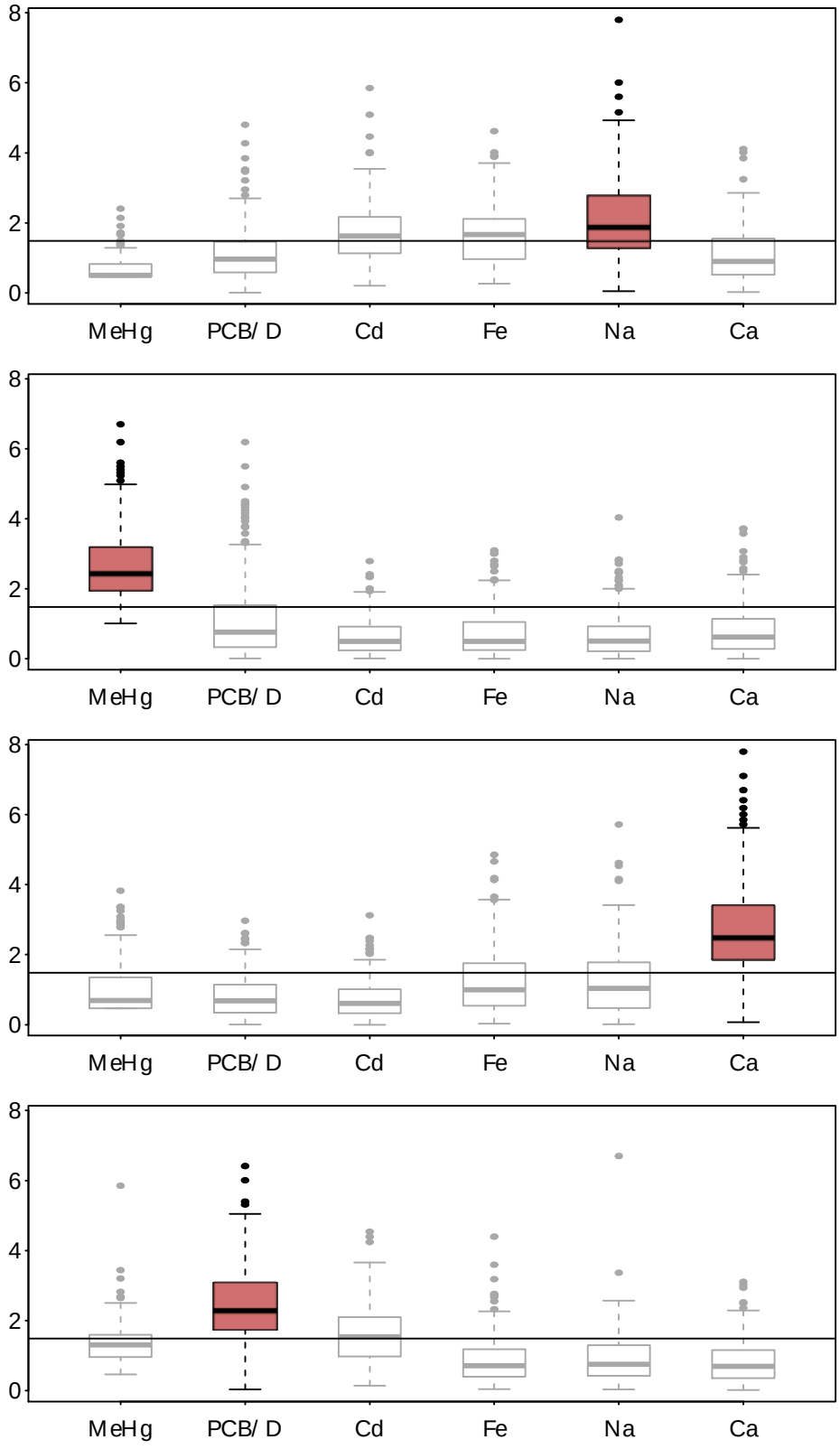


Figure 3.6 – Log-scaled marginal distributions of exposures to the 6 chemical elements within clusters 1, 2, 3 and 4, obtained for the couple $(k; \hat{\pi}(k))$ as defined in Section 3.4: distributions of contaminants with extreme exposures are displayed in red with black contours, while the others are white with grey contours. The thin horizontal line indicates k.

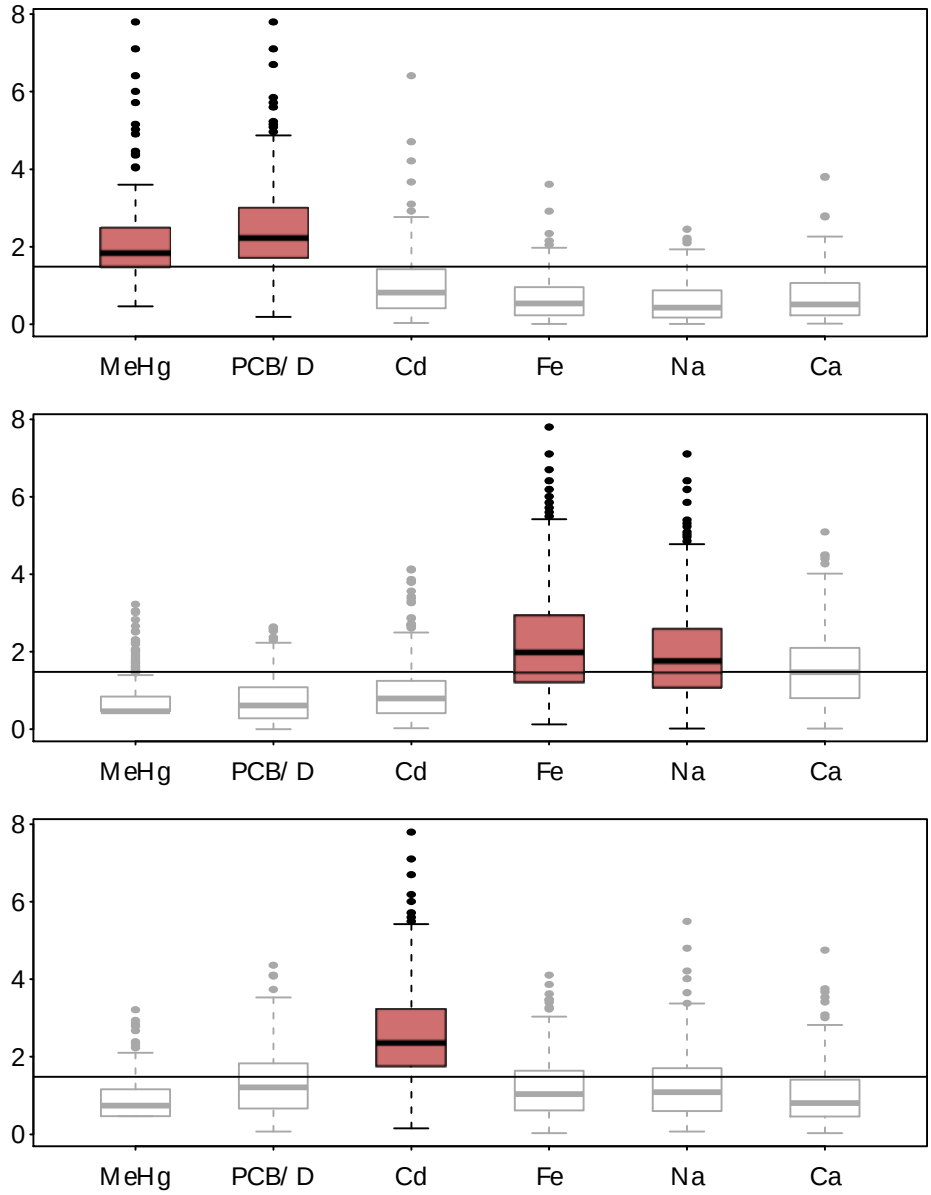


Figure 3.7 – Log-scaled marginal distributions of exposures to the 6 chemical elements within clusters 5, 6 and 7, obtained for the couple $(k; \Pi(k))$ as defined in Section 3.4: distributions of contaminants with extreme exposures are displayed in red with black contours, while the others are white with grey contours. The thin horizontal line indicates k.

3.8.4 A little more on PNS and spherical k-means

As a complement to the succinct description of Principal Nested Spheres and spherical k-means in Section 3.4.2, we provide here a more detailed overview of these algorithms accompanied by illustrative figures. With these additional specifications we are then able to discuss some technical choices in the implementation that were only briefly mentioned in the core of the chapter.

3.8.4.1 Principal Nested Spheres

Recall from Definition 3.4 that the geodesic distance between two points x and y of S^{d-1} (the unit sphere in R^d) is written

$$d_G(x;y) = \arccos x^1 y;$$

where x^1 stands for the transpose of vector x. Now consider any $(d-2)$-dimensional sub-sphere A_{d-2} in S^{d-1}. Relative to the geodesic distance on the sphere, its center and radius are respectively a point $v \in S^{d-1}$ and a distance $r \in (0; \pi/2]$ such that

$$A_{d-2} := A_{d-2}(v;r) := \{x \in S^{d-1} : d_G(v;x) = r\}.$$

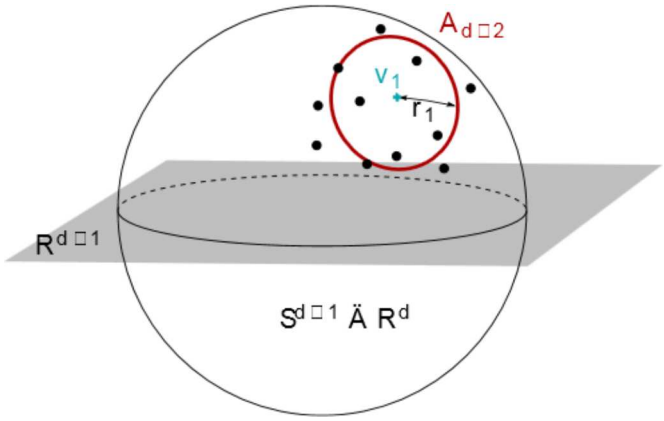
Given this representation, the signed distance between any point $x \in S^{d-1}$ and a sub-sphere $A_{d-2}(v;r)$ is naturally defined as

$$d_S(x; A_{d-2}(v;r)) := d_G(v;x) - r.$$

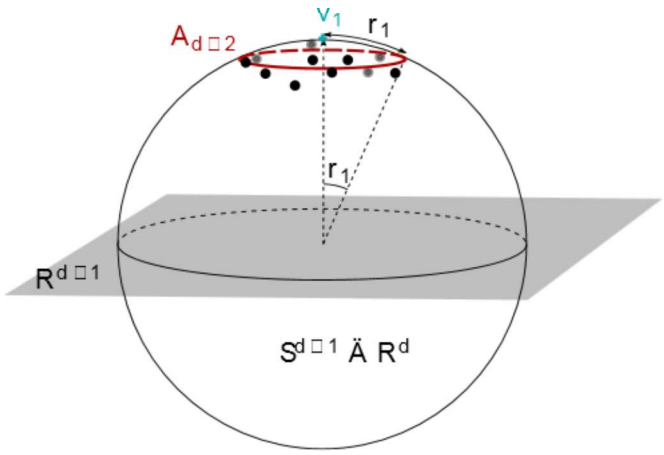
Equipped with these tools, the main steps of the PNS algorithm can be depicted as follows.

- **Step one** Find the sub-sphere of dimension $d-2$ that minimizes the mean squared residuals relative to the geodesic distance:

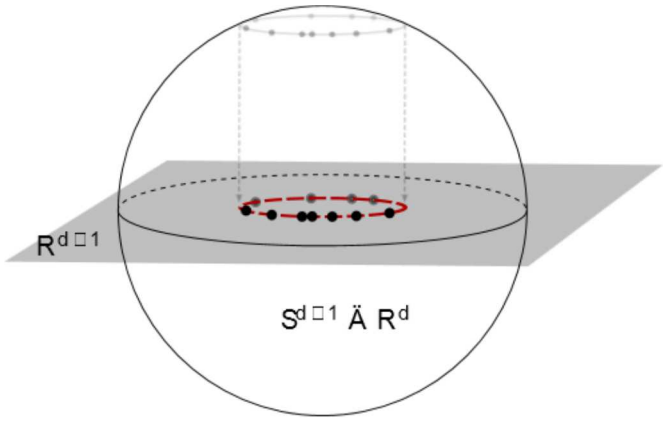
$$A_{d-2}(v_1;r_1) := \underset{\substack{v \in S^{d-1} \\ r \in (0;\frac{\pi}{2}]}}{\operatorname{argmin}} \sum_{i=1}^{n} d_S^2(x_i; A_{d-2}(v;r)).$$



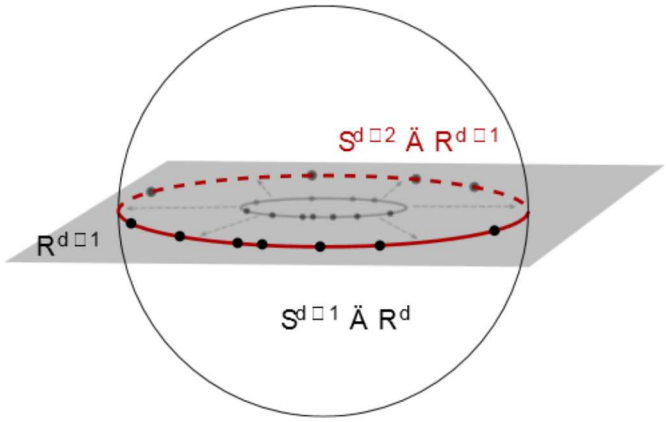
- **Step two** Use a rotation to place v_1 at the north pole.



Ú Step three Start by projecting the data on $A_{d\,\square 2}(v_1; r_1)$, then suppress the d-th dimension to place the sub-sphere equipped with the observations on $R^{d\,\square 1}$.



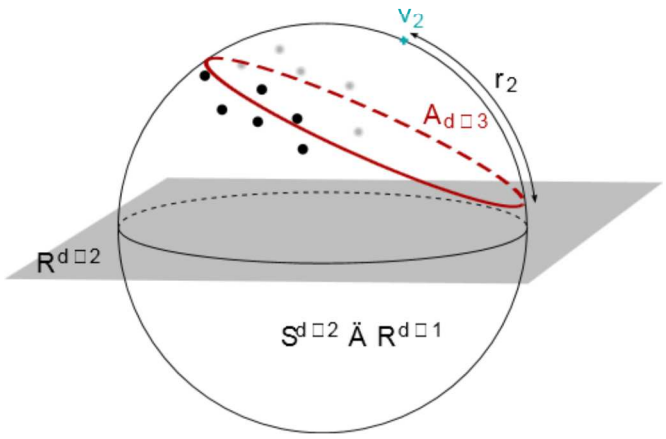
Ú Step four Scale $A_{d\,\square 2}$ so that it would coincide with the unit sphere $S^{d\,\square 2}$, and denote by $!_i^{(1)}$ P $S^{d\,\square 2}$ the new vectors of observations that are now elements of $R^{d\,\square 1}$.



Ú Step five Repeat the procedure to get

$$A_{d\,\square 3}(v_2; r_2) := \underset{\substack{v\,P\,S^{d\,\square 2} \\ r\,P(0;\frac{\square}{2}]}}{\operatorname{argmin}} \prod_{i=1}^{k} d_S^2(!_i^{(1)}; A_{d\,\square 3}(v; r))$$

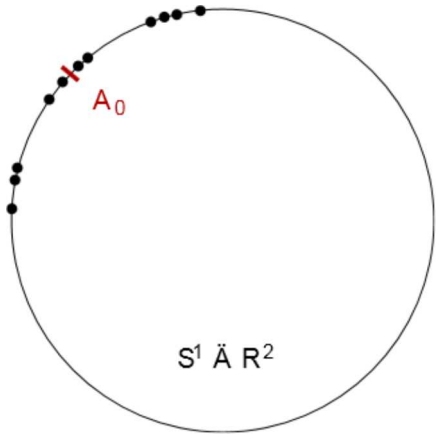
and identify it with $S^{d\,\square 3}$.



Ú Step six Continue the same way until the unit circle S^1 is reached, and define

$$A_0 := \underset{x \in S^1}{\operatorname{argmin}} \prod_{i=1}^{n_k} d_G^2(!_i^{(d-2)}; x);$$

the Fréchet mean of the data.



The computational algorithm designed to solve the least squares problem that defines each sub-sphere can be found in Section 3 of Jung et al. (2012) and the explicit formulas corresponding to the successive transformations of each PNS in Section 2. Many more details are provided in the supplementary materials associated with this seminal paper, concerning in particular the geometry of PNS. More importantly, a penalized version of the initial procedure is proposed to decide at each step whether small sub-spheres are more relevant candidates than those with maximal radius. Actually, both the numerical experiments in Section 3.5 and the case study in Section 3.6 are using this refined version of the PNS algorithm. We refer to the end of Section 1 (p.7) in the aforementioned supplementary materials for more details on the subject.

At the end of the procedure, we obtain a collection of unit spheres with dimensions ranging from 1 (the unit circle) to $d - 1$ (the space of the original data), which can be understood as a spherical equivalent of the principal components in PCA. In order to proceed with the rest of the analysis, we need to choose one of these $d - 1$ PNS and work with the corresponding projected angles. Recall that for all $j \in \{1; \ldots; d\}$ and $i \in \{1; \ldots; n_k\}$, $!_i^{(j-1)}$ denotes the projection of angle $!_i$ on S^{d-j} and define

$$\square_i^{(d-j)} := \prod_{\ell=1}^{j-1} \sin r_\ell \, d_G\left(!_i^{(j)}; A_{d-j}(v_j; r_j)\right);$$

the corresponding scaled residual. In short, scaling enables comparison between the deviations as if they were all measured on S_2^1 (see Jung et al., 2012, Sections 2.1 to 2.4). Then the relative variance encapsulated by PNS $d - j$, $1 \leq j \leq d$, is understood as

$$V_{d-j} := \frac{\sum_{i=1}^{n_k} \square_i^{(j-1)2}}{\sum_{\ell=1}^{d} \sum_{i=1}^{n_k} \square_i^{(\ell-1)2}}.$$

Given these notations, the selection heuristic evoked in Section 3.4.2 simply consists in picking the smallest sub-sphere S^{d-j} such that $V_{d-j-1} \bullet 0.1$, $V_{d-j} \bullet 0.1$ and $V_{d-j+1} \dagger 0.1$. Observe that contrary to PCA, there is no obvious link here between

$V_1, \ldots, V_{d-1}$ and the variance of the angles $V(\theta_i)$, $\theta_i \in \mathbb{R}^d$. Obviously, many refinements could and should be brought to our method in the near future, starting with a more adaptive way of identifying the "optimal" PNS. Moreover, though the situation was never encountered in our applications, many practical difficulties can arise and should receive appropriate attention. For instance it can so happen that there is no unique Fréchet mean: imagine a cloud of two points $(0,-1)$ and $(0,1)$, then there are two possible candidates for A_0, namely $(-1,0)$ and $(1,0)$. These intricate issues are left for future research.

3.8.4.2 Spherical k-means

Once a selected PNS, clustering is achieved using the spherical k-means algorithm. Before getting into detail, let us introduce a few additional notations. First, denote by $\theta_1, \ldots, \theta_{n_k}$ the cloud of angles projected onto the optimal PNS. For some fixed number of clusters H, the objective is to estimate the n_k vectors of class indicators $\delta_1, \ldots, \delta_{n_k}$, where $\forall i \in \{1, \ldots, n_k\}$, $\delta_i := (\delta_{i,1}, \ldots, \delta_{i,H})$. In spherical k-means, clusters are represented by their barycenter: for any set of $n \geq 1$ points $x_1, \ldots, x_n$ on the unit sphere $S^{d-1} \subset \mathbb{R}^d$ such that for all $i \in \{1, \ldots, n\}$, $x_i := (x_{i,1}, \ldots, x_{i,d})^t$, the barycenter function is written

$$B(x_1, \ldots, x_n) := \left(\frac{1}{n} \sum_{i=1}^{n} x_{i,1}, \ldots, \frac{1}{n} \sum_{i=1}^{n} x_{i,d} \right)^t,$$

and its projection on the unit sphere

$$SB(x_1, \ldots, x_n) := \frac{B(x_1, \ldots, x_n)}{\|B(x_1, \ldots, x_n)\|_2}.$$

For simplicity, we denote by b_h the barycenter of all angles in class h, i.e.

$$\begin{aligned} b_h &:= B\left(\theta_1 \mid \delta_{1,h} = 1, \ldots, \theta_{n_k} \mid \delta_{i,h} = 1 \right) \\ &:= \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \theta_{i,1} \mid \delta_{i,h} = 1, \ldots, \frac{1}{n_k} \sum_{i=1}^{n_k} \theta_{i,d} \mid \delta_{i,h} = 1 \right)^t, \end{aligned}$$

and by c_h its projection on the unit sphere, also called the concept vector:

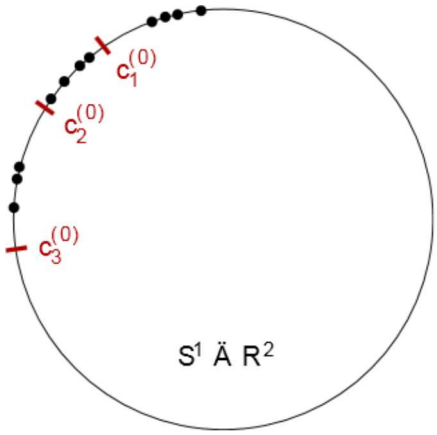
$$c_h := SB\left(\theta_1 \mid \delta_{1,h} = 1, \ldots, \theta_{n_k} \mid \delta_{i,h} = 1 \right) = \frac{b_h}{\|b_h\|_2}.$$

Given these notations, the spherical k-means algorithm tries to find the collection of indicators $\{\delta_{i,h}, 1 \leq i \leq n, 1 \leq h \leq H\}$, that minimize the intra-class geodesic variance, namely

$$GV\left(\delta_{i,h} \right)_{\substack{1 \leq i \leq n \\ 1 \leq h \leq H}} := \sum_{h=1}^{H} \sum_{i=1}^{n_k} d_G^2(\theta_i, c_h) \mid \delta_{i,h} = 1,$$

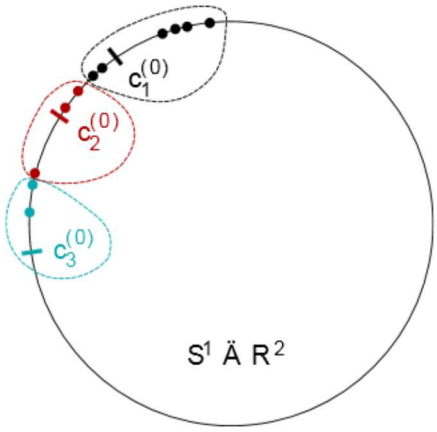
by implementing the following basic steps.

Ú Step one Choose a collection of initial concept vectors $c_1^{(0)}; \ldots; c_H^{(0)}$.



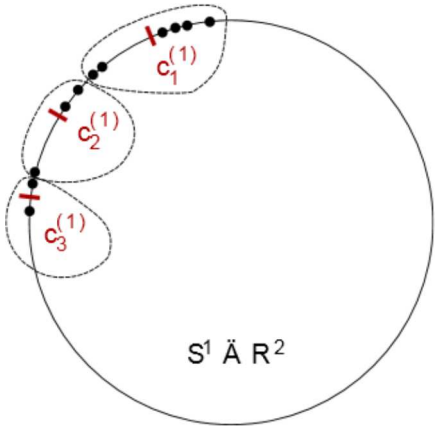
Ú Step two Affect the data to the class represented by the closest concept vector: for all $i \in \{1; \ldots; n_k\}$ and all $h \in \{1; \ldots; H\}$,

$$\delta_{i;h}^{(0)} := \begin{cases} 1 & \text{if } h = \operatorname*{argmin}_{1 \le \ell \le H} d_G^2\left(x_i; c_\ell^{(0)}\right), \\ 0 & \text{otherwise.} \end{cases}$$



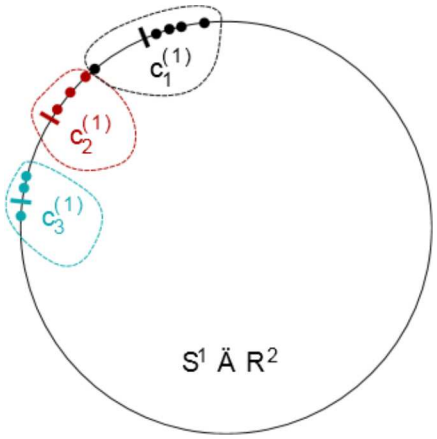
Ú Step three Identify the concept vectors in each new class: for all $h \in \{1; \ldots; H\}$,

$$c_h^{(1)} := \frac{\sum_{i=1}^{n_1} \delta_{i;h}^{(0)} x_i}{\sum_{i=1}^{n_1} \delta_{i;h}^{(0)} = 1}; \ldots; \frac{\sum_{i=1}^{n_k} \delta_{i;h}^{(0)} x_i}{\sum_{i=1}^{n_k} \delta_{i;h}^{(0)} = 1}.$$



Ú Step four Actualize the classes: for all $i \in \{1,\dots,n_k\}$ and $h \in \{1,\dots,H\}$,

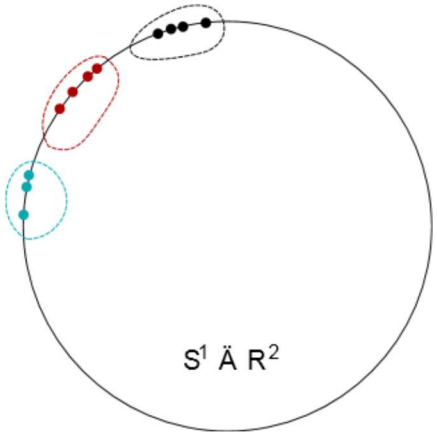
$$\Delta_{i,h}^{(1)} := \begin{cases} 1 & h = \operatorname*{argmin}_{1\le \ell\le H} d_G^2\left(x_i, c_\ell^{(1)}\right) \\ 0 & \text{otherwise.} \end{cases}$$



Ú Step five Repeat the procedure until

$$GV\left(\left\{\Delta_{i,h}^{(\ell-1)}\right\}_{\substack{1\le i\le n\\ 1\le h\le H}}\right) - GV\left(\left\{\Delta_{i,h}^{(\ell)}\right\}_{\substack{1\le i\le n\\ 1\le h\le H}}\right)$$

becomes less than a user-defined tolerance.



The option "s" in the R package skmeans that we used in our applications stipulates how the initial concept vectors $c_1^{(0)},\dots,c_H^{(0)}$ are to be chosen. Specifically, for $H \ge 2$ desired clusters, they are successively selected according to the following recurrence relation:

$$\begin{cases} c_1^{(0)} := A_0; \\ c_h^{(0)} := \operatorname*{argmax}_{x_i,\ 1\le i\le n_k} d_G^2\left(x_i, SB\left(c_1^{(0)},\dots,c_{h-1}^{(0)}\right)\right), & h \ge 2. \end{cases}$$

In words, the first initial concept vector is set as the Fréchet mean of the sample of projected angles, and the rest as the $H-1$ observations farthest away from all already picked concept vectors. This produces an initial clustering with centers as scattered as possible. Obviously, many other initialization techniques may have been applied, e.g. picking the first concept vectors at random. The present version has been chosen in accordance with our belief that after projection, the angles corresponding to different faces are disseminated on different regions of the retained PNS while those belonging to the same face are concentrated on a common neighborhood.

The main advantage of the spherical k-means algorithm is that it is very simple to implement. However, it can often happen that it remains stuck at a local minimum of

the intra-class geodesic variance function. To counteract this undesirable effect, many refinements have been proposed in the literature (see for instance Dhillon et al., 2002 and the references therein). As a first go, we confined ourselves to the basic version of spherical k-means, leaving further considerations on algorithmic improvement for future research.

4

A MINIMUM VOLUME SET APPROACH TO DIETARY RISK-BENEFIT ANALYSIS

In the same spirit as in Chapter 3, we propose an alternative method to inspect the multivariate distribution of the exposure to multiple food chemicals, which accounts for the variability of the contamination of foodstuffs. Directly inspired from typical statistical learning techniques, this non-parametric approach is no longer focused on extreme events. One of its advantages is that it can be very naturally extended to the identification of optimal dietary habits, in the sense that they realize a compromise between toxicological risk and nutritional benefit. From a practical point of view, such results would facilitate public communication of general dietary recommendations. The present chapter corresponds to a paper currently being written in collaboration with J.Tressou (INRA Met@risk, France) and S. Clémençon (Télécom ParisTech, France). It is not complete yet, in particular we are still working on the practical applications of the theoretical results introduced therein.

It is the major purpose of this chapter to show how to adapt recent (unsupervised) machine-learning techniques, specifically introduced to deal with very high-dimensional data in a non-parametric manner (avoiding thus the curse of dimensionality), to food risk/ benefit analysis. Precisely, a variant of the minimum-volume set methodology (MV-set in abbreviated form), originally investigated in Polonik (1997) (see also Scott and Nowak, 2006), is proposed in order to determine confidence or predictive regions for the joint dietary exposure to a variety of chemicals and nutrients present in the food. The main originality of this problem lies in the fact that a natural empirical counterpart of the dietary exposure is of the form of a generalized U-statistic, based on the combination of consumption survey data with a database gathering measures of contents of a variety of chemicals and nutrients in most food items. Mainly due to the large number of foodstuffs involved in the observed diets, this statistic cannot be calculated in general, its computation requiring to sum over a prohibitive number of terms. Following in the footsteps of Bertail and Tressou (2006), we replace the latter with an incomplete U-statistic (Blom, 1976) the computation of which is numerically feasible, and we establish a novel uniform deviation result, which shows that this approximation stage does not damage the learning rate of the MV-set procedure. Next, similar concepts and results are applied to identify regions where the multivariate distribution of dietary habits is mostly concentrated and where types of exposure simultaneously remain within toxicological values of reference (limitation of the risk) and recommended dietary allowances (preservation

of the benefit) with maximum probability. Statistical results, i.e. rate bounds guaranteeing the performance of the generic learning techniques we propose, are stated.

The chapter is structured as follows. A detailed account of the statistical issues related to food risks and benefits tackled in the chapter and the learning methods proposed to deal with them is given in Section 4.1, together with theoretical results claiming their validity. Technical proofs are postponed to Section 4.2.

4.1 theoretical analysis and methods

4.1.1 Assessment of dietary exposure to chemicals and nutrients: a MV-set formulation

We are concerned here with dietary types of exposure to $d \bullet 1$ different food chemicals, contaminants or nutrients, over a certain statistical population of interest during a given period of time, say a week like in the INCA2 database. Foods are classified according to some given nomenclature that accounts for H families of products, indexed by label $h \mathrel{P} t1;\dots;Hu$. The joint dietary exposure can be then described by a random vector $X := (X_1;\dots;X_d)$, where

$$X_j := \prod_{h=1} C_h \square Q_{h;j}. \tag{4.1}$$

for $1 \S j \S d$, denoting by C_h the quantity of food item h consumed per week by an individual drawn at random in the studied population and by $Q_{h;j}$ the (random) content related to food item h and component j. In the field of food safety, risk assessors are interested in building confidence/ predictive regions for the exposure X in R_+^d:

$$R_\square := \arg\min_{} \left(L(R) : P(X \mathrel{P} R) \bullet \square;\ R \ÄR_+^d \text{ Borelian} \right); \tag{4.2}$$

where $\square \mathrel{P} (0;1)$ and the Lebesgue measure on R_+^d is denoted by L. For values of the level $\square$ close to 1, such minimum volume sets (MV-sets in short, see Scott and Nowak, 2006) describe regions where the distribution of exposure is most concentrated, those lying in their complementary sets being possibly interpreted as "abnormal". The construction of predictive/ confidence regions for the dietary exposure is based on the observation of the dietary habits of $n \bullet 1$ individuals independently drawn from the population, yielding an iid sample $tC_i := (C_{i;1};\dots;C_{i;H}); 1 \S i \S nu$. They are combined with databases where $m_{h;j} \bullet 1$ iid measures of the amount of pollutant or nutrient j present in food h are listed for all $h \mathrel{P} t1;\dots;Hu$ and $j \mathrel{P} t1;\dots;du$. The corresponding vectors of contents are denoted by $Q_{h;j} := (Q_{h;j}^1;\dots;Q_{h;j}^{m_{h;j}})$. Usually, the following hypotheses are supposed to hold true.

Assumption 4.1 For all couples $(h;\tilde{h}) \in \{1;\dots;H\}^2$ such that $h \neq \tilde{h}$ and any $j \in \{1;\dots;d\}$, level $Q_{h;j}$ is independent from $Q_{\tilde{h};j}$.

Assumption 4.2 For all couples $(j;\tilde{j}) \in \{1;\dots;d\}^2$ such that $j \neq \tilde{j}$ and any $h \in \{1;\dots;H\}$, level $Q_{h;j}$ is independent from $Q_{h;\tilde{j}}$.

Assumption 4.3 For any $(h;j) \in \{1;\dots;H\} \times \{1;\dots;d\}$, level $Q_{h;j}$ is independent from consumption vector C_i.

Assumption 4.1, which formally states that nutrients and contaminants are independently assimilated by foodstuffs, could easily be relaxed without any substantial impact on the present approach. Its applicability depends mostly on the level of detail provided by the database(s) at hand. On the contrary, Assumption 4.2 and Assumption 4.3 are necessary. They respectively stipulate that the contents of one foodstuff are independent from that of others and that dietary habits are not dictated by nutritional or toxicological characteristics of the food. From a more practical point of view, Assumption 4.2 is likely to be true if consumers get their supplies from various productions and Assumption 4.3 if they do not base their consumption decision on a systematic scrutiny of the composition of the food.

Based on these data, the probability involved in the constraint of the MV-set problem stipulated in Equation (4.2) is estimated by

$$\hat{P}(X \in R) := \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell_{1;1}=1}^{m_{1;1}} \cdots \sum_{\ell_{H;d}=1}^{m_{H;d}} \mathbb{I}\left\{ \sum_{h=1}^{H} C_{i;h} Q_{h;j}^{\ell_{h;j}} \Big|_{1 \leq j \leq d} \in R \right\}, \tag{4.3}$$

with

$$\alpha := n \prod_{j=1}^{d} \prod_{h=1}^{H} m_{h;j}$$

and $\mathbb{I}\{u\}$ the indicator function. In practice, all types of sets cannot be explored; the search is restricted to a class $\mathcal{R}$ of Borelian sets, the complexity of which is controlled, like hypercubes or ellipses (see the next subsection). The level α is in turn replaced by $\alpha - \varepsilon$, where ε is some tolerance level that depends, roughly speaking, on the order of magnitude of

$$\sup_{R \in \mathcal{R}} \left| \hat{P}(X \in R) - P(X \in R) \right|.$$

Hence, one should ideally try to solve the constrained minimization problem:

$$\min_{R \in \mathcal{R}} \mathrm{Leb}(R) \text{ subject to } \hat{P}(X \in R) \geq \alpha - \varepsilon. \tag{4.4}$$

The major difference with the formulation in Scott and Nowak (2006) lies in the fact that the estimate of the probability involved in the mass constraint is here of the form

of a (generalized) U-statistic. Thus, the study of the performance of solutions of Equation (4.4) includes the proof of concentration results for U-processes (i.e. collections of U-statistics). Unfortunately, averaging over the $n \cdot \sum_{j=1}^{d} \sum_{h=1}^{H} m_{h,j}$ terms appearing in Equation (4.3) is generally numerically unfeasible, even for moderate sample sizes. In Bertail and Tressou (2006) for instance, where the estimation of the probability that the exposure to Ochratoxin A exceeds a critical threshold is considered, this corresponds to $4 \cdot 10^{21}$ terms! As shall be seen below, the statistic in Equation (4.3) can be uniformly approximated by a "Monte-Carlo" version, the computation cost of which is drastically reduced.

Remark 4.4 – Alternative approaches. We underline that the MV-set methodology is by no means the sole possible approach for constructing predictive regions. For instance, density sub-level sets can be built by means of non-parametric density estimation techniques, see Tsybakov (1997) and the references therein. However, when trying to implement such "plug-in" alternatives, even for moderate dimensions, one faces significant computational problems inherent to the curse of dimensionality. This motivates the machine-learning approach promoted here, which avoids a preliminary density estimation stage, while focusing directly on performance optimization. One may also refer to Vert and Vert (2006) or Steinwart et al. (2005) for closely related techniques.

4.1.2 Uniform approximation of generalized U-statistics by their incomplete versions

For clarity, we recall below the definition of generalized U-statistics. Properties and asymptotic theory of U-statistics can be found in Lee (1990).

Definition 4.5 – Generalized U-statistic. Let $K \geq 1$, $(d_1, \ldots, d_K) \in \mathbb{N}^K$ and consider the vectors $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$, $1 \leq k \leq K$, corresponding to K independent samples of iid random variables, taking their values in some space $\mathcal{X}_k$ with distribution P_k respectively. The generalized (or K-sample) U-statistic of degrees $(d_1, \ldots, d_K)$ with kernel $\varphi : \mathcal{X}_1^{d_1} \times \cdots \times \mathcal{X}_K^{d_K} \to \mathbb{R}$, square integrable with respect to the probability distribution $P_1^{\otimes d_1} \otimes \cdots \otimes P_K^{\otimes d_K}$, is defined as

$$U_n(\varphi) := \frac{1}{\prod_{k=1}^{K} \binom{n_k}{d_k}} \sum_{I_1} \cdots \sum_{I_K} \varphi(X_{I_1}^{(1)}, X_{I_2}^{(2)}, \ldots, X_{I_K}^{(K)}), \tag{4.5}$$

where $\sum_{I_k}$ refers to the summation over all $\binom{n_k}{d_k}$ subsets $X_{I_k}^{(k)} := (X_{i_1}^{(k)}, \ldots, X_{i_{d_k}}^{(k)})$ related to a set I_k of d_k indexes $1 \leq i_1 < \cdots < i_{d_k} \leq n_k$. It is said symmetric when φ is permutation symmetric in each set of d_k arguments $X_{I_k}^{(k)}$.

Observe that the functional in Equation (4.3) corresponds to a K-sample U-statistic of degrees $(1;1;\dots;1)$, with $K = d - H + 1$ and kernel given by:

$$_R(c;q) := I_{\%}\left(\prod_{h=1}^{H} c_h \cdot q_{h;j_{1\leq j\leq d}} \geq R\right);$$

for $c := (c_1;\dots;c_H) \in R_+^H$ and $q := (q_{h;j})_{j=1;\dots;d} \in R_+^{H\times d}$. Beyond this example, many statistics used for pointwise estimation or hypothesis testing are actually U-statistics (e.g. the sample variance, the Gini mean difference, the Wilcoxon Mann-Whitney statistic, the Kendall tau). Their popularity mainly arise from their "reduced variance" property: the statistic $U_n(\ )$ has minimum variance among all unbiased estimators of the parameter

$$\theta(\ ) := E\left(X_1^{(1)};\dots;X_{d_1}^{(1)};\dots;X_1^{(K)};\dots;X_{d_K}^{(K)}\right).$$

Classically, the limit properties of these statistics (law of large numbers, central limit theorem, etc.) are investigated in an asymptotic framework stipulating that, as the full sample size

$$n := n_1 + \cdots + n_K$$

tends to infinity, we have $n_k/n \to \lambda_k > 0$ for all $k \in \{1;\dots;K\}$. They can be established by means of a linearization technique (see Hoeffding, 1948), permitting to write $U_n(\ )$ as a sum of K basic sample mean statistics (of the order $O_P(1/\sqrt{n})$ each, after recentering), plus possible degenerate terms called degenerate U-statistics.

As mentioned before, in practice, the number $\pm_{k=1}^{K}\binom{n_k}{d_k}$ of terms to be summed up to compute Equation (4.5) is generally prohibitive. As a remedy to this computational issue, in the seminal contribution of Blom (1976), the concept of incomplete generalized U-statistic has been introduced: the summation in Equation (4.5) is replaced by a summation involving much less terms, extending only over low cardinality subsets of the $\binom{n_k}{d_k}$ d_k-tuples of indexes, $1 \leq k \leq K$. In the simplest formulation, the subsets of indexes are obtained by sampling with replacement, leading to the following definition.

Definition 4.6 – Incomplete Generalized U-statistic. Let $B \geq 1$. The incomplete version of the U-statistic in Equation (4.5) based on B terms is defined by:

$$\tilde{U}_B(\ ) := \frac{1}{B}\sum_{(I_1;\dots;I_K)\in D_B} (X_{I_1}^{(1)};\dots;X_{I_K}^{(K)}); \tag{4.6}$$

where D_B is a set of cardinality B built by sampling with replacement in the set

$$L := \left\{(i_1^{(1)};\dots;i_{d_1}^{(1)});\dots;(i_1^{(K)};\dots;i_{d_K}^{(K)})\right\}_{1\leq i_1^{(k)}<\dots<i_{d_k}^{(k)}\leq n_k;\ 1\leq k\leq K}.$$

Remark 4.7 – Alternative sampling schemes. We point out that, as proposed in Janson (1984), other sampling schemes could be considered, in particular sampling without replacement or Bernoulli sampling. The results of this chapter could be extended to these situations. For the sake of simplicity, we restrict our attention here to the sampling with replacement scheme.

In practice, B should be chosen much smaller than the cardinality of L, namely $\#L := \prod_{k=1}^{K} \binom{n_k}{d_k}$, in order to overcome the computational issue previously mentioned. We emphasize that the cost related to the computation of the value taken by the kernel φ at a given point $(x_{I_1}^{(1)}, \ldots, x_{I_K}^{(K)})$ depending on the form of φ is not considered here, focus is solely on the number of terms involved in the summation. As an estimator of $\theta(P)$, the statistic in Equation (4.6) is still unbiased but its variance is naturally larger than that of Equation (4.5). Precisely, we have

$$V\left[U_B(\varphi)\right] = \left(1 - \frac{1}{B}\right) V\left(U_n(\varphi)\right) + O\left(\frac{1}{B}\right) \quad \text{as } B \to +\infty,$$

refer to Lee (1990, p.193). Incidentally, we underline that the empirical variance of Equation (4.5) is not easy to compute since it involves summing approximately $\#L$ terms and bootstrap techniques should be used for this purpose, as proposed in Bertail and Tressou (2006). The asymptotic properties of incomplete U-statistics have been investigated in several articles, see Janson (1984); Brown and Kildea (1978); Enqvist (1978). The angle embraced in the present chapter is of different nature: the key idea we promote here is to use incomplete versions of collections of U-statistics in learning problems such as that described in Section 4.1.1. The following result shows that this approach solves the numerical problem, while not damaging the learning rates. It reveals that, under adequate complexity assumptions on the considered collection Φ of (symmetric) kernels φ (refer to Dudley, 1999), concentration results established for U-processes (i.e. collections of U-statistics) may extend to their incomplete versions.

Theorem 4.8 – Maximal deviation. Let Φ be a collection of bounded symmetric kernels on $\mathcal{X} := \prod_{k=1}^{K} \mathcal{X}_k^{d_k}$. Suppose that Φ is a VC major class of functions with finite Vapnik-Chervonenkis dimension V and that $M_\Phi := \sup_{(\varphi, x) \in \Phi \times \mathcal{X}} |\varphi(x)| < +\infty$. Then, the following assertions hold true.

i) For all $\varepsilon > 0$, we have: $\forall n = (n_1, \ldots, n_K) \in \mathbb{N}^K$, $\forall B \geq 1$,

$$\mathbb{P}\left[\sup_{\varphi \in \Phi} \left|U_B(\varphi) - U_n(\varphi)\right| > \varepsilon\right] \leq 2(1 + \#L)^V e^{-B\varepsilon^2 / M_\Phi^2}.$$

ii) For all $\alpha \in (0,1)$, with probability at least $1 - \alpha$, we have: $\forall n_k \geq 1, 1 \leq k \leq K$,

$$\sup_{P} \left| U_B(\theta) - \mathbb{E}\left[U_B(\theta)\right] \right| \leq M \left(2 \sqrt{\frac{2V \log(1+\#L)}{n}} + \sqrt{\frac{\log(2/\alpha)}{n}} + \frac{V \log(1+\#L) + \log(4/\alpha)}{B} \right),$$

where

$$n := \min\{\lfloor n_1/d_1 \rfloor, \ldots, \lfloor n_K/d_K \rfloor\}$$

and $\lfloor x \rfloor$ denotes the integer part of any real number x.

We refer to Section 4.2 for the proof. Observe that, with the asymptotic settings previously specified, $n \asymp n$ and $\log(\#L) \asymp \log(n)$ as $n \to +\infty$. The bounds stated above show that, for a number $B := B_n$ of terms tending to infinity as $n \to +\infty$ at a rate $O(n)$, the maximal deviation $\sup_{P} |U_B(\theta) - I(\theta)|$ is asymptotically of the order $O_P(\sqrt{\log(n)/n})$, just like $\sup_{P} |U_n(\theta) - I(\theta)|$. Remarkably, except in the case $K = 1$ and $d_K = 1$ solely, using such incomplete U-statistics thus yields a significant gain in terms of computational cost and fully preserves the order of the probabilistic upper bounds for the uniform deviations. Before showing how these results apply to the analysis of the dietary exposure distribution, a few remarks are in order.

Remark 4.9 – On the complexity assumption. We point out that, as can be seen by examining their proof in Section 4.2, the results above could be extended to other complexity measures than the VC dimension, such as Rademacher averages (Boucheron et al., 2005). However, the confidence regions we shall consider in practice being of the form of the union of a limited number of hypercubes, the finite VC dimension hypothesis is sufficient to provide a validity framework to the present analysis.

Remark 4.10 – Learning tasks based on optimization of U-statistics. The uniform deviation result stated in Theorem 4.8 can be proved very useful much beyond the framework described in Section 4.1.1. Indeed, statistical learning problems where the empirical performance criterion one seeks to optimize is of the form of a (generalized) U-statistic have recently been the subject of a good deal of attention in the literature: supervised ranking (Clémençon et al., 2008), learning on graphs (Biau and Bleakley, 2006) or dissimilarity-based clustering (Clémençon, 2011) for instance. The result above permits to show that, in such problems, the empirical criterion can be replaced by an incomplete version of much simpler computation, with only a slight impact on the learning rate, provided that the parameter B is suitably chosen.

Let us now come back to the specific learning task formulated in Section 4.1.1. In order to avoid the computation of Equation (4.3), which involves summing over $\Box := n \prod_{j=1}^{d} \prod_{h=1}^{H} m_{h;j}$ terms based on $N := n + \prod_{j=1}^{d} \prod_{h=1}^{H} m_{h;j}$ observations, Theorem 4.8 suggests to draw with replacement B times in the set of indexes

$$L := \{1;\ldots;n\} \times \prod_{j=1}^{d} \prod_{h=1}^{H} \{1;\ldots;m_{h;j}\}, \quad \mathrm{Card}\, L = \Box;$$

yielding an index set D_B of cardinality B. For any borelian $R \subset R^d$, the probability that the dietary exposure lies in the region R is estimated by the incomplete U-statistic:

$$\hat{P}_B(X \in R) := \frac{1}{B} \sum_{(i;\ell_{h;j}) \in D_B} \mathbb{I}\left\{ \sum_{h=1}^{H} C_{i;h} Q_{h;j}^{\ell_{h;j}} \in R, \; 1 \leq j \leq d \right\}.$$

Suppose that the class R is of finite VC dimension V. Let $\alpha \in (0;1)$ be the target mass and $\beta \in (0;1)$ the desired confidence level. Notice that in the present case,

$$\Box = \min\left\{ n; \min\{ m_{h;j} : 1 \leq j \leq d; 1 \leq h \leq H \} \right\}.$$

Further define the complexity penalty by

$$\Phi(B;N;\beta) := 2\left(\sqrt{\frac{2V \log(1+\Box)}{\Box}} + \sqrt{\frac{\log(2/\beta)}{\Box}} + \sqrt{\frac{V \log(1+\Box) + \log(4/\beta)}{B}} \right); \tag{4.7}$$

and consider the solution $\hat{R}_{\Box}$ of the constrained optimization problem:

$$\min_{R \in \mathcal{R}} L(R) \text{ subject to } \hat{P}(X \in R) \geq \alpha - \Phi(B;N;\beta). \tag{4.8}$$

The result below shows that, if the number B of exposure values computed through the sampling scheme and $\Box$ are of the order O(N), the performance of $\hat{R}_{\Box}$ is then comparable to that of the region the selection of which is based on the quantities in Equation (4.3) (see Section 4.2 for a sketch of proof).

Corollary 4.11 *For all $\alpha \in (0;1)$, we have with probability at least $1 - \beta$:*

$$L\left(\hat{R}_{\Box}\right) \leq \inf_{R \in \mathcal{R}: P(X \in R) \geq \alpha} L(R) \text{ and } P\left[X \in \hat{R}_{\Box} \right] \geq \alpha - 2\Phi(B;N;\beta).$$

From a practical perspective, the constrained optimization program in Equation (4.8) will be here performed over a collection of sets obtained as the union of small hypercubes by the means of a simplistic sorting algorithm, detailed at length in the next subsection. Alternative techniques, whose implementation is less immediate, such as those based on Dyadic Decision Trees (DDT) could be also considered, see Scott and Nowak (2006, Subsection 6.2) for further details.

4.1.3 Empirical MV-set estimation based on hypercubes

We now turn to the issue of solving Equation (4.8) from a practical perspective. For simplicity, such regions are built by binding together hypercubes of the positive orthant $\mathbb{R}_+^d$. Suppose that the exposure X takes its values in the compact set $[0;1]^d$, even if it means dividing each component of the exposure random vector by a supposedly finite essential upper bound. Observe that this assumption makes the present approach quite different from that introduced in Chapter 3, where the support of the distribution of individual types of exposure could extend to infinity. Since extremes are no longer of interest here, such refinements are of little concern. Now let $k \geq 1$ and consider the partition of the unit cube $\prod_{j=1}^{d} [s_j/k\,;\,(s_j+1)/k]$ made of sub-cubes of side length $1/k$, with $k \geq 1$ and $s_j \in \{0,\ldots,k-1\}$ for $1 \leq j \leq d$. Denote by $C_1,\ldots,C_M$ these sub-cubes, with $M = k^d$, and consider the collection $\mathcal{R}_k$ of subsets obtained as the union of such cubes. Observe that $\mathcal{R}_k$ is of finite cardinality: $\#\mathcal{R}_k = 2^{k^d}$. In this situation, the theoretical results established in the previous subsection apply, with the penalty

$$\Phi_k(B,N,\alpha) := 2\left(c\,\frac{2k^d\log(1+k^d)}{\alpha} + c\,\frac{\log(2/\alpha)}{\alpha} + c\,\frac{k^d\log(1+k^d)+\log(4/\alpha)}{B}\right),$$

and a solution of the constrained minimization problem in Equation (4.8) can be obtained in two steps, as follows. Let α, ε in $(0,1)$ and $B \geq 1$.

1. Sort the sub-cubes C_m, $1 \leq m \leq M$, so that:

$$\widehat{P}_B(X \in C_{1,M}) \geq \cdots \geq \widehat{P}_B(X \in C_{M,M}).$$

2. Bind together the cubes sequentially, until the incomplete U-statistic estimating the mass of the resulting set exceeds $\alpha - \varepsilon - \Phi_k(B,N,\alpha)$, yielding the region:

$$\widehat{R}_{k,\alpha} := \bigcup_{m=1}^{M_\alpha} C_{m,M},$$

where

$$M_\alpha := \min\left\{M \geq 1 : \sum_{m=1}^{M} \widehat{P}_B(X \in C_{m,M}) \geq \alpha - \varepsilon - \Phi_k(B,N,\alpha)\right\}.$$

Let $1 \leq k_* \leq k^* < +\infty$. The issue of selecting the "resolution level" $k \in \{k_*,\ldots,k^*\}$ automatically can be handled through complexity penalization, as shown in Scott and Nowak (2006, Section 4). Precisely, one should pick the value

$$\widehat{k} := \operatorname*{argmin}_{k_* \leq k \leq k^*} \left(L\left(\widehat{R}_{k,\alpha}\right) + \Phi_k(B,N,\alpha 2^{-k})\right),$$

minimizing thus a complexity-penalized version of the volume, in order to approximate the MV-set over $\bigcup_k \mathcal{R}_k$ as accurately as possible without overfitting the data. An oracle inequality showing that the chosen set $\widehat{R}_{\widehat{k},\alpha}$ corresponds to an optimal trade-off between excess volume and missing mass can be straightforwardly derived from the analysis carried out in Section 4.1.2, just like in Scott and Nowak (2006, Theorem 7). Details are left to the reader.

4.1.4 Optimal regions in the consumption space in regard to dietary risks and benefits

In food safety, one topic of crucial interest is to determine optimal regions in the dietary consumption space $\mathbb{R}_+^H$ in the sense that they achieve a compromise between toxicological risk and nutritional benefit. Indeed, for each nutrient and contaminant, experts define maximum threshold levels of exposure, generally called dietary intake limits (DIL), above which health issues due to excessive supply of chemicals are likely to occur. Equivalent lower bounds are defined for nutrients, which indicate nutritional deficiency. From a public health point of view, it is then of particular interest to identify the dietary habits that offer the best chance of respecting both lower and upper DIL and draw general, easily comprehensible dietary guidelines to the concerned population, in the same spirit as "eat at least five fruits and vegetables a day". Since it is very difficult to recommend people to completely alter their habits, public health institutes are more interested in determining realistic food baskets, already consumed by a non-negligible amount of persons, which they could set as examples for the rest of the population. This problem can be addressed by means of concepts and tools very similar to those investigated in Section 4.1.1. Consider $d \bullet 1$ pollutants and nutrients, present in a nomenclature of foodstuffs indexed by $h \in \{1,\dots,H\}$ at (random) concentration levels $Q_{h;1},\dots,Q_{h;d}$. We denote by $\ell^{\square}(j)$ (resp. $\ell_{\square}(j)$) the maximum (resp. minimum) recommended exposure for element j. By convention, if j is a contaminant, we set $\ell_{\square}(j) = 0$. Notice that whereas those limits are given in terms of amounts of nutrients, regarding contaminants they usually also depend on the body weights of consumers. Hence, from now on, when we write $Q_{h;j}$ to designate the amount of chemical j in food family h, if j is a contaminant then we implicitly refer to its standardized version $Q_{h;j}/w_i$, with w_i the body weight of individual i.

Equipped with these notations, the "safe" situation in regard to nutritional benefits and dietary chemical contamination is described by the random subset of the consumption space

$$S_Q := \{c \in \mathbb{R}_+^K : \forall h \in \{1,\dots,H\},\ c'Q_j \in [\ell_{\square}(j), \ell^{\square}(j)]\} \tag{4.9}$$

with c' the transpose of vector c, i.e. the set of dietary habits that yield types of exposure respecting all d considered DIL. Conversely, the unsafe consumption zone is denoted by S_Q^c, the complementary of S_Q in $\mathbb{R}_+^H$. The subscript Q emphasizes their dependence on contents $Q_{h;j}$, $1 \leq h \leq H$, $1 \leq j \leq d$, (and body weights for contaminants), which justifies the random nature of both sets. Further set

$$S_{Q_j} := \{c \in \mathbb{R}_+^K : c'Q_j \in [\ell_{\square}(j), \ell^{\square}(j)]\}; \tag{4.10}$$

the safe zone for chemical element j and $S^c_{Q_j}$ its complementary, then according to Equation (4.9), we have $S_Q := \bigcap_{j=1}^{d} S_{Q_j}$ and $S^c_Q := \bigcup_{j=1}^{d} S^c_{Q_j}$. Graphical examples of some possible realizations of S^c_Q are given in Figure 4.1.

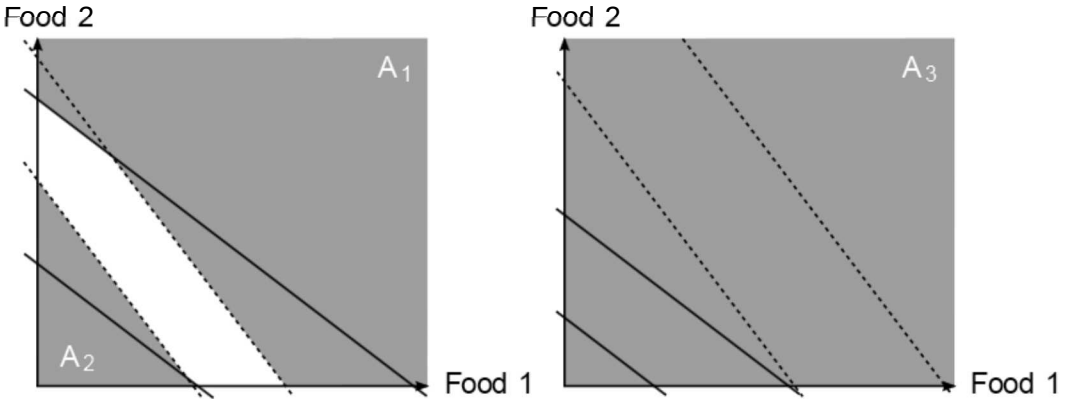


Figure 4.1 – Illustration in dimension 2 of the form of S^c_q for some fixed contamination levels q_1 and q_2 of 2 nutrients. The exterior of the two plain (resp. dotted) lines defines the set $S^c_{q_1}$ (resp. $S^c_{q_2}$). Consequently, S^c_q coincides with $A_1 \cup A_2$ in the left hand graph and with A_3 (the entire space) in the right hand graph.

Again, we assume in this section that we dispose of independent samples: iid consumption vectors $\{C_i := (C_{i,1}, \ldots, C_{i,H}) : 1 \leq i \leq n\}$ are gathered together with a number $m_{h,j}$ of iid measures of content of component j occurring in food item h, namely $\{Q_{h,j} := (Q^1_{h,j}, \ldots, Q^{m_{h,j}}_{h,j})\}$ for $1 \leq h \leq H$, $1 \leq j \leq d$. For simplicity, dietary habits are divided by some large enough constant so that they may fit into the unit cube of $\mathbb{R}^d_+$. The issue mentioned above can be then formulated as follows. Denote by $\mathcal{R}^H$ the set of all subspaces in $[0,1]^H$ that can be written as a finite (with reasonable cardinal, say inferior to some $r \in \mathbb{N}$) union of hyperrectangles, and fix $\alpha \in [0,1]$, a desired level of diet frequency in the studied population. Optimal diets $\mathcal{R}^*$ are then defined as solutions of the optimization program:

$$\mathcal{R}^* := \operatorname*{argmin}_{\mathcal{R} \in \mathcal{R}^H} \varphi(\mathcal{R}) \quad \text{subject to} \quad \mathbb{P}(C \in \mathcal{R}) \geq \alpha, \tag{4.11}$$

where φ is a dietary risk measure. Notice that exploring sub-spaces in the form of hyperrectangles has the advantage of facilitating the eventual communication of the results: it enables interpretations such as "one should eat more or less than this and that amount of specific food". The main question now is to choose an explicit expression for the volume φ. One very natural way of modeling dietary risks would be to define $\varphi(\mathcal{R})$ as the conditional probability $\mathbb{P}\left[C \in S^c_Q \,\middle|\, C \in \mathcal{R}\right]$ for any $\mathcal{R} \in \mathcal{R}^H$. Going back to Equation (4.11), notice that in the case where solutions $\mathcal{R}^*$ are such that $\mathbb{P}(C \in \mathcal{R}^*) = \alpha^*$, it is equivalent to minimizing the joint probability $\mathbb{P}\left[C \in S^c_Q \cap \mathcal{R}\right]$ over the set $\mathcal{R}^H$ of region candidates, subject to $\mathbb{P}(C \in \mathcal{R}) \geq \alpha$. This problem is very similar to that tackled in Section 4.1.1, except that the target criterion one seeks to optimize now is not a (Lebesgue) volume anymore, but an unknown

probability measure. As in Scott and Nowak (2005) and Clémençon and Vayatis (2010), the criterion must then also be replaced by an empirical estimate. Following the approach proposed in Section 4.1.2, we consider the incomplete U-statistic

$$\hat{P}_B\left(C \mid X S_Q^c\right) := \frac{1}{B} \sum_{(i;\ell_{h;j}) \in D_B} \mathbb{I}\{C_i \in u\} \left\{ \prod_{h=1}^{H} C_{i;h} Q_{h;j}^{\ell_{h;j}} R[\ell_{(1)}; \ell_{(1)}] \text{ or } \ldots \right.$$
$$\left. \ldots \text{ or } \prod_{h=1}^{H} C_{i;h} Q_{h;j}^{\ell_{h;j}} R[\ell_{(d)}; \ell_{(d)}] \right\};$$

where D_B is the index set of cardinality B obtained by drawing with replacement B times in the set of indexes $L := \{1; \ldots; n\} \times \pm_{h=1}^{H} \pm_{j=1}^{d} \{1; \ldots; m_{h;j}\}$. This leads to the constrained optimization problem:

$$\min_{R \in \mathcal{R}^H} \hat{P}_B\left(C \mid X S_Q^c\right) \text{ subject to } \hat{P}(C) \geq \alpha - \varepsilon; \tag{4.12}$$

where

$$\hat{P}(C) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{C_i \in u\}. \tag{4.13}$$

The following theorem describes the properties of solutions of the problem in Equation (4.12), involving statistical quantities the computation of which is feasible, the issue of finding such a solution in practice shall be tackled subsequently.

Theorem 4.12 Let $(\alpha; \beta) \in (0;1)^2$. Suppose that the collection $\mathcal{R}^H$ is of finite VC dimension $V < +\infty$ and set:

$$\varepsilon_1(n; \beta) := 4 \sqrt{\frac{2 \log(8) + V \log(n+1) + \log(2/\beta)}{n}}.$$

Then, if R_B is a solution of Equation (4.12), we have with probability at least $1 - \beta$:

$$P\left(C \mid R_B\right) \geq \alpha + 2\varepsilon_1(n; \beta/2) \text{ and } \Phi(R_B) \leq \Phi(R^*) + 2\varepsilon(B; N; \beta/2); \tag{4.14}$$

denoting by Φ the penalty given by Equation (4.7) and by R^* any minimizer of $\Phi(.)$ among the elements R of $\mathcal{R}^H$ such that $P(C) \geq \alpha$.

The (sketch of the) proof is given in Section 4.2. The statistical procedure described in Section 4.1.3 cannot be extended in a straightforward manner, since the quantity $P\left(C \mid S_Q^c\right) \times P(C)$ is far from being constant over the collection of cubes C of fixed side length $1/k$, which paves the supposedly compact support of the distribution of C.

Ú Statistically equivalent hypercubes Our proposal to solve Equation (4.12) approximately is to partition $[0;1]^H$ into a finite number of hypercubes $C_1; \ldots; C_M$ (of variable side length) such that $\hat{P}_B\left(C \mid C_m \times S_Q^c\right)$, $m \in \{1; \ldots; M\}$, remains (approximately) constant, equal to $\omega := M^{-1} \hat{P}_B\left(C \mid S_Q^c\right)$. This can be achieved by means of

a variety of greedy procedures, see section 21.4 in Devroye et al. (1996) and the references therein. Here we use a slightly modified version of the celebrated Gessaman's rule (Gessaman, 1970) as follows. Recall that the quantities involved are built from the sampled dataset $D_B := \{(i^{(b)}; (\ell^{(b)}_{1;1}, \ldots, \ell^{(b)}_{H;d})) : 1 \leq b \leq Bu$ of cardinality B. For all $i \in \{1; \ldots; Bu$, assign to the observation C_i the weight

$$\omega_i := \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left(i = i^{(b)}\right) \left\{ \prod_{h=1}^{H} C_{i^{(b)};h} Q^{\ell^{(b)}_{h;j}}_{h;j} R[\ell_h(1); \ell_h(1)] \text{ or } \ldots \right.$$
$$\left. \ldots \text{ or } \prod_{h=1}^{H} C_{i^{(b)};h} Q^{\ell^{(b)}_{h;j}}_{h;j} R[\ell_h(d); \ell_h(d)] \right\}.$$

Equipped with this notation, observe that $\hat{P}_B(C \mid S^c_Q, C_m) = \sum_{i=1}^{n} \omega_i \mathbb{1}\{C_i \in C_m\}$ for any $m \in \{1; \ldots; M\}$. Starting with all the consumption data, consider the food item h and a "split value" $s \in [0; 1]$ and consider the regions of the consumption space

$$R_1(h; s) := \{c \in [0; 1]^H : c_h \leq s\} \text{ and } R_2(h; s) := \{c \in [0; 1]^H : c_h > s\}.$$

Then find the threshold value s^* that solves

$$\min_{s \in [0;1]} \left[\sum_{i : C_i \in R_1(h;s)} \omega_i \cdot \sum_{i : C_i \in R_2(h;s)} \omega_i \right]$$

over $\{C_{i;h}) : 1 \leq i \leq nu$. For each food item h, the minimization problem in s can be solved very rapidly by scanning through all of the data projected onto the h-th axis. Having thus determined the best "split value", one repeat the binary splitting procedure on both regions in a recursive manner. Let $d_1; \ldots; d_H \geq 1$. Starting from the whole consumption space $[0; 1]^H$, identified as the root node, we propose to build a recursive partition, which can be represented by a binary tree of depth $D = \sum_{h=1}^{H} d_h$, by using first C_1 as "split variable" to build a complete binary tree of depth d_1, then using C_2 to grow the tree until depth $d_1 + d_2$ and continuing in the same way with the remaining food items. Its terminal leaves correspond to the cells $C_1; \ldots; C_{2^D}$ of the data-dependent partition. The collection of such regions, obtained by combining hierarchically $2^D - 1$ splits perpendicular to the coordinate axes is of finite VC dimension $V_D < +1$ (namely, its shatter coefficient for $m \geq 1$ points is classically bounded by $(m + 1)^{2H(2^D - 1)}$). In addition, by construction, the latter are such that $\hat{P}_B(C \mid C_m, S^c_Q)$ is approximately equal to $\hat{P}_B(C \mid S^c_Q) = 2^D$ for $m = 1; \ldots; 2^D$.

Now, from a practical perspective, the target region of the consumption space can be assessed by using a strategy similar to that described in the previous subsection, as follows.

1. Sort the terminal leaves of the tree representing the partition in a way that

$$\hat{p}(C \mid C_{1;2^D}) \geq \ldots \geq \hat{p}(C \mid C_{2^D;2^D}).$$

2. Bind together the cells sequentially, until the empirical estimate of the mass of the resulting set exceeds $\alpha - \beta_D(B, N, \alpha)$, yielding the region:

$$\widehat{R}_{D,\alpha} := \bigcup_{m=1}^{M_\alpha} C_{m,2^D},$$

where

$$M_\alpha := \min\left\{ M \geq 1 : \sum_{m=1}^{M} \widehat{P}(C_{m,2^D}) \geq \alpha - \beta_D(B, N, \alpha) \right\}.$$

Beyond its computational efficacy and simplicity, a crucial advantage of the approach described above lies in its capacity to produce regions which can be visually summarized by a binary tree, the terminal leaves of which can be described by combining elementary rules of the form "$c_h \leq s$" or "$c_h > s$" in a hierarchical manner. This point is of major importance when designing dietary guidelines to improve nutrition over the population of interest.

4.2 proofs and supplements

4.2.1 Maximal deviation

We start by establishing the following intermediary result, which extends Corollary 3 in Clémençon et al. (2008) to the K-sample situation.

Lemma 4.13 Suppose that the hypotheses in Theorem 4.8 are fulfilled. For all $\delta \in (0,1)$, we have with probability at least $1 - \delta$,

$$\sup_{P \in \mathcal{P}} |U_n(P) - \alpha(P)| \leq M_\alpha \left(2 \sqrt{\frac{2V \log(1+K)}{n}} + \frac{\log(1/\delta)}{n} \right).$$

Proof Set $n = \min\{n_1/d_1, \ldots, n_K/d_K\}$ and let

$$V\left(X_1^{(1)}, \ldots, X_{n_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{n_K}^{(K)}\right) := \frac{1}{n} \left(\left(X_1^{(1)}, \ldots, X_{d_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{d_K}^{(K)}\right) + \left(X_{d_1+1}^{(1)}, \ldots, X_{2d_1}^{(1)}, \ldots, X_{d_K+1}^{(K)}, \ldots, X_{2d_K}^{(K)}\right) + \cdots + \left(X_{(n-1)d_1+1}^{(1)}, \ldots, X_{nd_1}^{(1)}, \ldots, X_{(n-1)d_K+1}^{(K)}, \ldots, X_{nd_K}^{(K)}\right) \right);$$

for any $P \in \mathcal{P}$. Recall that the K-sample U-statistic $U_n(P)$ can be expressed as

$$U_n(P) = \frac{1}{n_1! \cdots n_K!} \sum_{\substack{\sigma_1 \in \mathcal{S}_{n_1} \\ \vdots \\ \sigma_K \in \mathcal{S}_{n_K}}} V\left(X_{\sigma_1(1)}^{(1)}, \ldots, X_{\sigma_1(n_1)}^{(1)}, \ldots, X_{\sigma_K(1)}^{(K)}, \ldots, X_{\sigma_K(n_K)}^{(K)}\right);$$

where S_m denotes the symmetric group of order m for any $m \geq 1$. This representation as an average of sums of n independent terms is known as the (first) Hoeffding's decomposition (Hoeffding, 1948). Then, using Jensen's inequality in particular, one may easily show that, for any nondecreasing convex function $G : \mathbb{R}_+ \to \mathbb{R}$, we have:

$$\mathbb{E}\left[G\left(\sup_{P\in\mathcal{P}} |U_n(P) - \bar{P}|\right)\right] \leq \mathbb{E}\left[G\left(\sup_{P\in\mathcal{P}} |V - \mathbb{E}(V)|\right)\right], \quad (4.15)$$

where we set $\bar{P} := \int P(x)\,dx$ for all $P \in \mathcal{P}$. Now, using standard symmetrization and randomization arguments (see Giné and Zinn (1984) for instance) and Equation (4.15), we obtain that

$$\mathbb{E}\left[G\left(\sup_{P\in\mathcal{P}} |U_n(P) - \bar{P}|\right)\right] \leq \mathbb{E}\left(G(2R_n)\right), \quad (4.16)$$

where

$$R_n := \sup_{P\in\mathcal{P}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i P\left(X_1^{(1)}, \ldots, X_{d_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{d_K}^{(K)}\right),$$

is a Rademacher average based on the Rademacher chaos $\varepsilon_1, \ldots, \varepsilon_n$ (independent random symmetric sign variables), independent from the $X_i^{(k)}$'s. We now apply the bounded difference inequality (McDiarmid, 1989) to the functional R_n, seen as a function of the iid random variables $(\varepsilon_i, X_1^{(1)}, \ldots, X_{d_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{d_K}^{(K)})$, $1 \leq i \leq n$: changing any of these random variables changes the value of R_n by at most M_n/n. One thus obtains from Equation (4.16) with $G(x) = \exp(\lambda x)$, where $\lambda > 0$ is a parameter which shall be chosen later, that:

$$\mathbb{E}\left[\exp\left(\lambda \sup_{P\in\mathcal{P}} |U_n(P) - \bar{P}|\right)\right] \leq \exp\left(2\lambda \mathbb{E}(R_n) + \frac{M_n^2 \lambda^2}{4n}\right).$$

Applying Chernoff's method, one then gets:

$$\mathbb{P}\left[\sup_{P\in\mathcal{P}} |U_n(P) - \bar{P}| > t\right] \leq \exp\left(-\lambda t + 2\lambda \mathbb{E}(R_n) + \frac{M_n^2 \lambda^2}{4n}\right). \quad (4.17)$$

Using the bound (see Equation (6) in Boucheron et al. (2005) for instance)

$$\mathbb{E}(R_n) \leq M_n \sqrt{\frac{2V \log(1+n)}{n}}$$

and taking $\lambda = 2n(t - 2\mathbb{E}(R_n))/M_n^2$ in Equation (4.17), one finally establishes the desired result. □

Now we shall prove Theorem 4.8, the statement of which is recalled below.

Theorem – Maximal deviation. Let $\mathcal{F}$ be a collection of bounded symmetric kernels on $\mathcal{X} := \prod_{k=1}^{K} \mathbb{X}_k^{d_k}$. Suppose that $\mathcal{F}$ is a VC major class of functions with finite Vapnik-Chervonenkis dimension V and that $M_{\mathcal{F}} := \sup_{(\varphi;x)\in\mathcal{F}\times\mathcal{X}} |\varphi(x)| \vee 1$. Then, the following assertions hold true.

i) For all $\varepsilon > 0$, we have: $\forall n = (n_1,\ldots,n_K) \in \mathbb{N}^K$, $\forall B \geq 1$,

$$\mathbb{P}\left[\sup_{\varphi\in\mathcal{F}} \left|U_B(\varphi) - U_n(\varphi)\right| > \varepsilon\right] \leq 2(1+\#\mathcal{L})^V e^{-B\varepsilon^2/M_{\mathcal{F}}^2}.$$

ii) For all $\alpha \in (0,1)$, with probability at least $1-\alpha$, we have: $\forall n_k \geq 1$, $1 \leq k \leq K$,

$$\sup_{\varphi\in\mathcal{F}} \left|U_B(\varphi) - \mathbb{E}\left[U_B(\varphi)\right]\right| \leq M_{\mathcal{F}} \left(c\sqrt{\frac{2V\log(1+\#\mathcal{L})}{\#\mathcal{L}}} + c\sqrt{\frac{\log(2/\alpha)}{\#\mathcal{L}}} + c\frac{V\log(1+\#\mathcal{L}) + \log(4/\alpha)}{B}\right);$$

where

$$\#\mathcal{L} := \min\{\lfloor n_1/d_1 \rfloor, \ldots, \lfloor n_K/d_K \rfloor\}$$

and $\lfloor x \rfloor$ denotes the integer part of any real number x.

Proof For convenience, we introduce the random sequence $\mathcal{W} := ((\omega_b(l))_{l\in\mathcal{L}})_{1\leq b\leq B}$, where $\omega_b(l)$ is equal to 1 if the tuple $l := (l_1,\ldots,l_K)$ has been selected at the b-th draw and to 0 otherwise: the ω_b's are iid random vectors and, for all $(b;l)$ in $\{1,\ldots,B\}\times\mathcal{L}$, the random variable $\omega_b(l)$ has a Bernoulli distribution with parameter $1/\#\mathcal{L}$. We also set $X_l := (X_{l_1}^{(1)},\ldots,X_{l_K}^{(K)})$ for any l in $\mathcal{L}$. Equipped with these notations, observe first that one may write: $\forall B \geq 1$, $\forall \varphi \in \mathcal{F}^K$,

$$U_B(\varphi) - U_n(\varphi) = \frac{1}{B}\sum_{b=1}^{B} Z_b(\varphi);$$

where $Z_b(\varphi) := \sum_{l\in\mathcal{L}} (\omega_b(l) - 1/\#\mathcal{L})\, \varphi(X_l)$ for any $(b;l) \in \{1,\ldots,B\}\times\mathcal{L}$. It follows from the independence between the X_l's and the $\omega(l)$'s that, for all $\varphi \in \mathcal{F}$, conditioned upon the X_l's, the variables $Z_1(\varphi),\ldots,Z_B(\varphi)$ are independent, centered and almost-surely bounded by $2M_{\mathcal{F}}$ (notice that $\sum_{l\in\mathcal{L}} \omega_b(l) = 1$ for all $b \geq 1$). By virtue of Sauer's lemma, since $\mathcal{F}$ is a VC major class with finite VC dimension V, we have, for fixed X_l's:

$$\#\{(\varphi(X_l))_{l\in\mathcal{L}} : \varphi \in \mathcal{F}\} \leq (1+\#\mathcal{L})^V.$$

Hence, conditioned upon the X_l's, using the union bound and next Hoeffding's inequality applied to the independent sequence $Z_1(\varphi),\ldots,Z_B(\varphi)$, for all $\varepsilon > 0$, we obtain that:

$$\mathbb{P}\left[\sup_{\varphi\in\mathcal{F}} \left|U_B(\varphi) - U_n(\varphi)\right| > \varepsilon \,\middle|\, (X_l)_{l\in\mathcal{L}}\right] \leq \mathbb{P}\left[\sup_{\varphi\in\mathcal{F}} \left|\frac{1}{B}\sum_{b=1}^{B} Z_b(\varphi)\right| > \varepsilon \,\middle|\, (X_l)_{l\in\mathcal{L}}\right]$$

$$\leq 2(1+\#L)^{V} e^{-B\varepsilon^2/M_\gamma^2};$$

which proves the first assertion of the theorem. Notice that this can be formulated: for any $\alpha \in (0,1)$, we have with probability at least $1-\alpha$:

$$\sup_{P\in\mathcal{P}} \left| U_B(\cdot) - U_n(\cdot) \right| \leq M_\gamma \sqrt{\frac{V\log(1+\#L) + \log(2/\alpha)}{B}}. \tag{4.18}$$

Turning to the second part of the theorem, it straightforwardly results from the first part combined with Lemma 4.13. □

4.2.2 Maximal deviation in dietary risk analysis

We shall prove Corollary 4.11, the statement of which is recalled below.

Corollary For all $\alpha \in (0,1)$, we have with probability at least $1-\alpha$:

$$L\left(\hat{R}_\alpha\right) \leq \inf_{R\in\mathcal{R}:\, P(X\in R)\geq\alpha} L(R) \text{ and } P\left(X \in \hat{R}_\alpha\right) \geq \alpha - 2\kappa(B,N,\alpha).$$

Proof Observe first that the assumptions of Theorem 4.8 are fulfilled when taking $\mathcal{G} = \{\mathbb{1}_R : R \in \mathcal{R}\}$: the collection $\mathcal{R}$ of indicator functions is a VC major class of functions with finite VC dimension V and $M_\mathcal{G}^2 = 1$. Applying thus Theorem 4.8, the proof is derived by following line by line the argument of Corollary 6 in Scott and Nowak (2006). Details are left to the reader. □

4.2.3 Optimal dietary habits

We shall prove Theorem 4.12, the statement of which is recalled below.

Theorem Let $(\alpha,\gamma) \in (0,1)^2$. Suppose that the collection $\mathcal{R}^H$ is of finite VC dimension $V < +\infty$ and set:

$$\delta_1(n,\alpha) := 4\sqrt{\frac{2\log(8) + V\log(n+1) + \log(2/\alpha)}{n}}.$$

Then, if $\hat{r}_B$ is a solution of Equation (4.12), we have with probability at least $1-\alpha$:

$$P\left(C_P\left(\hat{r}_B\right) \geq \gamma + 2\delta_1(n,\alpha/2)\right) \text{ and } \Phi(\hat{r}_B) \leq \Phi(r^*) + 2\kappa(B,N,\alpha/2); \tag{4.19}$$

denoting by Φ the penalty given by Equation (4.7) and by r^* any minimizer of $\Phi(.)$ among the elements r of $\mathcal{R}^H$ such that $P(C_P(r)) \geq \gamma$.

Proof The result immediately follows from the argument of Theorem 10 in Clémençon and Vayatis (2010) combined with Theorem 4.8 and Vapnik-Chervonenkis inequality (see Theorem 12.5 in Devroye et al., 1996 for instance) to control the deviations of the supremum:

$$\sup_{P \in \mathcal{P}^H} \left| \hat{P}(C_{P,\alpha}) - P(C_{P,\alpha}) \right|.$$

□

5

EMPIRICAL PROCESSES IN SURVEY SAMPLING

Like INCA2, most consumption databases are now constructed with some survey design to produce representative samples. Formally, this means that in the population of interest, the probability that an individual may be selected is taken into account in the form of a survey weight. For institutional data, these weights often correspond to the so-called true inclusion probabilities, but statisticians may sometimes have at their disposal calibrated or post-stratification weights (e.g. minimizing some discrepancy with the inclusion probabilities subject to some margin constraints). In most cases, the survey scheme is ignored, potentially yielding a significant sampling bias. When considering some functional of the empirical process such as the empirical distribution function, this may cause severe drawbacks and completely jeopardize the estimation, as can be revealed by simulation experiments. In the context of dietary risk analysis, the impact of such an omission would be for instance an erroneous, greatly biased estimation of the true proportion of over-exposed people. To avoid such undesirable outcomes, many estimators have been developed in the branch of survey sampling theory (Tillé, 2006; Gourieroux, 1981; Droesbeke et al., 1987), which take into account these survey weights and make up for the induced bias of the sampling phase. Unfortunately, to our knowledge, except in the specific case of stratified sampling, there is still no general functional result that would guarantee the asymptotic normality of a large family of estimators in the context of survey sampling. In particular, when estimating the distribution function of the exposure to some food chemical, the construction of confidence bands, as opposed to point confidence intervals, has not been made possible yet. We started addressing this issue in the paper presented from Section 5.1 to Section 5.4, in the specific case of Poisson-like survey plans with no post-calibration. It is the result of a collaboration with P. Bertail (Université Paris X, France) and S. Clémençon (Télécom ParisTech, France) and has been submitted for publication. Unfortunately, the results that are presented there do not apply to the complex sampling of INCA2; they have to be understood as a first step towards the elaboration of a more general theory that would encompass a wider range of survey techniques.

The main goal of this chapter is to investigate how to incorporate the survey scheme into the inference procedure dedicated to the estimation of a probability measure P on a measurable space (viewed as a linear operator acting on a certain class of functions F), in order to guarantee its asymptotic normality. This problem

has been addressed by Breslow and Wellner (2007) and Saegusa and Wellner (2011) in the particular case of a stratified survey sampling, where individuals are selected at random (without replacement) in each stratum, by means of bootstrap limit results. Our approach is different and follows that of Hàjek (1964), extended next by Berger (1998, 2011), and is applicable to more general sampling surveys, namely those with unequal first order inclusion probabilities which are of the Poisson type or sequential/ rejective. The main result of the chapter is a Functional Central Limit Theorem (FCLT) describing the limit behavior of an adequate version of the empirical process (referred to as the Horvitz-Thompson empirical process throughout the article) in a superpopulation statistical framework. The key argument involved in this asymptotic analysis consists in approximating the distribution of the extended empirical process by that related to a much simpler sampling plan. In order to illustrate the reach of this result, statistical applications are considered, where the extensions of the empirical process are used to construct confidence bands around the Horvitz-Thompson estimator of the cumulative distribution function.

The chapter is organized as follows. In Section 5.1 and Section 5.2, the statistical framework is described at length, notations are set out and some basics on survey sampling theory are recalled, together with important examples of survey schemes to which the subsequent asymptotic analysis can be applied. The main result of the chapter, a FCLT for the Horvitz-Thompson empirical process, is stated in Section 5.3, while applications of the latter to non-parametric functional estimation are displayed in Section 5.4. Finally, technical details are deferred to Section 5.5.

5.1 background and preliminaries

We start off with recalling some crucial notions in survey sampling and in modern empirical process theory, which shall be extensively used in the subsequent analysis. Throughout the article, the Dirac mass at x in some vector space X is denoted by δ_x and the indicator function of any event E by $\mathbb{1}\{E\}$. We also denote by $\#E$ the cardinality of any finite set E, and by $\mathcal{P}(E)$ its power set.

5.1.1 Survey sampling: some basics

The purpose of survey sampling is to study some characteristics of a population U_N of $N \geq 1$ units (or individuals) identified by an arbitrary collection of labels: $U_N := \{1, \ldots, N\}$. For various reasons (limited budget, geographical constraints, etc.), it is usually not possible to reach the whole population, and the features of interest have to be estimated from a finite, relatively small number of its elements, namely a sample $s := \{i_1, \ldots, i_{n(s)}\} \subset U_N$ of size $n(s) \ll N$. So as to provide handy ways of controlling the accuracy of estimation, sample units are picked randomly among

U_N (see for instance Tillé, 2006, Chapter 1, Tillé, 1999 or Gourieroux, 1981 for an introduction to the origins of random sampling). Equipped with this representation, a sampling scheme (design/ plan) is determined by a discrete probability measure R_N on $P(U_N)$, the set of all possible samples in U_N. Depending on the adopted point of view, like in superpopulation models, the characteristics of the population can be considered random too. In the next paragraphs, while introducing crucial concepts and notations, we shall discuss both sources of hazard and their classical modeling in survey sampling theory.

5.1.1.1 Survey schemes without replacement

Consider a sampling scheme R_N where individuals are only selected once, i.e. a design without replacement. Our analysis is restricted to this popular family of survey plans. By definition, the two conditions below are always fulfilled,

1. $\forall s \in P(U_N), R_N(s) \geq 0$,
2. $\sum_{s \in P(U_N)} R_N(s) = 1$,

and the mean survey sample size is given by

$$E_{R_N}(n(S)) = \sum_{s \in P(U_N)} n(s) R_N(s).$$

Here, the notation $E_{R_N}(.)$ denotes the expectation taken with respect to the random sample S with distribution R_N. In a similar fashion, $P_{R_N}(S \in \mathcal{S})$ refers to the probability of the event $\{S \in \mathcal{S}\}$ with $\mathcal{S} \subset P(U_N)$, when S is drawn from R_N. In particular, $R_N(s) = P_{R_N}(S = s)$. Such distributions are entirely characterized by the concepts listed below.

Inclusion probabilities For any $i \in U_N$, the quantity usually referred to as the i-th (first order) inclusion probability,

$$\pi_i(R_N) := P_{R_N}(i \in S) = \sum_{s \in P(U_N)} R_N(s) \mathbb{I}_{\{i \in s\}},$$

is the probability that the individual labeled i belongs to a random sample S under the survey scheme R_N. When there is no ambiguity on the sampling design, we will simplify notations and write π_i instead of $\pi_i(R_N)$. In the subsequent analysis, first order inclusion probabilities are assumed to be strictly positive: $\forall i \in U_N, \pi_i(R_N) > 0$. We shall even require the stronger hypothesis that they never get either too small or too large, as formally stated below.

Assumption 5.1 There exist $\alpha > 0$ and $N_0 \in \mathbb{N}^*$ such that for all $N \geq N_0$ and $i \in U_N$,

$$\pi_i(R_N) > \alpha.$$

In addition,

$$\limsup_{N \to +\infty} \frac{1}{N} \prod_{i=1}^{N} \pi_i(R_N) < 1.$$

When the first condition holds, the rate of convergence of the estimators considered in Section 5.2 and Section 5.3 will be shown to be typically of order $1/\sqrt{N}$. One could possibly relax it and allow $\pi_\cdot$ to depend on N, with $\pi_\cdot = \pi_\cdot(N)$ decaying to zero as N tends to infinity at a specific rate, and still be able to establish limit results. The analysis would be however much more technical; this is left for further research.

Conditions involving the second order inclusion probabilities shall also be used in our asymptotic analysis. They are denoted by

$$\pi_{i,j}(R_N) := P_{R_N}\left[(i,j) \in S^2\right] = \sum_{s \in \mathcal{P}(U_N)} R_N(s)\, \mathbb{I}\{i,j \in s\},$$

for all $(i,j) \in U_N^2$. In other words, $\pi_{i,j}(R_N)$ is the probability that two distinct individuals labeled i and j are jointly selected under design R_N. Again, we may eventually write $\pi_{i,j}$ when there is no need to emphasize the dependency on the sampling plan R_N. Notice that higher order inclusion probabilities may be defined in a similar way, up to the maximal order for which the entire population is selected.

Inclusion indicators The information related to the observed sample S is encapsulated by the random vector $\mathcal{I}_{(N)} := (\mathcal{I}_1, \ldots, \mathcal{I}_N)$, where

$$\mathcal{I}_i := \mathbb{I}\{i \in S\} = \begin{cases} 1 & \text{with probability } \pi_i, \\ 0 & \text{with probability } 1 - \pi_i. \end{cases}$$

Notice indeed that the set $\mathcal{P}(U_N)$ of all possible samples is in one-to-one correspondence with $\{0,1\}^N$, which provides a handy alternative representation of sampling schemes. Again, for simplicity, we will omit the subscript (N) when no ambiguity is possible. By definition, the distribution of $\mathcal{I} := \mathcal{I}_{(N)}$ has univariate marginals that correspond to the Bernoulli distributions $\mathcal{B}(\pi_i)$, $i \in U_N$, and covariance matrix given by

$$\Gamma_N := \left(\pi_{i,j} - \pi_i \pi_j\right)_{1 \leq i,j \leq N}.$$

Incidentally we have $\sum_{i=1}^{N} \mathcal{I}_i = n(S)$ and thus $\sum_{i=1}^{N} \pi_i = E_{R_N}(n(S))$.

Before considering the issue of extending the concept of empirical process in the context of survey sampling, we recall a few important classes of survey schemes, to which the results established in Section 5.2 and Section 5.3 can be applied. One may refer to Deville (1987) for instance for an excellent account of survey theory, including many more examples of sampling designs.

Example 5.2 – Simple Random Sampling Without Replacement. A simple random sampling without replacement (SRSWOR in abbreviated form) is a sampling design of fixed size $n(S) = n$, according to which all samples with cardinality n in the population U_N are equally likely to be chosen, with probability $(N - n)!/n!$. It follows that all units of U_N have the same chance of being selected, n/N namely, and all second order probabilities are equal to $n(n-1)/(N(N-1))$.

Example 5.3 – Poisson survey sampling. The Poisson sampling plan without replacement (POISSWOR), denoted here by T_N, is one of the simplest survey schemes. In this case, the N elements of Ω are independent Bernoulli random variables with respective parameters $\pi_i(T_N) =: p_i$, $i \in \{1, \ldots, N\}$ so that for any sample $s \in \mathcal{P}(U_N)$,

$$T_N(s) = \prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i).$$

Notice that the size $n(S)$ of sample S with distribution T_N is random (except in the sole situation where $p_i \in \{0, 1\}$ for $i = 1, \ldots, N$) and that the corresponding survey plan is fully characterized by the first order inclusion probabilities. In the specific situation where they are all equal, i.e. $p_1 = \cdots = p_N = p$, the design is called Bernoulli.

Example 5.4 – Stratified sampling. A stratified sampling design permits to draw a sample S of fixed size $n(S) = n \leq N$ within a population U_N that can be partitioned into $K \geq 1$ distinct strata $U_{N_1}, \ldots, U_{N_K}$ (known a priori) of respective sizes $N_1, \ldots, N_K$ adding up to N. Let $n_1, \ldots, n_K$ be non-negative integers such that $n_1 + \cdots + n_K = n$, then the drawing procedure is implemented in K steps: within each stratum U_{N_k}, $k \in \{1, \ldots, K\}$, perform a SRSWOR of size $n_k \leq N_k$ yielding a sample S_k. The final sample is obtained by assembling these sub-samples: $S = \bigcup_{k=1}^{K} S_k$. The probability of drawing a specific sample s by means of this survey scheme is

$$R_N^{str}(s) = \prod_{k=1}^{K} \binom{N_k}{n_k}^{-1}.$$

Naturally, first and second order inclusion probabilities depend on the stratum to which each unit belong: for all $i \neq j$ in U_N,

$$\pi_i(R_N^{str}) = \prod_{k=1}^{K} \frac{n_k}{N_k} \mathbb{I}_{i \in U_{N_k}} \quad \text{and} \quad \pi_{i,j}(R_N^{str}) = \prod_{k=1}^{K} \frac{n_k(n_k-1)}{N_k(N_k-1)} \mathbb{I}_{(i,j) \in U_{N_k}^2}.$$

Example 5.5 – Canonical Rejective Sampling. Let $n \leq N$ and consider a vector $\pi^R := (\pi_1^R, \ldots, \pi_N^R)$ of first order inclusion probabilities. Further define $S_n := \{s \in \mathcal{P}(U_N) : \#s = n\}$, the set of all samples in population U_N with car-

dinality n. The rejective sampling (Hàjek, 1964; Berger, 1998), sometimes called conditional Poisson sampling (CPS), exponential design without replacement or maximum entropy design (Tillé, 2006, Section 5.6), is the sampling design R_N^R that selects samples of fixed size $n(s) = n$ so as to maximize the entropy measure

$$H(R_N) = \sum_{s \in S_n} R_N(s) \log R_N(s);$$

subject to the constraint that its vector of first order inclusion probabilities coincides with π^R. It is easily implemented in two steps.

1. Draw a sample S with a POISSWOR plan $T_N = T_N^p$, with properly chosen first order inclusion probabilities vector $p := (p_1, \ldots, p_N)$. The representation is called canonical if $\sum_{i=1}^N p_i = n$. In that case, relationships between each p_i and π_i^R, $1 \leq i \leq N$, are established in Hàjek (1964).
2. If $n(S) \neq n$, then reject sample S and go back to step one, otherwise stop.

Vector p must be chosen in a way that the resulting first order inclusion probabilities coincide with π^R, by means of a dedicated optimization algorithm (Tillé, 2006, Algorithms 5.5 to 5.9). The corresponding probability distribution is given for all $s \in P(U_N)$ by

$$R_N^R(s) = \frac{T_N^p(s)\, \mathbb{I}\{\#s = n\}}{\sum_{s' \in S_n} T_N^p(s')} \propto \prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i)\, \mathbb{I}\{\#s = n\};$$

where $\propto$ denotes the proportionality. We refer to Hàjek (1964, p.1496) for more details on the links between rejective and Poisson sampling plans.

Example 5.6 – Rao-Sampford Sampling. The Rao-Sampford sampling design generates samples $s \in P(U_N)$ of fixed size $n(s) = n$ with respect to some given first order inclusion probabilities $\pi^{RS} := (\pi_1^{RS}, \ldots, \pi_N^{RS})$, fulfilling $\sum_{i=1}^N \pi_i^{RS} = n$, with probability

$$R_N^{RS}(s) = \alpha \prod_{i \in s} \pi_i^{RS} \prod_{j \notin s} \frac{\pi_j^{RS}}{1 - \pi_j^{RS}}.$$

Here, $\alpha > 0$ is chosen such that $\sum_{s \in P(U_N)} R_N^{RS}(s) = 1$. In practice, the following algorithm is often used to implement such a design (Berger, 1998):

1. select the first unit i with probability π_i^{RS}/n,
2. select the remaining $n - 1$ units j with drawing probabilities proportional to $\pi_j^{RS}/(1 - \pi_j^{RS})$, $j = 1, \ldots, N$,
3. accept the sample if the units drawn are all distinct, otherwise reject it and go back to step one.

5.1.1.2 Superpopulation models

The characteristics of interest in population U_N are modeled as follows. We consider the probability space $(U_N; \mathcal{P}(U_N); P)$ and a random variable/ vector X defined on the latter, taking its values in a Banach space $(\mathbb{X}; \|.\|)$, with probability measure P. We set

$$X : \begin{array}{ccc} U_N & \longrightarrow & \mathbb{X} \\ i & \longmapsto & X(i) =: X_i \end{array};$$

and the σ-algebra induced by the normed vector space topology structure of $\mathbb{X}$ is denoted by $\mathcal{A}$. For instance, X could represent the amounts of K food products consumed on a specific day. In that case, we would have $\mathbb{X} = \mathbb{R}_+^K$, $\mathcal{A}$ the associated Borel algebra, $\|.\|$ the euclidean norm and $X_i = (X_{i;1}; \ldots; X_{i;K})$ would give the daily intakes of individual $i \in U_N$. Then, the studied features correspond to some synoptic mapping $(X_1; \ldots; X_N) \mapsto f(X_1; \ldots; X_N)$. In our example, we could consider $f(X_1; \ldots; X_N) = (N^{-1} \sum_{i=1}^N X_{i;1}; \ldots; N^{-1} \sum_{i=1}^N X_{i;K})$, the average consumption of the K foods in U_N.

In survey sampling, a superpopulation is basically an imaginary infinite population, U_1 say, from which U_N is supposed to be issued. In a model-based approach, it is assumed that the random vectors of interest $X_1; \ldots; X_N$ are in fact realizations of N random vectors $X_j : U_1 \longrightarrow \mathbb{X}$, $1 \leq j \leq N$, with joint distribution Q. Then, a superpopulation model is simply a set of conditions that characterize Q (Droesbeke et al., 1987, Chapter 4). The main advantage of such a framework is that it often facilitates statistical inference; in particular, it permits the development of an asymptotic theory, when sample and population sizes grow conjointly to infinity. The superpopulation model we consider here stipulates that all N random vectors X_i, $i \in U_N$, are independent and identically distributed (iid) with common distribution P, i.e. $Q = P^{\otimes N}$, where $\otimes$ denotes the tensor product of measures.

Remark 5.7 The most celebrated iid superpopulation model that we adopt here establishes a setting very similar to that of weighted bootstrap (Arcones and Giné, 1992; Barbe and Bertail, 1995): the original iid N-sample there would correspond to the complete vector $(X_1; \ldots; X_N)$, from which sub-samples are drawn according to some procedure likened to the survey scheme. Actually, both approaches are completely equivalent if the survey weights $(\pi_1/\pi_1; \ldots; \pi_N/\pi_N)$ are exchangeable (i.e. the N-variate distribution of this vector is invariant to the order of its elements). For instance, in the specific case of stratified sampling, drawing units with equal probabilities in each stratum (with a finite and given stratum-size) amounts to bootstrapping (without replacement) in some given cell. It is not surprising then that both Breslow and Wellner (2007) and Saegusa and Wellner (2011) construct a general asymptotic theory in two-phase sampling by using bootstrap type results.

5.1.1.3 Auxiliary information

In practice, sampling from a population U_N is only possible if all individuals are listed somehow, and can be identified once selected. Such documents are called survey frames; in the case of social surveys, they are collected by government institutions and often provide some minimal information about its components. For instance, in France, the geographical situation, the age, the genre and the profession and socio-professional category ("Profession et Catégorie Socioprofessionnelle", PCS, in French) of citizens are displayed in files managed by INSEE (Institut National de la Statistique et des Etudes Economiques). These auxiliary variables, supposedly known for all $i \in U_N$, can sometimes be used to optimize in some sense the survey scheme. In a superpopulation framework, we denote by W the auxiliary random vector, valued in some measurable space $\mathcal{W}$, and set $W_{(N)} := (W_1, \ldots, W_N)$. Again, the subscript (N) shall be dropped when no confusion is possible. As soon as W is correlated with X, the vector of interest, it becomes possible to boost the efficiency of estimators by defining inclusion probabilities as a function of $W_{(N)}$ (Droesbeke et al., 1987).

In the present analysis, we denote by $P_{X,W}$ the joint distribution of (X, W) and by P_W the marginal distribution of W. Like in most applications, we assume that the W_i's are independent (or exchangeable) random variables/ vectors, linked to the variable of interest X through a linear model (notice that W may be constant over the population). It is required though that W is not proportional to X (in a deterministic sense) to avoid degenerate situations; in such a case, knowing W on the whole population would mean knowing the empirical process without any error. For the sake of simplicity, the dependence of survey weights in W will only be emphasized when it is necessary, starting in Section 5.3.

5.1.2 Empirical process indexed by classes of functions

In the context of iid realizations $X_1, \ldots, X_N$ of a probability measure P, empirical process theory (Ledoux and Talagrand, 1991) consists in the study of the fluctuations of random processes of the type

$$\{G_N f, f \in \mathcal{F}\}, \text{ where } G_N := P_N - P.$$

There, class $\mathcal{F}$ designates a certain set of P-integrable real-valued functions,

$$P_N := \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i}$$

is the "classical" empirical measure, and for any signed measure Q on a measurable space $(\mathcal{X}, \mathcal{A})$, $Qf := \int_{\mathcal{X}} f(x)\, Q(dx)$ when the integral is well-defined. We assume that class $\mathcal{F}$ admits a square integrable envelope H as defined below.

Assumption 5.8 There exists a measurable function $H : X \to R$ such that $\int_X H^2(x)\,P(dx) < \infty$ and $|f(x)| \leq H(x)$ for all $x \in X$ and any $f \in F$.

As a consequence, F is a subset of the space

$$L_2(P) := \left\{ h : X \to R;\ h \text{ measurable and } \|h\|_{2;P}^2 := E_P\left[h^2(X)\right] < +\infty \right\}.$$

Notice that we may assume without loss of generality that there exists $\delta > 0$ such that $H(x) > \delta$ for every $x \in X$, even if it entails replacing H by $H + \delta$ in the condition above.

5.1.2.1 Donsker classes

When viewed as a linear operator acting on F, a probability measure P satisfying Assumption 5.8 may be considered as an element of $\ell^\infty(F)$, i.e. the space of all maps $\Phi : F \to R$ such that

$$\|\Phi\|_F := \sup_{f \in F} |\Phi(f)| < +\infty;$$

equipped with the uniform convergence norm (or, equivalently, with Zolotarev metric), namely

$$\|P - Q\|_F := d_F(P;Q) = \sup_{h \in F} \left| \int h\,dP - \int h\,dQ \right|;$$

for any couple of probability measures P and Q. The main purpose of empirical process theory is to find conditions on the class of functions F guaranteeing that the distribution of $\sqrt{N}\,G_N$ converges, as $N \to +\infty$, to that of a Gaussian, Banach space valued process in $\ell^\infty(F)$. Such collections of functions are called Donsker classes by analogy to the classical results on the empirical distribution function that analyze $\sqrt{N}(F_N - F)$, where

$$F_N(x) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left\{X_i \in (-\infty, x_1] \times \cdots \times (-\infty, x_d]\right\}$$

and

$$F(x) := P\left(X \in (-\infty, x_1] \times \cdots \times (-\infty, x_d]\right)$$

for $x := (x_1, \ldots, x_d) \in R^d$ (see Example 5.9). In particular, the study of the uniform deviations over F

$$\sqrt{N}\,\|P_N - P\|_F$$

is of great interest, with a variety of applications in statistics, see Shorack and Wellner (1986). A nearly exhaustive review of asymptotic results ensuring that F is a Donsker class of functions is available in van der Vaart and Wellner (1996). The purpose of this chapter is to extend typical empirical processes results obtained for iid data to the framework of survey sampling.

5.1.2.2 On measurability issues

Recall that the normed vector space $(\ell^\infty(F); \|.\|_F)$ is (generally) a non-separable Banach space. The major problem one faces when dealing with sums of random variables taking their values in such an infinite-dimensional non-separable space concerns the measurability of events. For instance, the "classical" empirical process $\sqrt{N}(F_N - F)$, which can be viewed as a random sequence in the Skorokhod space $D([0;1])$ of càd-làg functions endowed with the supremum norm, is not Borel-measurable. In this specific case, the topology induced by the sup-norm on $D([0;1])$ can be classically replaced by the Skorokhod metric in order to overcome this technical difficulty. Alternative approaches can be found in Pollard (1984). The ideas developed in Hoffmann-Jørgensen (1991) have led to a general solution, based on the concept of outer probability, extending the original probability measure P to non-measurable events by setting $P^*(A) := \inf\{P(B) : A \subset B; B \text{ measurable}\}$. Then, the related concept of Hoffman-Jørgensen weak convergence permits somehow to forget the measurability assumptions. Hence, expectations and probabilities must now be understood as outer expectations and probabilities for non-measurable events. For simplicity, the same notations are kept to denote original and outer probabilities (resp. expectations). Here, weak convergence is metrized through the bounded Lipchitz metric on the space $\ell^\infty(F)$: for all random functions X and Y valued in $\ell^\infty(F)$,

$$d_{BL}(X;Y) = \sup_{b \in BL_1(\ell^\infty(F))} \left| E^*(b(X)) - E^*(b(Y)) \right|,$$

where $BL_1(\ell^\infty(F))$ is the set of all 1-Lipchitz functions on $\ell^\infty(F)$ bounded by 1. In the following we define the ρ_P semi-metric under P as

$$\rho_P(f;g) := E_P\left[(f(X) - g(X))^2\right] =: \|f - g\|_{2;P}^2.$$

We refer to van der Vaart and Wellner (1996) for technical details and general results.

5.1.2.3 Uniform covering numbers

A key concept in the study of empirical process is the covering number $N(\varepsilon; F; |.|)$, which corresponds to the minimal number of balls of radius $\varepsilon > 0$ for a given semi metric $|.|$ needed to cover F. Donsker classes of functions are often characterized by some integrability conditions of the form

$$\int_0^1 \sqrt{\log N(\varepsilon; F; |.|)}\, d\varepsilon < \infty,$$

arising from maximal inequalities. Such a condition essentially ensures that the size of class F is not too big and that one may be able to approximate any of its elements (up to ε) by functions in a set of finite cardinality. In our non-iid setting, we will essentially consider the $L_2(P)$ norm for $|.|$ and use uniform covering numbers

$$\sup_{Q \in D} N\left(\varepsilon \|H\|_{2;Q}; F; \|.\|_{2;Q}\right),$$

where D is the set of all discrete probability measures Q such that $0 < \int_{\mathbb{X}} H^2 dQ < +\infty$. Explicit calculus of (uniform) covering numbers for general classes of functions may be found in several textbooks, see van der Vaart and Wellner (1996) or van de Geer (2000).

5.2 empirical process in survey sampling

We now introduce two different empirical processes built from survey data, the asymptotic behaviors of which shall be investigated at length in Section 5.3.

5.2.1 The Horvitz-Thompson empirical process

In the context of survey data drawn through a general survey plan R_N, the empirical measure P_N cannot be computed since the whole statistical population is not observable. Hence, a variant based on the observations must be naturally considered. For any measurable set $M \subset \mathbb{X}$, the Horvitz-Thompson estimator of the empirical probability $P_N(M) = N^{-1} \sum_{i=1}^{N} \delta_{x_i}(M)$ based on the survey data described above is defined as follows, see Horvitz and Thompson (1951):

$$P_{R_N}^{\Phi(R_N)}(M) := \frac{1}{N} \sum_{i=1}^{n} \frac{\pi_i}{\pi_i} \delta_{X_i}(M) = \frac{1}{N} \sum_{i \in PS} \frac{I_{\{i \in PS\}}}{\pi_i} \delta_{X_i}(M). \tag{5.1}$$

We highlight the fact that the measure $P_{R_N}^{\Phi(R_N)}$ is an unbiased estimator of P (resp. P_N, when conditioned upon $(X_1, \ldots, X_N)$) although it is not a probability measure. For a fixed subset M, the consistency and asymptotic normality of the estimator in Equation (5.1) are established in Robinson (1982) and Berger (1998), as N tends to infinity. When considering the estimation of measure P_N (the measure of interest in survey sampling) over a class of functions $\mathcal{F}$, we are led to the asymptotic study of the collection of random processes

$$G_{R_N}^{\Phi(R_N)} := \left\{ G_{R_N}^{\Phi(R_N)} f \right\}_{f \in \mathcal{F}},$$

where

$$G_{R_N}^{\Phi(R_N)} f := \sqrt{N} \left(P_{R_N}^{\Phi(R_N)} - P_N \right) f = \frac{1}{\sqrt{N}} \sum_{i=1}^{n} \left(\frac{\pi_i}{\pi_i(R_N)} - 1 \right) f(X_i); \tag{5.2}$$

which shall be referred to as the $\mathcal{F}$-indexed Horvitz-Thompson empirical process (HT-empirical process, in short). The seemingly redundant notation $G_{R_N}^{\Phi(R_N)}$ is motivated by the fact that extensions involving first order probabilities related to a different sampling scheme T_N will be considered in the sequel. Precisely, $G_{R_N}^{\Phi(T_N)}$ shall denote the process obtained when replacing all $\pi_i(R_N)$ by $\pi_i(T_N)$, $1 \leq i \leq N$, in Equation (5.2).

The main purpose of this chapter is to establish the convergence of the re-weighted empirical process $(G_{R_N} f)_{f \in \mathcal{F}}$ under adequate hypotheses involving some properties of measure P, certain characteristics of the sequence of sampling plans (R_N), and the "complexity" of class $\mathcal{F}$ (in the classical metric entropy sense) as well. In particular, such a result would permit to describe the asymptotic behavior of the quantity below (assumed to be almost-surely finite, see Assumption 5.8):

$$\left\| G_{R_N(R_N)} \right\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| G_{R_N(R_N)} f \right|.$$

By virtue of Cauchy-Schwarz inequality and Assumption 5.1 and Assumption 5.8, we almost-surely have, $\forall N \geq 1$,

$$\left\| G_{R_N(R_N)} \right\|_{\mathcal{F}}^2 \leq \frac{1}{N} \sum_{i=1}^{n} \left(\frac{\varepsilon_i - \pi_i}{\pi_i} \right)^2 \sum_{i=1}^{n} H^2(X_i)$$
$$\leq \frac{1}{\pi^2} \sum_{i=1}^{n} H^2(X_i) + 1 .$$

Under Assumption 5.1 and Assumption 5.8, the $\mathcal{F}$-indexed HT-empirical process in Equation (5.2) may thus be seen as a sequence of random elements of $\ell^\infty(\mathcal{F})$.

Example 5.9 – Empirical cumulative distribution function. In the case where $\mathcal{X} = \mathbb{R}^d$ with $d \geq 1$ for instance, a situation of particular interest is that where $\mathcal{F}$ is the class of indicator functions of rectangles of the type

$$(-\infty ; x] := \prod_{j=1}^{d} (-\infty ; x_j] \; ; \; x = (x_1, \ldots, x_d) \in \mathbb{R}^d .$$

Then, the empirical process can be identified with the Horvitz-Thompson version of the empirical cumulative distribution function (cdf) $F_{R_N}^{(R_N)}(x) := P_{R_N}^{(R_N)}(-\infty ; x]$, $x \in \mathbb{R}^d$, and the goal pursued boils down to investigating conditions under which uniform versions of the Law of Large Numbers (LLN) and of the Central Limit Theorem (CLT) hold for $F_{R_N}^{(R_N)}(x) - F_N(x)$, where $F_N(x) := P_N(-\infty ; x]$. As shall be seen later, the study of the asymptotic behavior of this empirical process lies at the center of the validity of the confidence band construction considered in Section 5.4.

5.2.2 Alternative estimate in the Poisson sampling case

The Poisson sampling scheme T_N (see Example 5.3) has been the subject of much attention, especially in Hàjek (1964), where asymptotic normality of (pointwise) Horvitz-Thompson estimators have been established in this specific case. Following in the

footsteps of this seminal contribution, we consider the following Poisson version of the empirical process rather than the original process :

$$\mathbb{G}^{p}_{T_N} f := \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\delta_i - p_i) \left(\frac{f(X_i)}{p_i} - \mu_{N;p}(f) \right) ; \quad f \in \mathcal{F}; \tag{5.3}$$

where for all $f \in \mathcal{F}$,

$$\mu_{N;p}(f) := \frac{1}{d_N} \sum_{i=1}^{N} (1 - p_i) f(X_i) \quad \text{and} \quad d_N := \sum_{i=1}^{N} p_i (1 - p_i).$$

Under the assumption that $d_N \longrightarrow +\infty$ as $N \longrightarrow +\infty$, it has been established in Hàjek (1964, Lemma 3.2) that conditioned upon $(X_1, \ldots, X_N)$, for fixed $f \in \mathcal{F}$, when N tends to infinity and under a Lindeberg-Feller type condition, the weighted sum of independent random variables in Equation (5.3) can be approximated by a centered Gaussian random variable with (conditional) variance

$$V_N^2(f) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{f(X_i)}{p_i} - \mu_{N;p}(f) \right)^2 p_i (1 - p_i).$$

As claimed by Theorem 5.14 in the next section, this result can be extended to a functional framework under adequate hypotheses.

Although the subscript T_N in $\mathbb{G}^{p}_{T_N} f$ could have been dropped since the process above only depends on vector p, we keep it in order to emphasize that the corresponding inclusion vector δ is distributed according to the sampling scheme T_N. In this subsection, the weights $p_i := \pi_i(T_N)$, $1 \leq i \leq N$, correspond to the inclusion probabilities of the Poisson sampling plan. Later on, when investigating a general sampling scheme R_N, we shall consider the Poisson-like empirical process defined by

$$\mathbb{G}^{p}_{R_N} f := \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\delta_i - p_i) \left(\frac{f(X_i)}{p_i} - \mu_{N;p}(f) \right) ;$$

where $p := (p_1, \ldots, p_N)$ is the vector of first order inclusion probabilities of a Poisson design. In general, it will not coincide with those of R_N, namely $\pi(R_N)$, but the subscript specifies that δ is still distributed according to R_N (in particular, $E(\delta_i) = \pi_i(R_N)$ for $i = 1, \ldots, N$). In the subsequent analysis, we start off by establishing that the process $\mathbb{G}^{p}_{T_N}$ can be asymptotically approximated by a Gaussian process.

5.3 asymptotic results

The main results of the chapter are stated in the present section. As a first go, we establish a FCLT for the empirical process variant of Equation (5.3) in the Poisson survey scheme case. Combined with an approximation result, it will serve as the main tool for proving next a similar result in the context of rejective sampling.

5.3.1 Limit of the empirical process for the Poisson survey scheme

The purpose of this section is to obtain a Gaussian approximation of the empirical process $\sqrt{N}G^{p}_{T_N}$ related to a Poisson survey plan T_N with first order inclusion probabilities $p = (p_1; \dots; p_N)$ depending on some auxiliary variable W (see Section 5.1.1.3). The proof relies on Theorem 2.11.1 in van der Vaart and Wellner (1996), applied to the triangular collection of independent variables defined for all $f \in \mathcal{F}$ by

$$Z_{N;i}(f) := Z_{N;i}(f;\omega) := \frac{1}{\sqrt{N}}(\delta_i - p_i)\,\frac{f(X_i)}{p_i} - \alpha_{N;p}(f) \quad \text{for } i \in \{1; \dots; N\}.$$

For clarity, the result is recalled below.

Theorem 5.10 – Triangular arrays (van der Vaart and Wellner, 1996).

Let $Z_{N;i}(f)$, $1 \leq i \leq N$ be independent $\mathcal{F}$-indexed stochastic processes defined on the product probability space $\bigotimes_{i=1}^{N}(\Omega; \mathcal{A}; P)(\Omega; \mathcal{B}(\Omega_i(R_N)))$ where the process $Z_{N;i}(f) := Z_{N;i}(f;\omega)$ only depends on the i-th coordinate of $\omega := (\omega_1; \dots; \omega_N)$. Assume that the maps

$$(\omega_1; \dots; \omega_N) \mapsto \sup_{\rho_P(f;g)<\delta} \left| \sum_{i=1}^{N} \varepsilon_i \left(Z_{N;i}(f;\omega) - Z_{N;i}(g;\omega)\right) \right|$$

and

$$(\omega_1; \dots; \omega_N) \mapsto \sup_{\rho_P(f;g)<\delta} \sum_{i=1}^{N} \varepsilon_i \left(Z_{N;i}(f) - Z_{N;i}(g)\right)^2$$

are measurable for every $\delta > 0$, every $(\varepsilon_1; \dots; \varepsilon_N) \in \{-1; 0; 1\}^N$ and every $N \in \mathbb{N}$.

Further define the random semi-metric

$$d_N^2(f;g) := \sum_{i=1}^{N} \left(Z_{N;i}(f) - Z_{N;i}(g)\right)^2;$$

and suppose that the following conditions are fulfilled.

i) $\sum_{i=1}^{\infty} E\left[\{Z_{N;i}(f)\}^2_F \, \mathbb{1}\{Z_{N;i}(f)\}_F > \eta\sqrt{N}\right] \xrightarrow[N\to\infty]{} 0$ for every $\eta > 0$.

ii) $\sup_{\rho_P(f;g)<\delta} \sum_{i=1}^{\infty} E\left[\left(Z_{N;i}(f) - Z_{N;i}(g)\right)^2\right] \xrightarrow[N\to\infty]{} 0$ as $\delta \to 0$.

iii) $\int_0^{\infty} \sqrt{\log N(\varepsilon; \mathcal{F}; d_N)}\, d\varepsilon \xrightarrow[N\to\infty]{} 0$ as $\delta \to 0$.

iv) The sequence of covariance functions $\mathrm{cov}(Z_{N;i}(f); Z_{N;i}(g))$ converges pointwise on $\mathcal{F} \times \mathcal{F}$ as $N \to \infty$ to a non degenerate limit $\Gamma(f;g)$.

Then the sequence $\sum_{i=1}^{N}(Z_{N;i}(f) - E(Z_{N;i}(f)))$ is ρ_P-equicontinuous and converges in $\ell^1(\mathcal{F})$ to a Gaussian process with covariance function $\Gamma(f;g)$.

5.3.1.1 Convergence of the covariance operator

The following intermediary results show that condition iv) in Theorem 5.10 is fulfilled in the particular case of Poisson survey plans. For $(f;g) \in F^2$, set

$$\mathrm{cov}_{N;p}(f;g) := \frac{1}{N}\sum_{i=1}^{N}\left(\frac{f(X_i)}{p_i} - \tau_{N;p}(f)\right)\left(\frac{g(X_i)}{p_i} - \tau_{N;p}(g)\right)p_i(1-p_i).$$

Due to the independence of the δ_i's, it is clear that

$$\mathrm{cov}_{T_N}\left(\sqrt{N}G^p_{T_N}(f);\sqrt{N}G^p_{T_N}(g)\right) := \mathrm{cov}\left(\sqrt{N}G^p_{T_N}(f);\sqrt{N}G^p_{T_N}(g)\right)(X_i;W_i)_{1\leq i\leq N} = \mathrm{cov}_{N;p}(f;g).$$

We thus essentially have to determine conditions ensuring that $\mathrm{cov}_{N;p}(f;g)$ has a non-degenerate limit. The following assumptions are by no means necessary but provide a useful framework to derive such conditions. Similar types of assumptions may be found in Bonnéry et al. (2011) or Hàjek (1964) for instance.

Recall that inclusion probabilities were defined relative to some auxiliary variable W. An additional assumption on the latter is required in the subsequent result.

Assumption 5.11 The couples of random vectors $(X_1;W_1);\ldots;(X_N;W_N)$ are iid (exchangeable at least) with distribution $P_{X;W}$. Moreover, the conditional inclusion probabilities $p := (p_1;\ldots;p_N)$ are then given for all $i \in \{1;\ldots;N\}$ and $W_{(N)} \in W^N$ by

$$p_i := p(W_i) := E\left(\delta_i \mid W_{(N)}\right).$$

Remark 5.12 It can happen that p_i not only depends on W_i, but on the entire vector $W_{(N)}$. It is the case, for instance, when there is a unique auxiliary variable W to which weights are proportional:

$$p_i := n\frac{W_i}{\sum_{j=1}^{N} W_j}.$$

In such situations the iid property of the vectors $(X_i;W_i)$, $1 \leq i \leq N$, can be used to bypass the part involving all $(W_1;\ldots;W_N)$ in the subsequent asymptotic analysis.

Under this supplementary condition, we have the following result, the proof of which can be found in Section 5.5.1.

Lemma 5.13 – Limit of the covariance operator. Suppose that Assumption 5.1, Assumption 5.8 and Assumption 5.11 are fulfilled and that

$$0 < \int_W p^2(w)\,P_W(dw) < +\infty.$$

Then we have

$$\frac{1}{N} d_N \underset{N\to\infty}{\longrightarrow} D_p := \int_W (1 - p(w))\, p(w)\, P_W(dw) > 0$$

and

$$\mathrm{cov}_{N,p}(f,g) \underset{N\to\infty}{\longrightarrow} \Gamma(f,g),$$

where for all $(f,g) \in F^2$,

$$\Gamma(f,g) := \int_{X\times W} f(x)g(x) \left(\frac{1}{p(w)} - 1\right) P_{X,W}(dx,dw) - \Gamma_p(f)\Gamma_p(g) D_p, \tag{5.4}$$

with

$$\Gamma_p(f) := \frac{1}{D_p} \int_{X\times W} (1 - p(w))\, f(x)\, P_{X,W}(dx,dw).$$

5.3.1.2 Functional Central Limit Theorem

Applying Theorem 5.10 to the empirical process $\tilde{G}^p_{T_N} f$ defined in Equation (5.3) thus leads to the theorem below, proved in Section 5.5.2.

Theorem 5.14 – FCLT in the Poisson survey case. Suppose that Assumption 5.1, Assumption 5.8 and Assumption 5.11 hold, as well as the following conditions.

i) Lindeberg-Feller type condition: $\forall \varepsilon > 0$,

$$E\left[(Z_{N,i})^2 \mathbb{I}\left\{Z_{N,i} > \varepsilon \sqrt{N}\right\}\right] \underset{N\to\infty}{\longrightarrow} 0,$$

with $Z_{N,i} := (\pi_i - p(W_i)) \sup_{f\in F} \left|\frac{f(X_i)}{p(W_i)} - \Gamma_{N,p}(f)\right|$.

ii) Uniform entropy condition: let D be the set of all finitely discrete probability measures defined in Section 5.1.2.3, and assume

$$\int_0^1 \sup_{Q\in D} \sqrt{\log(N(\varepsilon \|F\|_{2,Q}, F, \|\cdot\|_{2,Q}))}\, d\varepsilon < +\infty.$$

Then there exists a Γ_P-equicontinuous Gaussian process G in $\ell^\infty(F)$ with covariance operator Γ given by Equation (5.4) such that

$$\tilde{G}^p_{T_N} \Rightarrow G \text{ weakly in } \ell^\infty(F), \text{ as } N \to \infty.$$

Remark 5.15 – On the Lindeberg-Feller condition. Observe that, as can be proved using Hölder's inequality, condition i) in Theorem 5.14 can be replaced by the simpler condition: $\exists \delta > 0$ such that

$$\text{i)} \quad E_{P_{X,W}} \left[\frac{H(X_i)}{p(W_i)} \right]^{2+\delta} < \infty \quad E_{T_N} \left[(\pi_i \circ p(W_i))^{2+\delta} \circ \Phi(X_i, W_i) \right]^{\frac{1}{\delta}} < +\infty .$$

5.3.2 The case of rejective sampling

As shall be shown herein-after, the result obtained above in the case of a Poisson sampling scheme may carry over to more general survey plans, as originally proposed in the seminal contribution of Hàjek (1964).

5.3.2.1 Reduction to simpler sampling designs

The lemma stated below, following in the footsteps of Berger (1998), shows that the study of the empirical process related to a general sampling design R_N may be reduced to that related to a simpler sampling design, T_N say, which is close to R_N with respect to some metric and entirely characterized by its first order inclusion probabilities. The only "drawback" is that the estimator involved in this approximation result is not the Horvitz-Thompson estimator, since it does not involves the inclusion probabilities of the sampling plan of interest but those related to a Poisson scheme (Hàjek, 1964). However, as will be shown next, the two estimators may asymptotically coincide, as N tends to $+\infty$.

In order to formulate the approximation result needed in the sequel, we introduce, for two sampling designs R_N and T_N, the total variation metric

$$\| R_N - T_N \|_1 := \sum_{s \in \mathcal{P}(U_N)} |R_N(s) - T_N(s)| ;$$

as well as the entropy

$$D(T_N ; R_N) := \sum_{s \in \mathcal{P}(U_N)} T_N(s) \log \left[\frac{T_N(s)}{R_N(s)} \right] .$$

In practice, T_N will typically be the Poisson sampling plan investigated in the previous subsection and $\mathbb{G}_{T_N}^{\Phi(T_N)}$ the corresponding empirical process.

Lemma 5.16 – Approximation result. Let R_N and T_N be two sampling designs and assume that T_N is entirely characterized by its first order inclusion probabilities, $\Phi(T_N)$. Then, the empirical processes $\mathbb{G}_{T_N}^{\Phi(T_N)}$ and $\mathbb{G}_{R_N}^{\Phi(T_N)}$ valued in $\ell^1(\mathcal{F})$ satisfy the relationships:

$$d_{BL}\left(\mathbb{G}_{T_N}^{\Phi(T_N)}, \mathbb{G}_{R_N}^{\Phi(T_N)} \right) \leq \| R_N - T_N \|_1 \leq \sqrt{2 D(T_N ; R_N)}.$$

Consequently, if the sequences $(R_N)_{N\geq 1}$ and $(T_N)_{N\geq 1}$ are such that $\|R_N \Delta T_N\|_1$ tends to 0 or $D(T_N; R_N) \to 0$ as $N \to \infty$ and if there exists a Gaussian process G such that

$$d_{BL}(\mathbb{G}_{T_N}^{\pi(T_N)}; G) \xrightarrow[N\to\infty]{} 0;$$

then we also have

$$d_{BL}(\mathbb{G}_{R_N}^{\pi(T_N)}; G) \xrightarrow[N\to\infty]{} 0.$$

The same result holds true when replacing $\mathbb{G}_{T_N}^{\pi(T_N)}$ and $\mathbb{G}_{R_N}^{\pi(T_N)}$ by $G_{T_N}^{\pi(T_N)}$ and $G_{R_N}^{\pi(T_N)}$ respectively.

This result, proved in Section 5.5.3, reveals that as soon as a possibly complicated survey design R_N can be approximated by a simpler one T_N through some coupling argument ensuring that the $\|.\|_1$ distance between them decays to zero (as in Hàjek, 1964), then an asymptotic approximation result possibly holding true for the empirical process related to T_N immediately extends to that related to R_N, when built with the inclusion probabilities $p = \pi(T_N)$. As shall be seen below, a typical situation where this result applies corresponds to the case where R_N is a rejective sampling design, while T_N is a Poisson sampling design, as in Hàjek (1964). Other natural applications may arise in the framework of post-stratification, which can be connected with empirical likelihood results.

5.3.2.2 Empirical process for the rejective sampling and its variants

The Central Limit Theorem for rejective sampling and some variants of this survey scheme has been studied at length in Hàjek (1964) and Berger (1998, 2011). Consider the rejective sampling scheme defined in Example 5.5 from a given vector π^R corresponding to the vector $p := (p_1, \ldots, p_N) = (p(W_1), \ldots, p(W_N)) =: p(W)$. Assume in addition that the representation is canonical, i.e. is such that $\sum_{i=1}^N p(W_i) = n$. The key argument for proving a CLT in the rejective sampling case consists in exhibiting a certain coupling $((\mathbb{I}_1, \ldots, \mathbb{I}_N); (\mathbb{I}_1^*, \ldots, \mathbb{I}_N^*))$ of the Poisson sampling scheme with inclusion probabilities $p(W_1), \ldots, p(W_N)$ and the rejective sampling scheme with corresponding inclusion probabilities π^R such that $\|R_N \Delta T_N\|_1 \to 0$, see Hàjek (1964, p. 1503-1504) for further details. A straightforward application of Lemma 5.16 will then immediately yield a functional CLT in our framework. We point out that, under the rejective sampling scheme, the survey size is fixed, so that

$$\sum_{i=1}^N (\mathbb{I}_i - p(W_i)) = n - n = 0.$$

Thus, we have:

$$\mathbb{G}_{R_N}^p f := \frac{1}{\sqrt{N}} \sum_{i=1}^N (\mathbb{I}_i - p(W_i)) \left[\frac{f(X_i)}{p(W_i)} - \Phi_{N;p}(f) \right]$$

$$= \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i}{p(W_i)}f(X_i) - \int f\right)$$
$$=: G_{R_N}^{p(W)} f.$$

Hence, the Poisson-like empirical process coincides, in that case, with the original HT-empirical process where the weights p(W) are involved instead of the true inclusion probabilities π^R, the latter being however asymptotically equivalent to the former, see Hàjek (1964).

The next theorem is obtained by combining Lemma 5.16 with Theorem 5.14.

Theorem 5.17 – FCLT in the rejective survey with Poisson weights case. Suppose that Assumption 5.1, Assumption 5.8, Assumption 5.11 and conditions i) and ii) of Theorem 5.14 are satisfied. Then, there exists a ρ_P-equicontinuous Gaussian process G in $\ell^1(\mathcal{F})$ with covariance operator Γ given by Equation (5.4) such that

$$G_{R_N}^{p(W)} \Rightarrow G \text{ weakly in } \ell^1(\mathcal{F}), \text{ as } N \to \infty.$$

It has been established in Berger (1998) that for a variety of sampling plans R_N, including the Rao-Sampford scheme defined in Example 5.6, we have $D(R_N; T_N) \to 0$ as $N \to \infty$. By virtue of Lemma 5.16, Theorem 5.17 naturally extends to these sampling schemes.

Going back to the original HT-empirical process in Equation (5.2) related to the plan R_N, the corollary below reveals that the asymptotic result still holds true for the latter (see the proof in Section 5.5.4). This essentially follows from the fact that the weights p(W) and the inclusion probabilities corresponding to the rejective sampling are asymptotically equivalent.

Corollary 5.18 – FCLT in the rejective survey case. Suppose that Assumption 5.1, Assumption 5.8, Assumption 5.11 and conditions i) and ii) of Theorem 5.14 are satisfied. Then, there exists a ρ_P-equicontinuous Gaussian process G in $\ell^1(\mathcal{F})$ with covariance operator Γ given by Equation (5.4) such that

$$G_{R_N}^{\pi(R_N)} \Rightarrow G \text{ weakly in } \ell^1(\mathcal{F}), \text{ as } N \to \infty.$$

5.4 application to non-parametric statistics

For illustration purpose, we consider now several statistical applications of the asymptotic results previously established.

5.4.1 Hadamard differentiable functionals

We first highlight that the FCLT stated above permits to establish the asymptotic normality of any statistic that can be expressed as the empirical version of some Hadamard differentiable functional, see Shorack and Wellner (1986). For the sake of clarity, we recall the definition of uniform Hadamard differentiability in Definition 5.19, adapted from Pons and de Turkheim (1991). Our results apply to many situations considered in their paper, related in particular to certain functionals of censored data. Other examples are treated in Gill (1989), van der Vaart and Wellner (1996) (see Chapter 3.9 p. 379 therein, in particular refer to the discussion about the validity of the bootstrap for uniform Hadamard differentiable functionals). Define $B(F;P)$ as the set of measures Q in $\ell^1(F)$ whose paths $f \in F \mapsto Qf := \int f dQ$ are $\|\cdot\|_{2;P}$-uniformly continuous and bounded. This is the smallest natural space containing G. We consider the uniform Hadamard differentiability tangentially to the subspace $B(F;P)$ because it weakens the notion of differentiation and is easier to check in practice.

Definition 5.19 A functional $T : \ell^1(F) \to \mathbb{R}^q$ is said to be uniformly Hadamard differentiable at P tangentially to $B(F;P)$; if and only if there exists a continuous linear mapping dT_P such that for any sequence P_N converging to P, any h_N converging to $h \in B(F;P)$ and every t_N converging to 0 such that $P_N + t_N.h_N \in \ell^1(F)$, we have:

$$\frac{T(P_N + t_N.h_N) - T(P_N)}{t_N} - dT_P.h \xrightarrow[t_N \to 0]{} 0.$$

Notice that T may be defined not on the entire space $\ell^1(F)$ but on a subset L only. In this case, one must check that $P_N + t_N.h_N \in L$.

Remark 5.20 We may in addition assume that the differential dT_P admits an integral representation, i.e.

$$dT_P.h = \int_a T^{(1)}(x;P)\, h(dx);$$

where $T^{(1)}(.;P)$ is the influence function defined from X to B_1 such that we have

$$E_P\left[T^{(1)}(X;P)\right] = 0.$$

We recall that in the robustness terminology (Hampel et al., 1986), the influence function of the parameter $T(P)$ may be calculated directly by computing the derivative of the functional taken at the contaminated distribution $(1-t)P + t\delta_x$, i.e.

$$T^{(1)}(x;P) := \lim_{t \to 0} \frac{T((1-t)P + t\delta_x) - T(P)}{t}.$$

In this case, the limiting distribution may be calculated more easily.

Theorem 5.21 – CLT for Hadamard differentiable functionals. Suppose that the assumptions of Theorem 5.14 hold and that functional $T : L Ä `^1(F) — R^q$ is Hadamard differentiable at P with differential dT_P and influence function $T^{(1)}(x;P)$. Then, as $N — +1$, we have:

$$? \overline{N} \, \square T(P_{R_N}^{\square(R_N)}) \square T(P_N) \, \square Ò \, dT_P.G;$$

where G is a Gaussian process with covariance operator $\square$, as in Equation (5.4).

The result above, the proof of which is available in Section 5.5.5, applies in particular to the following statistics.

Example 5.22 – Expectation and variance. It is well-known that for some appropriate choice of F, the functionals $T(P) = E_P(X)$ and $T(P) = V_P(X)$ are uniformly Hadamard-differentiable. When $T(P) = E_P(X)$, Theorem 5.21 exactly reduces to the Central Limit Theorem established in Hàjek (1964).

Example 5.23 – Cumulative distribution function. In a univariate setting, the functional $T(P) = F(x) := P(X P(\square 1 ; x])$ can be dealt with by simply considering the class of indicator functions u fi– I t u § xu with x P R and applying next Theorem 5.17 and Corollary 5.18. We provide illustrations of this specific example in Section 5.4.3.

5.4.2 Fréchet differentiable functionals

Hadamard differentiability is sometimes difficult to prove and it does not yield a precise control of the remainder for further approximations like Edgeworth expansions. Another approach followed by Dudley (1990) and Barbe and Bertail (1995) is to assume Fréchet differentiability with respect to a metric d_F indexed by a class of function F, for which some uniform entropy conditions hold. A functional is said to be Fréchet differentiable at P for such a metric if there exists a gradient (for instance the influence function $T^{(1)}(x;P)$, which fulfills $E_P \, \square T^{(1)}(x;P) \, \square = 0$) and a continuous function $"(.)$, null at 0, such that for any probability Q,

$$T(Q) \square T(P) = \overset{a}{} T^{(1)}(x;P)(Q \square P)(dx) + d_F(Q;P)"(d_F(Q;P)).$$

It is generally possible to choose the class of functions according to the functional of interest, see for instance Arcones and Giné (1992) for general classes of M-estimators. Notice that in that case, by applying Fréchet differentiability twice, we have

$$? \overline{N} \, \square T(P_{R_N}^{\square(R_N)}) \square T(P_N) \, \square = ? \overline{N} \overset{a}{} T^{(1)}(x;P)(P_{R_N}^{\square(R_N)} \square P_N)(dx) + r_N$$

$$= \frac{1}{N}\sum_{i=1}^{n}\frac{\pi_i}{\pi_i}\,T^{(1)}(X_i;P) + r_N;$$

with a remainder

$$r_N = \sqrt{N}\,d_F(P_{R_N}^{\pi(R_N)};P)-(d_F(P_{R_N}^{\pi(R_N)};P)) + \sqrt{N}\,d_F(P_N;P)-(d_F(P_N;P)).$$

By virtue of the results in Hàjek (1964), it is then obvious that the linear term in this approximation is asymptotically Gaussian with known variance. Controlling the remainder essentially amounts to controlling the behavior of $\sqrt{N}\,d_F(P_{R_N}^{\pi(R_N)};P)$ or alternatively, by the triangular inequality, that of $\sqrt{N}\,d_F(P_{R_N}^{\pi(R_N)};P_N)$, which was the purpose of Section 5.2 and Section 5.3.

5.4.3 Simulation-based Gaussian asymptotic confidence regions

A straightforward application consists in the building of Gaussian confidence regions for the (univariate) empirical cumulative distribution function in the entire population, denoted by $F_N(x)$, $x \in \mathbb{R}$, when the survey scheme is of the rejective type. Indeed, consider the class of functions $\mathcal{F} := \{f_x(.) := \mathbb{I}\{.\leq x\};\ x\in\mathbb{R}\}$. Provided Assumption 5.1 is fulfilled, it respects the required conditions for Corollary 5.18 to hold (see Van der Vaart, 2000, Example 19.16 for the uniform entropy condition and take $H(x)=1$ and $\delta=1$ when checking condition i) in Remark 5.15), which implies in particular that $\{G_{R_N}^{\pi(R_N)}\}_{\mathcal{F}}$ converges in distribution to $\{G\}_{\mathcal{F}}$ as $N\to+\infty$ (Van der Vaart, 2000, Corrolary 19.21). This yields the following asymptotic uniform confidence band of level $\alpha\in(0;1)$ for the population cdf F_N:

$$CB_\alpha := \left[F_{R_N}^{\pi(R_N)} - \frac{q_\alpha}{\sqrt{N}};\ F_{R_N}^{\pi(R_N)} + \frac{q_\alpha}{\sqrt{N}}\right];$$

where $F_{R_N}^{\pi(R_N)}$ is the Horvitz-Thompson estimator of the cdf based on the rejective sample and q_α the α-quantile of random variable $\{G\}_{\mathcal{F}}$. Since in practice q_α is unknown, it needs to be estimated. It can be achieved by means of Monte-carlo simulations, using a simple technique based on the Cholesky decomposition of the covariance matrix (Kroese et al., 2011, Algorithm 5.1).

Algorithm 5.24 – Simulation of the limit process G and estimation of q_α.

1. Choose a grid of real values $\{x_1;\ldots;x_K\}$, $K\geq 1$, and compute the Horvitz-Thompson estimator of $\Gamma(f_{x_k};f_{x_{k'}})$ for each couple $(x_k;x_{k'})$ in $\{x_1;\ldots;x_K\}^2$, namely

$$\Gamma_{R_N}^{p}(f_{x_k};f_{x_{k'}}) := \frac{1}{N}\sum_{i=1}^{n}\frac{\pi_i}{1-p_i}\,\mathbb{I}\{X_i\leq\min(x_k;x_{k'})\}$$

$$\square \frac{1}{N} \frac{\sum_{i=1}^{N} \square_i \frac{1 \square p_i}{p_i} I\{X_i \S x_k\} \sum_{i=1}^{N} \square_i \frac{1 \square p_i}{p_i} I\{X_i \S x_k \square u\}}{\sum_{i=1}^{N} \square_i (1 \square p_i)}.$$

2. Derive the Cholesky decomposition $\square^{p}_{R_N}(f_{x_k}; f_{x_k \square}) = LL^{1}$, where L is a lower-triangular Cholesky matrix.
3. Generate $B \degree 1$ independent copies $Y_1; \ldots; Y_B$ of the Gaussian random vector $Y := (Y_1; \ldots; Y_K)^{1}$ with null expectation and covariance I, the identity matrix.
4. Compute $Z_b := LY_b$, $b \P \{1; \ldots; Bu$, which are considered as realizations of the limit process G.
5. For each $b \P \{1; \ldots; Bu$, calculate $Z^{\square}_b := \}Z_b\}_1$ the maximum absolute distance to 0 of the path Z_b and sort the obtained sample $Z^{\square}_{1;B} \S \square\square\S Z^{\square}_{B;B}$.
6. Set $\hat{q}_{\square} := Z^{\square}_{\{B \square u+ 1;B}$, the empirical $\square$-quantile of the sample of maximum absolute deviations, where $\{u$ denotes the floor function.

In the next subsections, a set of numerical experiments is performed to provide illustrative examples of this technique.

5.4.3.1 Experiment setting

Simulations were based on the following model, chosen for its simplicity in terms of both computation and interpretation:

$$\begin{aligned} &X = \square W + U; \quad \square \P \{0; 1u \\ &W \; ; \; TN(\square; \square^2_W; w_{\square}; w^{\square}); \\ &U \; ; \; N(0; \square^2_U); \\ &P(W \S w; U \S u) = P(W \S w) P(U \S u); \end{aligned}$$

where X is the variable of interest, W the auxiliary information, U a white noise independent from W, and $TN(0; \square^2_W; w_{\square}; w^{\square})$ refers to the truncated Normal distribution over $[w_{\square}; w^{\square}]$, with expectation $\square$ and variance $\square^2_W$. Such a representation enables a simple control of the dependence between X and W, since their correlation is then

$$corr(X; W) = \square \frac{\square_W}{\sqrt{\square^2_W + \square^2_U}}.$$

For a given population U_N of size N, where it is assumed that $\{W_i; i \P U_N u$ (resp. $\{U_i; i \P U_N u$) are independent (hence exchangeable) realizations of W (resp. U), inclusion probabilities of the Poisson sampling scheme are defined as

$$p_i = p(W_i) = n \frac{W_i}{\sum_{j=1}^{N} W_j}; \tag{5.5}$$

with n the desired expected sample size (Hàjek, 1964, Section 6, p.1512). When the inclusion probabilities are proportionate to the auxiliary variable like in Equation (5.5),

the stronger the correlation between X and W, the smaller the variance of the estimator of the population mean $\frac{1}{N} \sum_{i=1}^{N} X_i$ (or, equivalently, of the total $\sum_{i=1}^{N} X_i$). Recall that under Assumption 5.1, we have $n/N \to c \in (0;1)$ as both n and N tend to infinity. Hence, p_i can be viewed as the empirical version in the population of

$$p(W) := W \frac{c}{E(W)}.$$

Observe that thus defined, $p(W) \in [p_\star; p^\star]$, where $p_\star = cw_\star/\mu$ and $p^\star = cw^\star/\mu$, which offers an easy way of ensuring Assumption 5.1 is fulfilled.

Numerical experiments were conducted on a set of populations with increasing sizes $N = 10^2, 5 \times 10^2, 10^3, 5 \times 10^3$ and 10^4. Though the latter may seem quite small to study asymptotic properties, they are in fact representative of many practical situations, where populations under the microscope have moderate sizes in comparison to nationwide surveys. Several scenarios were investigated depending on both the variance parameter σ_U^2 and the coefficient β, so as to cover situations where corr(X; W) is high, low or null. They are summarized in Table 5.1. For each scenario, two sample sizes were considered: one small with $n = 0.1 \times N$ and one relatively large with $n = 0.5 \times N$. Parameters of the distribution of W were chosen to ensure that for all $i \in U_N$, $p_i \in [0.01; 1]$. Specifically, we set $\mu = 1$, $\sigma_W^2 = 0.09$, $w_\star = 0.1$ and $w^\star = 2$, thereby implying that $(p_\star; p^\star) = (0.01; 0.02)$ when $n = 0.1 \times N$ and $(p_\star; p^\star) = (0.05; 1)$ when $n = 0.5 \times N$.

Table 5.1 – List of scenarios depending on β and σ_U^2, and corresponding model characteristics

Scenario	β	σ_U^2	corr(X; W)
S_1	1	0.01	0.95
S_2	1	35.91	0.05
S_3	0	35.91	0

For each scenario, we drew 1000 samples according to a rejective sampling scheme, following Algorithm 5.9 in Tillé (2006). The true inclusion probabilities, denoted by π_i, $1 \leq i \leq N$, could have been deduced from their Poisson equivalents defined in Equation (5.5) using Formula (5.13) in Tillé (2006). Though a very popular and natural algorithm, due to the limits of computer precision, the successive approximations it involves can lead to unexpected results like negative inclusion probabilities. Other algorithms have been developed to compute exact inclusion probabilities in a rejective sampling scheme (e.g. Aires, 1999 or Tillé, 2006, Algorithms 5.8 and 5.9), but again, especially when N is large, they have a tendency to produce illogical estimates. This is why we adopted a simpler, although computationally expensive Monte-Carlo approximation technique, based on the repetition (10^5 times) of the basic algorithm stated in Example 5.5. Notice that since rejective sampling is a Poisson sampling conditioned upon its size, we have $(p_i = 1) \Rightarrow (\pi_i = 1)$.

We constructed asymptotic uniform 95% confidence bands of the population cdf F_N using Algorithm 5.24, with B = 1000 and K = 10, 20 or 100 depending on the sample size n. More precisely, for n • 100, the grid was made of the standard empirical percentile estimators of variable X based on each artificial samples. To enable computation of the lower-triangular Cholesky matrix, which requires that the covariance matrix has full rank, we confined ourselves to deciles for n = 10 and to the quantiles of levels 0.05; 0.1; : : : ; 0.95; 1 for n = 50.

5.4.3.2 Experiment results

The average and maximal width of the confidence bands over the 1000 simulated samples for each scenario are given in Table 5.2. Coverage probabilities were also estimated, the results of which are displayed in Table 5.3. Finally, some graphical illustrations are provided in Figure 5.1.

Table 5.2 – Average (Av) and maximal (Mx) width of confidence bands multiplied by 100

		N									
		10^2		$5 \cdot 10^2$		10^3		$5 \cdot 10^3$		10^4	
Scenario	c	Av	Mx	Av	Mx	Av	Mx	Av	Mx	Av	Mx
S_1	0.5	31.24	37.12	15.68	20.70	11.43	14.61	5.10	5.76	3.58	3.94
	0.1	87.68	116.62	42.67	68.80	32.67	45.64	14.46	18.15	10.19	12.33
S_2	0.5	27.74	33.55	13.77	16.52	9.95	12.88	4.48	5.05	3.17	3.56
	0.1	76.43	103.01	37.63	53.10	28.40	40.36	12.70	15.95	8.99	10.81
S_3	0.5	27.53	32.79	13.67	16.12	9.85	13.36	4.43	5.13	3.14	3.41
	0.1	75.65	105.18	37.20	53.93	27.93	38.51	12.57	15.92	8.90	10.61

Table 5.3 – Estimated coverage probabilities (in %)

		N				
Scenario	c	10^2	$5 \cdot 10^2$	10^3	$5 \cdot 10^3$	10^4
S_1	0.5	96.30	96.80	96.80	97.80	97.00
	0.1	94.74	97.20	97.12	98.30	98.50
S_2	0.5	92.50	97.30	95.60	96.40	96.40
	0.1	89.12	93.80	93.55	95.60	96.90
S_3	0.5	91.80	96.50	95.50	96.10	95.60
	0.1	87.82	93.50	93.48	94.80	96.70

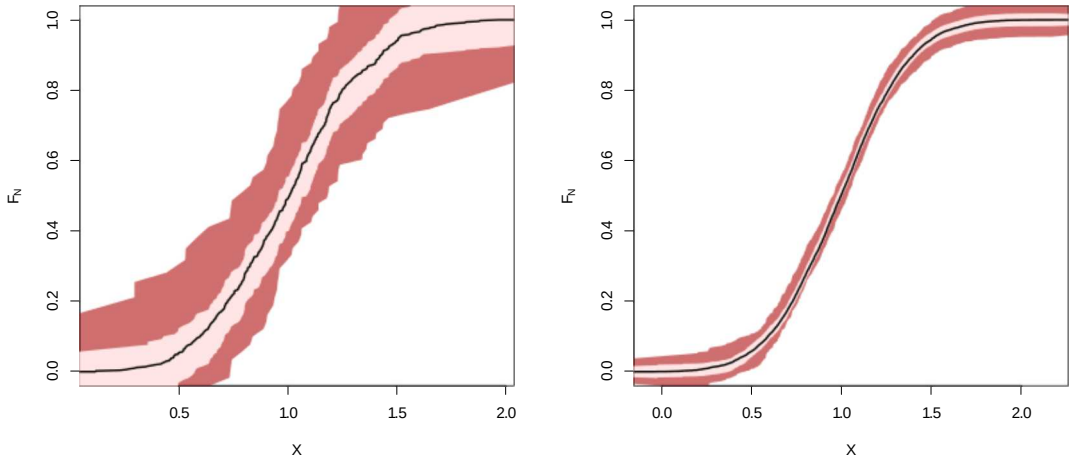


Figure 5.1 – Example of the 95% confidence bands of the empirical distribution function in the population F_N (black line) constructed on one of the 1000 simulated samples under scenario S_1 with $c = 0.1$ (dark pink area) and $c = 0.5$ (light pink area) for $N = 5 \cdot 10^2$ (left hand plot) and $N = 10^4$ (right hand plot)

As expected, the larger N and c, the smaller the confidence bands. Regarding coverage probabilities, they appear to be close to 95%, the desired level, for any N and c. The most remarkable variability is that observed between scenarios: confidence bands get significantly tighter as the correlation between X and W decreases. As a consequence, estimated coverage probabilities are systematically smaller in scenarios S_2 and S_3 than in scenario S_1, especially when $N = 10^2$ and $c = 0.1$. This phenomenon is due to the formula used to construct inclusion probabilities, in Equation (5.5). Let us dwell for a moment on this expression. It ensures that the Horvitz-Thompson estimator (based on the Poisson inclusion probabilities) of the expectation of W coincides with the classical empirical mean in the entire population:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\pi_i}{p_i}W_i = \frac{1}{N}\sum_{i=1}^{N}\frac{\pi_i}{n\,\frac{W_i}{\sum_{j=1}^{N}W_j}}W_i = \frac{1}{N}\sum_{i=1}^{N}\frac{\pi_i}{n}\sum_{j=1}^{N}W_j = \frac{1}{N}\sum_{j=1}^{N}W_j;$$

since $\sum_{i=1}^{N}\pi_i = n$ by definition. It is no surprise then that the stronger the correlation between X and W, the closer (in terms of variance) the weighted mean $\frac{1}{N}\sum_{i=1}^{N}\frac{\pi_i}{p_i}X_i$ is to its population counterpart. However, when considering empirical distribution functions, the standard and sample estimators for W are no longer equal. Hence, not only does the model in Equation (5.5) fail to improve the variance of the HT-cdf of X, but the deviations of $F^{p}_{R_N}$ are expected to grow as the link between X and W tightens. To counterbalance this drawback, we could for instance choose the inclusion probabilities p_i, $1 \leq i \leq N$, that minimize the uniform difference between the HT and the empirical cdf of W (see for instance Rueda et al., 2007). Such refinements are left for further research. Although not optimal, the confidence bands constructed on our

numerical experiments are still satisfactory and advocate the utility of our asymptotic results whatever the available inclusion probabilities.

5.5 proofs and supplements

5.5.1 Limit of the covariance operator

We shall prove Lemma 5.13, the statement of which is recalled below.

Lemma Suppose that Assumption 5.1, Assumption 5.8 and Assumption 5.11 are fulfilled and that

$$0 < \int_W p^2(w)\, P_W(dw) < 1 .$$

Then we have

$$\frac{1}{N} d_N \xrightarrow[N\to\infty]{} D_p := \int_W (1 - p(w))\, p(w)\, P_W(dw) > 0$$

and

$$\mathrm{cov}_{N;p}(f;g) \xrightarrow[N\to\infty]{} \Gamma(f;g);$$

where for all $(f;g) \in F^2$,

$$\Gamma(f;g) := \int_{X\times W} f(x)g(x) \left(\frac{1}{p(w)} - 1\right) P_{X;W}(dx;dw) - \ell_p(f)\ell_p(g) D_p ;$$

with

$$\ell_p(f) := \frac{1}{D_p} \int_{X\times W} (1 - p(w))\, f(x)\, P_{X;W}(dx;dw).$$

Proof Notice first that under Assumption 5.11, we have: $\forall (f;g) \in F^2$,

$$\mathrm{cov}_{N;p}(f;g) = \frac{1}{N} \sum_{i=1}^{N} \frac{f(X_i)}{p(W_i)} \frac{g(X_i)}{p(W_i)} p(W_i)\,(1 - p(W_i)) - \ell_{N;p}(f)\, \ell_{N;p}(g) \frac{1}{N} \sum_{i=1}^{N} p(W_i)\,(1 - p(W_i));$$

with

$$\ell_{N;p}(f) = \frac{\sum_{i=1}^{N} (1 - p(W_i))\, f(X_i)}{\sum_{i=1}^{N} (1 - p(W_i))\, p(W_i)}.$$

Now it is sufficient to apply the Strong Law of Large Numbers for exchangeable vectors to obtain that

$$\frac{1}{N} d_N = \frac{1}{N} \sum_{i=1}^{N} (1 - p(W_i))\, p(W_i)$$

$$\frac{1}{N-1} \int_W (1 - p(w))\, p(w)\, P_W(dw) \text{ almost-surely.}$$

The limit above is finite, positive under Assumption 5.1 (that implies there exists $p_0 > 0$ such that $p(w) \geq p_0$). Additionally, we have with probability one

$$\frac{1}{N} \sum_{i=1}^{N} \frac{f(X_i)}{p(W_i)} \frac{g(X_i)}{p(W_i)} p(W_i)\,(1 - p(W_i)) \frac{1}{N-1} \int_{X \times W} f(x)\, g(x) \left(\frac{1}{p(w)} - 1 \right) P_{X,W}(dx, dw).$$

By virtue of Assumption 5.8, the latter integral is finite. Finally, observe that we almost-surely have

$$\Delta_{N,p}(f) \frac{1}{N-1} \Delta_p(f);$$

and the desired result follows. In particular notice that the limiting variance $V^2(f)$ is given by

$$V^2(f) := \int_{X \times W} f(x)^2 \left(\frac{1}{p(w)} - 1 \right) P_{X,W}(dx, dw) - \Delta_p(f)^2 \int_W (1 - p(w))\, p(w)\, P_W(dw);$$

which is strictly positive except in the degenerate case where $f(x) = p(w)$. Typically, this occurs when the inclusion probabilities are based directly on the variable of interest (or $W = c \cdot X$ for some $c \in \mathbb{R}$). Positivity of the operator results from Cauchy-Schwarz inequality. □

5.5.2 FCLT in the Poisson survey case

We shall prove Theorem 5.14, the statement of which is recalled below.

Theorem Suppose that Assumption 5.1, Assumption 5.8 and Assumption 5.11 hold, as well as the following conditions.

i) Lindeberg-Feller type condition: $\forall \varepsilon > 0$,

$$\mathbb{E}\left[(Z_{N,i})^2 \mathbb{I}\left\{ Z_{N,i} > \varepsilon \sqrt{N} \right\} \right] \frac{1}{N-1} 0;$$

with $Z_{N,i} := (\delta_i - p(W_i)) \sup_{f \in \mathcal{F}} \left| \frac{f(X_i)}{p(W_i)} - \Delta_{N,p}(f) \right|$.

ii) Uniform entropy condition: let $\mathcal{D}$ be the set of all finitely discrete probability measures defined in Section 5.1.2.3, and assume

$$\int_0^1 \sup_{Q \in \mathcal{D}} \sqrt{\log(N(\varepsilon \|\mathcal{H}\|_{2,Q}, \mathcal{F}, \|\cdot\|_{2,Q}))}\, d\varepsilon < \infty.$$

Then there exists a $\square_P$-equicontinuous Gaussian process G in $\ell^1(F)$ with covariance operator $\square$ given by Equation (5.4) such that

$$\mathbb{G}_{T_N}^p \Rightarrow G \text{ weakly in } \ell^1(F), \text{ as } N \to \infty.$$

Proof We essentially have to check hypotheses i) – iv) of Theorem 5.10.

Concerning hypothesis i), the Lindeberg-Feller condition can be written as

$$\frac{1}{N}\sum_{i=1}^{N} E\left[Z_{N,i}^2 \mathbf{1}\left\{|Z_{N,i}| > \varepsilon \sqrt{N}\right\}\right] \to 0 \quad \text{for every } \varepsilon > 0;$$

which reduces to $E\left[Z_{N,i}^2 \mathbf{1}\left\{|Z_{N,i}| > \varepsilon \sqrt{N}\right\}\right] \to 0$ by exchangeability of the components. This corresponds to condition i) in Theorem 5.14 above.

Recall that Assumption 5.8 stipulates the envelope of class F is square-integrable function H and that under Assumption 5.1, there is some $p_\square > 0$ such that for all $i \in U_N$, $p(W_i) \geq p_\square$. Hence, we have

$$\sum_{i=1}^{N} p(W_i)(1 - p(W_i)) \leq p_\square (N - E(n)) = p_\square N\left(1 - \frac{E(n)}{N}\right).$$

as well as

$$\square_{N,p}(f) - \square_{N,p}(H) \leq \frac{1}{p_\square}\frac{1}{N - E(n)}\sum_{i=1}^{N} H(X_i) - 1.$$

We thus obtain:

$$\begin{aligned}\sup_{f \in F}\left(\frac{f(X_i)}{p(W_i)} - \square_{N,p}(f)\right)^2 &\leq 2\sup_{f \in F}\left(\frac{f(X_i)}{p(W_i)}\right)^2 + \sup_{f \in F}\square_{N,p}(f)^2\Big) \\ &\leq 2\left(\left(\frac{H(X_i)}{p(W_i)}\right)^2 + \square_{N,p}(H)^2\right).\end{aligned}$$

Set

$$\begin{aligned}G_{1,i} &= |\square_i - p(W_i)|\left(\frac{H(X_i)}{p(W_i)} + \square_{N,p}(H)\right); \\ G_{2,i} &= (\square_i - p(W_i))^2\left(\left(\frac{H(X_i)}{p(W_i)}\right)^2 + \square_{N,p}(H)^2\right).\end{aligned}$$

Observe that it is thus sufficient to check that

$$E_{P_{X,W}}\left[E_{T_N}\left[G_{2,i}\mathbf{1}\left\{G_{1,i} > \varepsilon\sqrt{N}\right\}\right]\right] \to 0 \quad \text{for every } \varepsilon > 0.$$

Condition ii) can be checked immediately by noticing that, in the case of the Poisson process, the equicontinuity condition becomes

$$\sup_{\rho(f,g)<\delta} \sum_{i=1}^{N} E\left[(Z_{N,i}(f) - Z_{N,i}(g))^2\right] =$$

$$\sup_{\rho(f,g)<\delta} \frac{1}{N} \sum_{i=1}^{N} E_{P_{X,W}}\left[\frac{f(X_i) - g(X_i)}{p(W_i)} - \Delta_{N,p}(f) + \Delta_{N,p}(g)\right]^2 p(W_i)(1 - p(W_i))$$

$$\leq \sup_{\rho(f,g)<\delta} \frac{C_p}{4p^2} E_{P_{X,W}}\left[(f(X_1) - g(X_1))^2\right] \to 0, \text{ as } \delta \to 0.$$

In practice, condition iii) is checked in an easier manner by using the uniform entropy condition given here, see also Lemma 2.11.6 in van der Vaart and Wellner (1996).

Finally, condition iv) is a direct consequence of Lemma 5.13. □

5.5.3 Approximation result

We shall prove Lemma 5.16, the statement of which is recalled below.

Lemma Let R_N and T_N be two sampling designs and assume that T_N is entirely characterized by its first order inclusion probabilities, $\pi(T_N)$. Then, the empirical processes $\tilde{G}_{T_N}^{\pi(T_N)}$ and $\tilde{G}_{R_N}^{\pi(T_N)}$ valued in $\ell^1(\mathcal{F})$ satisfy the relationships:

$$d_{BL}\left(\tilde{G}_{T_N}^{\pi(T_N)}, \tilde{G}_{R_N}^{\pi(T_N)}\right) \leq \|R_N - T_N\|_1 \leq \sqrt{2D(T_N, R_N)}.$$

Consequently, if the sequences $(R_N)_{N\geq 1}$ and $(T_N)_{N\geq 1}$ are such that $\|R_N - T_N\|_1 \to 0$ or $D(T_N, R_N) \to 0$ as $N \to \infty$ and if there exists a Gaussian process G such that

$$d_{BL}(\tilde{G}_{T_N}^{\pi(T_N)}, G) \xrightarrow[N\to\infty]{} 0,$$

then we also have

$$d_{BL}(\tilde{G}_{R_N}^{\pi(T_N)}, G) \xrightarrow[N\to\infty]{} 0.$$

The same result holds true when replacing $\tilde{G}_{T_N}^{\pi(T_N)}$ and $\tilde{G}_{R_N}^{\pi(T_N)}$ by $G_{T_N}^{\pi(T_N)}$ and $G_{R_N}^{\pi(T_N)}$ respectively.

Proof Let $b \in BL_1(\ell^1(\mathcal{F}))$. We have

$$E_{T_N}\left[b\left(G_{R_N}^{\pi(T_N)}\right)\right] - E_{R_N}\left[b\left(G_{T_N}^{\pi(T_N)}\right)\right] = \sum_{s\in\mathcal{P}(U_N)} R_N(s)\, b\left(G_{T_N(s)}^{\pi(T_N)}\right)$$

$$- \sum_{s\in\mathcal{P}(U_N)} T_N(s) b\left(G_{R_N(s)}^{\pi(T_N)}\right)$$

$$\leq \sum_{s\in\mathcal{P}(U_N)} |T_N(s) - R_N(s)|$$

because b is bounded by 1 and

$$b\left(G_{T_N(s)}^{\pi(T_N)}\right) = b\left(G_{R_N(s)}^{\pi(T_N)}\right);$$

their expressions depending on the first order inclusion probabilities $\pi(T_N)$ solely.

The last inequality follows from the usual inequality between the total variation metric and the entropy (Berger, 1998, Lemma 2 p.219). □

5.5.4 FCLT in the rejective survey case

We shall prove Corollary 5.18, the statement of which is recalled below.

Corollary Suppose that Assumption 5.1, Assumption 5.8, Assumption 5.11 and conditions i) and ii) of Theorem 5.14 are satisfied. Then, there exists a ρ-equicontinuous Gaussian process G in $\ell^1(\mathcal{F})$ with covariance operator Γ given by Equation (5.4) such that

$$G_{R_N}^{\pi(R_N)} \Rightarrow G \text{ weakly in } \ell^1(\mathcal{F}), \text{ as } N \to \infty.$$

Proof Following in the footsteps of Hàjek (1964), in the rejective sampling situation where $p(W_i) = p_i$ for $i \in \{1, \dots, N\}$, we have

$$\max_{1 \le i \le N} \left(\frac{p_i}{R_i} - 1\right) \to 0.$$

We thus have

$$\max_{1 \le i \le N} \left(\frac{R_i}{p_i} - 1\right) \to 0$$

under the hypothesis that Assumption 5.1 is fulfilled by the p_i's. Then, we can write

$$\begin{aligned} G_{R_N}^{p(W)} f - G_{R_N}^{\pi(R_N)} f &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{\pi_i}{p(W_i)} - \frac{\pi_i}{R_i}\right) f(X_i) \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\pi_i}{p(W_i)} \left(1 - \frac{p_i}{R_i}\right) f(X_i); \end{aligned}$$

and

$$\begin{aligned} &\sup_{b \in B L_1(\ell^1(\mathcal{F}))} E\left[b\left(\sup_{f \in \mathcal{F}} G_{R_N}^{p(W)} f\right) - b\left(\sup_{f \in \mathcal{F}} G_{R_N}^{\pi(R_N)} f\right)\right] \\ &\le E\left[\sup_{f \in \mathcal{F}} G_{R_N}^{p(W)} f - \sup_{f \in \mathcal{F}} G_{R_N}^{\pi(R_N)} f\right] \\ &\le E\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\pi_i}{p(W_i)} \left(1 - \frac{p_i}{R_i}\right) f(X_i)\right)\right] \end{aligned}$$

$$\S\ ?\frac{1}{\overline{N}}\sum_{i=1}^{N}\left(\frac{R_i}{p(W_i)}\ \square\ 1\right)H(X_i);$$

which quantity vanishes asymptotically under Assumption 5.1, according to Theorem 5.1, Equations (5.7) and (5.26) in Hàjek (1964, p. 1508-1510).

The desired convergence is finally established by combining this result with Theorem 5.17 and the functional version of Slutsky's theorem (see Theorem 3.4 in Resnick, 2007 for instance). □

5.5.5 CLT for Hadamard differentiable functionals

We shall prove Theorem 5.21, the statement of which is recalled below.

Theorem Suppose that the assumptions of Theorem 5.14 hold and that functional $T : L\ Ä\ `^1(F) \mathrel{—} R^q$ is Hadamard differentiable at P with differential dT_P and influence function $T^{(1)}(x; P)$. Then, as $N \mathrel{—} +1$, we have:

$$?\ \overline{N}\ \square\ T(P_{R_N}^{\square(R_N)})\ \square\ T(P_N)\ \square\ Ò\ dT_P.G;$$

where G is a Gaussian process with covariance operator □, as in Equation (5.4).

Proof The idea is essentially to apply the Hadamard differentiability property to the sequence $h_N = ?\overline{N}(P_{R_N}^{\square(R_N)}\ \square\ P_N) =: G_{R_N}^{\square(R_N)}$, which converges to $h = G$ in $`^1(F)$ and $t_N = ?\frac{1}{\overline{N}} \mathrel{—} 0$. We thus have, as $N \mathrel{—} +1$:

$$?\ \overline{N}\ \square\ T(P_{R_N}^{\square(R_N)})\ \square\ T(P_N)\ \square\ = ?\ \overline{N}\ \square\ T(P_N + ?\frac{1}{\overline{N}}h_N)\ \square\ T(P_N)\ \square \mathrel{—} dT_P.G.$$

□

6

TAIL INDEX ESTIMATION BASED ON SURVEY DATA

In the previous chapter we investigated the asymptotic behavior of a variant of the Horvitz-Thompson type of the traditional empirical process for survey schemes of the Poisson type. The ensuing Functional Central Limit Theorems proved particularly convenient to derive asymptotic properties for numerous families of estimators, as was suggested in Section 5.4. Unfortunately, when interested in extreme phenomena like the very high exposure to some food chemical, these results are no longer sufficient and estimators need to be analyzed one by one. In extreme value theory, the survey design is usually ignored and the ensuing statistics, already suffering from the rarity of tail observations, possibly exhibit an additional bias due to the omission of the sampling phase. Whereas asymptotic analysis of the Horvitz-Thompson estimator (Horvitz and Thompson, 1951) has been the subject of much attention, in particular in the context of mean estimation and regression (see Hàjek, 1964; Rosen, 1972; Robinson, 1982; Gourieroux, 1987; Deville and Särndal, 1992; Berger, 1998 for instance), and the last few years have witnessed significant progress towards a comprehensive functional limit theory for the assessment of distribution functions (Gill et al., 1988; Breslow and Wellner, 2007, 2008; Breslow et al., 2009; Saegusa and Wellner, 2011; Bertail et al., 2013), no result on tail estimation has been documented in the survey sampling literature yet. In a modest attempt to start filling this gap, we make our contribution to the elaboration of an extreme value theory for survey data by focusing on a modified (Horvitz-Thompson) version of the most celebrated Hill estimator. This new statistic assesses the extreme value index of univariate heavy-tailed distributions, which essentially indicates to what extent extreme events are rare, while accounting for the sampling design by means of which the data have been collected. Therefore, contrary to the standard Hill estimator, it corrects the bias induced by the survey plan. Its consistency is established for any type of sampling scheme that fulfills some adequate assumptions on the first and second order inclusion probabilities. Then, following in the footsteps of Hàjek (1964) and along the lines of the previous chapter, its asymptotic normality is investigated for Poisson-like survey designs. The results presented in this chapter originate from a paper written in collaboration with P. Bertail (Université Paris X, France) and S. Clémençon (Télécom ParisTech, France) and has been submitted for publication. Regrettably, our analysis does not encompass complex designs such as that of the INCA2 database

yet, hence no concrete application to dietary risk analysis is provided in this chapter. Such refinements will hopefully be the object of further research in the near future.

The chapter is structured as follows. Basics about survey sampling and tail index estimation in the standard iid setup are briefly recalled and notations are set up in Section 6.1. In Section 6.2, we describe at length the proposed modification of the Hill estimator in the context of a general sampling plan and prove its consistency. Its asymptotic normality is investigated next in Section 6.3. Finally, practical issues such as the selection of an optimal number of largest observations on which to base the estimation are discussed in Section 6.4, together with illustrative numerical experiments. Technical proofs are deferred to Section 6.6.

6.1 background and preliminaries

We first recall the crucial notions in survey sampling that are extensively used in the subsequent analysis, as well as basic concepts of heavy-tail modeling, including Pareto-type distributions and standard strategies for statistical estimation of the related parameters. For the sake of clarity, most notations are kept identical to those in Chapter 5: the Dirac mass at $x \in \mathbb{R}$ is denoted by δ_x and the indicator function of any event E by $\mathbb{I}\{E\}$. We also denote by $\#E$ the cardinality of any finite set E, and by $\mathcal{P}(E)$ its power set. The (left-continuous) general inverse of any non-decreasing function $H : (a,b) \to \mathbb{R}$, $-\infty \leq a < b \leq +\infty$, is denoted by

$$H^{-}(x) := \inf\{y \in (a,b) : H(y) \geq x\},$$

$x \in \mathbb{R}$, with the convention that the infimum over an empty set is $-\infty$. When dealing with some multivariate distribution function $H : \mathbb{R}^d \to \mathbb{R}$ with marginals $H_1, \ldots, H_d$, we shall write $H^{-}(x) := \left(H_1^{-}(x_1), \ldots, H_d^{-}(x_d)\right)$ for any $x := (x_1, \ldots, x_d) \in \mathbb{R}^d$. Finally, the minimum (resp. maximum) of two real numbers x and y is denoted by $x \wedge y$ (resp. $x \vee y$).

6.1.1 Survey sampling

In this section we recall a few essential definitions and set out the notations relative to survey sampling. More details about the ins and outs of these concepts can be found in Section 5.1.1 of Chapter 5.

6.1.1.1 Population, sample, inclusion probabilities and indicators

Here and throughout, we consider a finite population of size $N \geq 1$, denoted by $\mathcal{U}_N := \{1, \ldots, N\}$. We call a sample of (possibly random) size $n \leq N$, any subset $s := \{i_1, \ldots, i_{n(s)}\}$ in $\mathcal{P}(\mathcal{U}_N)$ with cardinality $n =: n(s)$ less than N. A sampling

scheme (design/ plan) without replacement is determined by a probability distribution R_N on the set of all possible samples $s \in \mathcal{P}(U_N)$. For any $i \in \{1,\dots,N\}$, the following quantity, generally called (first order) inclusion probability,

$$\pi_i(R_N) := P_{R_N}(i \in S),$$

is the probability that the unit i belongs to a random sample S drawn from distribution R_N. In vectorial form, we shall write $\pi(R_N) := (\pi_1(R_N),\dots,\pi_N(R_N))$. First order inclusion probabilities are assumed to be strictly positive in the subsequent analysis: $\forall \in \{1,\dots,n\}$, $\pi_i(R_N) > 0$. Additionally, the second order inclusion probabilities are denoted by

$$\pi_{i,j}(R_N) := P_{R_N}\left((i,j) \in S^2\right),$$

for any $i \neq j$ in $\{1,\dots,N\}^2$. When no confusion is possible, we shall fail to mention the dependence in R_N when writing the first/ second order probabilities of inclusion. The information related to the observed sample $S = \{1,\dots,N\}$ is encapsulated by the random vector $I := (I_1,\dots,I_N)$, where

$$I_i = \begin{cases} 1 & \text{if } i \in S \\ 0 & \text{otherwise.} \end{cases}$$

The distribution of the sampling scheme R has 1-dimensional marginals that correspond to the Bernoulli distributions $\mathcal{B}(\pi_i)$, $1 \leq i \leq N$, and covariance matrix given by

$$\Delta_N := \left(\pi_{i,j} - \pi_i \pi_j\right)_{1 \leq i,j \leq N}.$$

Notice incidentally that, equipped with these notations, we have $\sum_{i=1}^{N} I_i = n(S)$.

The superpopulation model we consider here stipulates that a real-valued random variable X with distribution P and cdf F is observable on the population U_N, i.e. $X_1,\dots,X_N$ are iid realizations drawn from P. In practice, it is customary to determine the first order inclusion probabilities as a function of an auxiliary variable, which is observed on the entire population. Here, it is denoted by W with distribution P_W. Hence, for all $i \in \{1,\dots,N\}$ we can write $\pi_i = \pi(W_i)$ for some link function $\pi(.)$. When W and X are strongly correlated, thus proceeding helps select more informative samples and subsequently reduce the variance of estimators (we refer to Section 5.1.1.3 of Chapter 5 for a more detailed discussion on the use of auxiliary information in survey sampling). One may refer to Cochran (1977); Gourieroux (1981); Deville (1987) for accounts of survey sampling techniques.

6.1.1.2 A crucial example: the Poisson survey scheme

Though of extreme simplicity, the Poisson scheme (without replacement) plays a crucial role in sampling theory, insofar as it can be used to approximate a wide range of survey plans. This is indeed a key observation to establish general asymptotic results in the survey context, see Hàjek (1964), Chapter 5 and Section 6.3 of the

present chapter. For such a plan, denoted by T_N, the δ_i's are independent Bernoulli random variables with parameters $p_1, \ldots, p_N$ in $(0,1)$. Thus, the first order inclusion probabilities fully characterize such a plan. Observe in addition that the size $n(S)$ of a sample generated this way is random and goes to infinity as $N \to +\infty$ with probability one, provided that $\min_{1 \leq i \leq N} p_i$ remains bounded away from zero.

6.1.1.3 The Horvitz-Thompson empirical measure

We recall that the Horvitz-Thompson estimator of the empirical measure

$$P_N := \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i}$$

based on the survey data described above is defined as follows (Horvitz and Thompson, 1951):

$$P_{R_N}^{(R_N)} := \frac{1}{N} \sum_{i=1}^{N} \frac{\delta_i}{p_i} \delta_{X_i} = \frac{1}{N} \sum_{i \in S} \frac{1}{p_i} \delta_{X_i},$$

where the subscript R_N stipulates that the vector $p := (p_1, \ldots, p_N)$ is in correspondence with a sample S drawn at random from distribution R_N, and the superscript (R_N) indicates that the inclusion probabilities used in the formula are those of the design R_N. When there is no ambiguity, we shall simplify notations and write $P_N^{(p)}$ instead of $P_{R_N}^{(R_N)}$. We highlight the fact that, conditional on the complete set of observations $\{(X_i, W_i),\ 1 \leq i \leq N\}$, the latter is an unbiased estimator of P_N, although it is not a probability measure. Its (pointwise) consistency and asymptotic normality are established in Robinson (1982) and Berger (1998). Limit results of functional nature are established in Gill et al. (1988) for specific biased sampling models (see also Breslow and Wellner, 2007, 2008; Saegusa and Wellner, 2011; Bertail et al., 2013). The weighted quantity $F_N^{(p)}(x) := P_N^{(p)}(]-\infty, x])$ is naturally different from the empirical cumulative distribution function of the observations, $F_n(x) := n^{-1} \sum_{i \in S} \mathbb{I}\{X_i \leq x\}$ namely, whose asymptotic behavior is investigated in Bonnéry et al. (2011). It is then straightforward to deduce the following (unbiased) estimate of the probability of exceedance $\bar{F}(x) := P(X > x)$, $x \in \mathbb{R}$, given by:

$$\bar{F}_N^{(p)}(x) := \frac{1}{N} \sum_{i=1}^{N} \frac{\delta_i}{p_i} \mathbb{I}\{X_i > x\} = \frac{1}{N} \sum_{i \in S} \frac{1}{p_i} \mathbb{I}\{X_i > x\}. \tag{6.1}$$

6.1.2 Tail index inference - the Hill estimator

In a wide variety of situations, it is appropriate to assume that a statistical population is described by a heavy-tailed probability distribution (the field of heavy-tail analysis is well depicted in Resnick, 2007). A distribution with Pareto-like right tail is any probability measure P on $\mathbb{R}$ with cdf F such that for all $x \in \mathbb{R}$,

$$1 - F(x) = \bar{F}(x) = x^{-1/\alpha} L(x); \tag{6.2}$$

where $\gamma > 0$ is the extreme value index (EVI) of distribution P and $L(x)$ is a slowly varying function, i.e. a function such that $L(tx)/L(x) \to 1$ as $x \to +\infty$ for all $t > 0$. Notice that instead of the EVI, focus is often on $\alpha := 1/\gamma$, the tail index of the distribution P. Functions of the form introduced in Equation (6.2) are said to be regularly varying with index $-1/\gamma$; the set of such functions is denoted by $R_{-1/\gamma}$. One may refer to Bingham et al. (1987) for an account of the theory of regularly varying functions. The Hill estimator (Hill, 1975) provides a popular way of estimating the EVI γ. Its asymptotic behavior and the practical issues related to its computation are well-documented in the literature, see for instance Resnick (2007, Chapter 4) and the references therein. Given an iid population $X_1, \ldots, X_N$ of size $N > 1$ drawn from P and $K \in \{1, \ldots, N-1\}$, the largest observations, it is written

$$H_{K,N} := \frac{1}{K} \sum_{i=1}^{K} \log \left(\frac{X_{N-i+1,N}}{X_{N-K,N}} \right), \tag{6.3}$$

where $X_{1,N} \leq \cdots \leq X_{N,N}$ denote the order statistics related to the population. Whereas the theory has been extensively developed in the case where the observations are independent and identically distributed (including questions related to the choice of K in Equation (6.3)), to the best of our knowledge the Hill procedure has received no attention when data arise from a general survey. We point out that there exist alternative methods for tail index or EVI estimation, refer for instance to Beirlant et al. (2004, Chapter 4) for further details. The argument of the subsequent analysis paves the way for studying extensions of such techniques in the context of survey sampling models.

6.2 the hill estimator in survey sampling

Placing ourselves in the framework described in Section 6.1.1, we shall denote by $X_{1,n} \leq \cdots \leq X_{n,n}$ the order statistics related to the survey sample $(X_{i_1}, \ldots, X_{i_n})$, where $n = n(S)$ may be random. When unit j is such that $X_j = X_{i,N}$, the i-th largest observation in the population, $1 \leq i, j \leq N$, its inclusion indicator and probability are denoted by $\delta_{i,N} = \delta_j$ and $\pi_{i,N} = \pi_j$ respectively. Similarly, we write $\pi_{i,n} := \pi_j$ when $X_{i,n} = X_j$, $1 \leq i, j \leq n$. As a general rule, indexes in uppercase shall designate the full population, as opposed to those in lowercase, which shall refer to the sample. We assume in addition that the distribution function F has the semi-parametric form set out in Equation (6.2) with unknown parameter $\gamma > 0$ and, for the sake of simplicity, that its support is included in $(0, +\infty)$. Because it is destined to be extensively used in the sequel, we also introduce the tail quantile function, written for all $x \in [1, +\infty)$ as

$$U(x) := F^{-} \left(1 - \frac{1}{x} \right).$$

Its empirical and Horvitz-Thompson equivalents are respectively denoted by

$$U_N(x) := \overline{F}_N^{\leftarrow}\left(1 \wedge \frac{1}{x}\right) \quad \text{and} \quad U_N^{\pi}(x) := (\overline{F}_N^{\pi})^{\leftarrow}\left(1 \wedge \frac{1}{x}\right).$$

Notice that when $F \in \mathrm{RV}_{1/\gamma}$, the corresponding tail quantile function U is also regularly varying with index γ (Beirlant et al., 2004, Sections 2.3.2 and 2.9.3). The goal pursued here is to estimate the tail parameter γ based on the survey data $X_{i_1}, \ldots, X_{i_n}$ and the sampling plan R_N.

6.2.1 The Horvitz-Thompson variant of the Hill estimator

Notice first that, under the heavy-tail assumption above, we have:

$$\gamma = \lim_{x \to \infty} \int_x^{a+1} \frac{\overline{F}(u)}{\overline{F}(x)} \frac{du}{u}; \tag{6.4}$$

see Beirlant et al. (2004, Section 2.6) for instance. In the case of the iid population $X_1, \ldots, X_N$ drawn from P, one classically recovers the celebrated Hill estimator by substituting F with the empirical cdf F_N in Equation (6.4) and taking

$$x = U_N(N/K) = X_{N-K;N}$$

for some number $1 \leq K \leq N-1$ of largest observations, supposedly representative of the tail of the distribution. Indeed, we have:

$$\begin{aligned}
\int_{X_{N-K;N}}^{a+1} \frac{\overline{F}_N(u)}{\overline{F}_N(X_{N-K;N})} \frac{du}{u} &= \prod_{i=1}^{K} \int_{X_{N-i;N}}^{a} \frac{\overline{F}_N(u)}{\overline{F}_N(X_{N-K;N})} \frac{du}{u} \\
&= \prod_{i=1}^{K} \frac{i}{K} \left(\log X_{N-i+1;N} - \log X_{N-i;N}\right) \\
&= \frac{1}{K} \prod_{i=1}^{K} \log \frac{X_{N-i+1;N}}{X_{N-K;N}} \\
&= H_{K;N}.
\end{aligned}$$

Consistency of this estimator is classically guaranteed as soon as $K \to +\infty$ and $K = o(N)$ when $N \to +\infty$, see Mason (1982). In this case, the empirical threshold $X_{N-K;N}$ (equivalent in probability to $U(N/K)$) goes to infinity (again in probability) as $N \to +\infty$.

Going back to the survey data situation, one may naturally replace $U(N/K)$ by $U_N^{\pi}(N/K)$ and build a plug-in estimate of the EVI γ based on the Horvitz-Thompson estimator given in Equation (6.1) of the tail probability $\overline{F}(x)$. Observe that by definition $U_N^{\pi}(N/K)$ corresponds to one of the observations in the sample, say $X_{i^\star}$ with rank $\ell \in \{1, \ldots, n\}$. To this ℓ obviously corresponds an index $k \in \{0, \ldots, n-1\}$ such that $\ell = n - k$, implying $X_{n-k;n} = U_N^{\pi}(N/K)$, the Horvitz-Thompson estimator of

the quantile of order $1 - K/N$. We denote by π_N the map linking k to K in U_N under the sampling scheme R_N:

$$\pi_N : \begin{array}{ccc} \{1,\dots,N-1\} & \longrightarrow & \{1,\dots,n-1\} \\ K & \longmapsto & k := \pi_N(K) \end{array},$$

where

$$\pi_N(K) := n - \inf\left\{ i \in \{1,\dots,n-1\} : \sum_{j=1}^{i} \frac{1}{\pi_{j,n}} \geq N - K \right\}. \tag{6.5}$$

This leads to the quantity:

$$\begin{aligned} p_\pi &= \int_{X_{n-k,n}}^{a+1} \frac{\bar{F}_{R_N}^{(\pi)}(u)}{\bar{F}_{R_N}^{(\pi)}(X_{n-k,n})} \frac{du}{u} = \sum_{i=1}^{k} \int_{X_{n-i+1,n}}^{a} \frac{\bar{F}_{R_N}^{(\pi)}(u)}{\bar{F}_{R_N}^{(\pi)}(X_{n-k,n})} \frac{du}{u} \\ &= \sum_{i=1}^{k} \left(\frac{\sum_{j=1}^{i} \frac{1}{\pi_{n-j+1,n}}}{\sum_{j=1}^{k} \frac{1}{\pi_{n-j+1,n}}} \right) \cdot (\log(X_{n-i+1,n}) - \log(X_{n-i,n})) \\ &= \sum_{j=1}^{k} \frac{1}{\pi_{n-j+1,n}} \Bigg/ \sum_{i=1}^{k} \frac{1}{\pi_{n-i+1,n}} \log \frac{X_{n-i+1,n}}{X_{n-k,n}} \\ &=: H_{k,n}^{\pi}. \end{aligned}$$

Hence, k is to the sample what K is to the population: the number of upper values on which the estimation should rely. Notice that we may also write

$$H_{k,n}^{\pi} = \sum_{j=1}^{k} \frac{\pi_{N-j+1,N}}{\pi_{N-j+1,N}} \Bigg/ \sum_{i=1}^{k} \frac{\pi_{N-i+1,N}}{\pi_{N-i+1,N}} \log \frac{X_{N-i+1,N}}{X_{N-K,N}} =: H_{K,N}^{\pi}, \tag{6.6}$$

where K is the chosen number of largest observations in the population from which k was constructed. Observe that π_N in Equation (6.5) is a surjective, non-injective random map, which suggests that the subsequent asymptotic analysis better rely on some appropriately chosen K and its random image $k := \pi_N(K)$ rather than the contrary. Since in practice only k can be computed from the X_i's and π_i's, $i \in S$, the total population being partly unobserved, considerations about the choice of an appropriate k are discussed in detail in Section 6.4. From this point forward, the Horvitz-Thompson Hill estimator shall be written $H_{K,N}^{\pi}$ with $K \in \{1,\dots,N-1\}$ held fixed.

6.2.2 Consistency of $H_{K,N}^{\pi}$

Here we investigate the limit properties of the estimator $H_{K,N}^{\pi}$ as N and n simultaneously go to infinity, with $n \ll N$. The following assumptions, related to the sample design, shall be involved in the asymptotic analysis.

Assumption 6.1 There exist $\varepsilon > 0$ and $N_0 \in \mathbb{N}^*$ such that for all $N \geq N_0$ and $i \in U_N$,

$$\pi_i > \varepsilon.$$

Assumption 6.2 There exists $\ell > 2$ such that we almost-surely have $\forall N \geq 1$,

$$\max_{1 \leq i,j \leq N} \left| \Delta_{i,j} - \pi_i \pi_j \right| \leq \frac{\ell}{n}.$$

Assumption 6.1 guarantees that first order inclusion probabilities do not vanish asymptotically, while Assumption 6.2 corresponds to the situation where the second order inclusion probabilities are not too different from those in the case of independent sampling (it is thus fulfilled by the Poisson design, see Section 6.1.1.2).

Remark 6.3 – On Assumption 6.1 and Assumption 6.2. The two assumptions introduced herein-before are rather mild and are fulfilled in a wide variety of situations. Indeed, Assumption 6.2 is standard in asymptotic analysis of sampling techniques, see Hartley and Rao (1962) and Hàjek (1964) for instance. As for Assumption 6.1, recall that $\pi_i = \pi(W_i)$ with W an auxiliary variable and $\pi(.)$ a link function. Then, it is fulfilled as soon as $\pi(.)$ is continuous and the support of P_W is a compact subset of $(\mathbb{R}_+^*)^d$, where d denotes the dimension of the random vector W and $\mathbb{R}_+^*$ the set of positive real numbers.

Given this framework, following in the footsteps of Resnick (2007, Section 4.4.1), the consistency of $H_{K,N}^{\alpha}$ can be handled by exploiting the properties of regularly varying distributions. Indeed, under the heavy-tail assumption in Equation (6.2), provided that $K \to +\infty$ and $K/N \to 0$ as $N \to +\infty$, we have

$$\frac{N}{K} P\left[\frac{X}{U(N/K)} \in . \right] \xrightarrow{v} \nu_{1/\alpha}(.)$$

in the space of Radon measures on $(0, +\infty]$. There, "$\xrightarrow{v}$" stands for the vague convergence of measures[1] and $\nu_{1/\alpha}(.)$ is such that for all $x > 0$, $\nu_{1/\alpha}(x, +\infty] = x^{-1/\alpha}$ (see Resnick, 2007, Theorem 3.6 for instance). Its empirical counterpart in the population, usually called the tail empirical measure, is defined as follows:

$$\nu_N := \frac{1}{K} \sum_{i=1}^{N} \pi_i \delta_{X_i / U(N/K)}.$$

1. Recall that, in the space of non-negative Radon measures on $(0, +\infty]$, a sequence $(\mu_m)_{m \geq 1}$ is said to converge vaguely to μ iff for any compactly supported continuous function $h : (0, +\infty] \to \mathbb{R}$, we have: $\int_0^{+\infty} h(x) \mu_m(dx) \to \int_0^{+\infty} h(x) \mu(dx)$ as $m \to +\infty$.

When replacing $U(N/K)$ by its estimate $U_N(N/K)$ in the expression above and assuming that $K = K(N) \to +\infty$ where $K/N \to 0$ as $N \to +\infty$, it can be shown that it converges to $\gamma^{1/\alpha}$ in probability (Resnick, 2007, Equation (4.21)). Since we have

$$H_{K,N} = \int_1^{\infty} \Big[\frac{X_{N-K,N}}{U(N/K)} \Big]^{\alpha} \frac{dx}{x}$$

and

$$\gamma = \int_1^{\infty} \gamma^{1/\alpha} (x, +\infty] \frac{dx}{x};$$

the asymptotic properties of γ_N naturally convey the consitency of the Hill estimator. Generalizing this result to the Horvitz-Thompson tail empirical measure

$$\gamma^{(\alpha)}_N = \frac{1}{K} \sum_{i=1}^{N} \frac{\pi_i}{\pi_i} \delta_{X_i / U_N(N/K)} = \frac{1}{K} \sum_{i \in PS} \frac{1}{\pi_i} \delta_{X_i / U_N(N/K)} \tag{6.7}$$

would then yield the theorem below (see the proof in Section 6.6). It reveals that, in regard to the asymptotic statistical estimation of the EVI γ, the Horvitz-Thompson variant of the Hill estimator $H^{(\alpha)}_{K,N}$ is consistent.

Theorem 6.4 – Consistency. Let $K = K(N)$ be a sequence of integers such that $K \to +\infty$ and $K/N \to 0$ as $N, n \to +\infty$. Provided that Assumption 6.1 and Assumption 6.2 are fulfilled, we then have, as N and n tend to $+\infty$:

$$H^{(\alpha)}_{K,N} \xrightarrow{P} \gamma. \tag{6.8}$$

6.3 asymptotic normality of $H^{(\alpha)}_{K,N}$

Whereas the consistency of the standard Hill estimator in Equation (6.3) can be proved for any sequence K going to infinity at a reasonable rate, asymptotic normality cannot be guaranteed at such a level of generality. Higher-order regular variation properties of the heavy-tail model in Equation (6.2) are required (de Haan and Peng, 1998; de Haan and Stadtmüller, 1996). More specifically, consider the hypothesis below, referred to as the Von Mises condition (Goldie and Smith, 1987).

Assumption 6.5 The regularly varying tail quantile function $U \in RV_{\gamma}$ with $\gamma > 0$ is such that there is a real parameter $\rho < 0$, referred to as the second order parameter, and a positive or negative function A with $\lim_{x \to +\infty} A(x) = 0$ such that for any $t > 0$,

$$\frac{1}{A(x)} \Big[\frac{U(tx)}{U(x)} - t^{\gamma} \Big] \xrightarrow[x \to \infty]{} t^{\gamma} \frac{t^{\rho} - 1}{\rho};$$

or equivalently

$$\frac{1}{A\left(\frac{1}{\bar{F}(x)}\right)}\left(\frac{\bar{F}(tx)}{\bar{F}(x)}-t^{-1/\gamma}\right)\to t^{-1/\gamma}\,\frac{t^{\rho/\gamma}-1}{\gamma\rho}.$$

This condition simply establishes some constraints about the slowly varying function $L(.)$ in Equation (6.2) to ensure its influence vanishes quickly enough not to interfere with the Pareto form $x^{-1/\gamma}$ of $\bar{F}$.

The limit distribution of the standard Hill estimator in Equation (6.3) has been investigated by means of Rényi's exponential representation of log-spacings under Assumption 6.5. Of course, this condition can hardly be checked in practice and the choice of the number of extremal observations is generally selected so as to minimize an estimate of the asymptotic mean squared error (MSE), see Section 6.4 and the references therein.

Remark 6.6 – On Hill and the tail empirical process. Other approaches than the Rényi decomposition in log-spacings have been developed to prove the asymptotic normality of the Hill estimator $H_{K;N}$ (see Resnick, 2007, Chapters 4 and 9 and the references therein). Along the lines of the study of empirical processes in Chapter 5, the version presented at length in Resnick (2007) involves the preliminary study of the tail empirical process

$$T_N := \sqrt{K}\left(\frac{N}{K}\,\bar{P}_N\left(X_{N-K;N}\,x^{1/\gamma},\infty\right]\right),\quad x>0,$$

from which the asymptotic properties of the Hill estimator are later deduced. We refer to Theorem 9.1 and Section 9.1.2 in Resnick (2007) for more details on this seemingly simple but actually quite intricate procedure.

Unfortunately, contrary to the classical empirical process, the asymptotic properties of T_N cannot be extended to its Horvitz-Thompson equivalent. This is essentially due to the fact that the population U_N contains a finite number of observations. Hence, there is always a maximum $X_{N;N}<\infty$ bounding P_N and sampling fails to distinguish between a distribution with finite support such as those in the Weibull domain of attraction and a heavy-tailed distribution with no endpoint (see Chapter 2 for an introduction to extreme value theory and maximum domains of attractions). Actually, these arguments are exactly the same as those introduced when discussing the applicability of bootstrap in extreme value analysis. Indeed, as explained in Remark 5.7 of Chapter 5, survey sampling under a superpopulation model can be viewed as a generalization of weighted bootstrap. The interested reader may refer to Resnick (2007, Section 6.4) for a brief introduction to the difficulty of bootstrapping heavy-tailed phenomena.

In this section, we shall aim at proving first that under Poisson survey schemes, the Horvitz-Thompson version of the Hill estimator computed on the sample is asymptotically close to its standard version calculated over the entire population. This result is next extended to rejective sampling plans through a coupling argument, similar to that used in Hàjek (1964) and in Chapter 5.

6.3.1 The case of the Poisson survey scheme

In this section we assume that the vector $\square$ corresponds to that of a Poisson survey scheme, such as depicted in Section 6.1.1.2. The distribution of this design is denoted by T_N and the first order inclusion probabilities by $p_1; : : : ; p_N$. Under this setting, the Horvitz-Thompson variant of the Hill estimator is naturally denoted by $H^{p}_{K;N}$ and the $\square_i$'s in Equation (6.6) are to be replaced by the corresponding p_i's. In addition, just like we previously set $\square_i = \square(W_i)$, we write $p_i = p(W_i)$ for all $i \in U_N$ and p the Poisson link function. In keeping with the results obtained in Chapter 5, so as to prove the asymptotic normality of $H^{p}_{K;N}$, we shall require the p_i's to fulfill Assumption 6.1 with lower bound $p_{\square}$ and the auxiliary variable from which they are built to satisfy the condition below.

Assumption 6.7 The random vectors $W_1; : : : ; W_N$ are iid with continuous distribution P_W on $W \subset \mathbb{R}^d$, d-variate cdf F_W with marginals $F_{W_1}; : : : ; F_{W_d}$ and density f_W. The joint distribution of the entailed iid sequence $t(X_i; W_i); 1 \leq i \leq N$ is denoted by $P_{X;W}$ with corresponding cdf $F_{X;W}$.

The following result reveals that under the Poisson survey scheme, when based on the K largest values among the whole population $X_1; : : : ; X_N$, $H^{p}_{K;N}$ converges at the same rate ($1/\sqrt{K}$ namely) to the same limit distribution as $H_{K;N}$, up to a multiplicative term in the asymptotic variance induced by the sampling scheme. Further details about the convergence of the classical Hill estimator can be found e.g. in de Haan and Peng (1998, Theorem 1) and Resnick (2007, Section 9).

Theorem 6.8 – Limit distribution in the Poisson survey case. Suppose that Assumption 6.5 is fulfilled by the underlying heavy-tailed model and that Assumption 6.1 is satisfied by the considered sequence of Poisson inclusion probabilities $p_1; : : : ; p_N$, $N \geq 1$, constructed from some set of auxiliary variables as in Assumption 6.7. Further assume the conditions introduced herein-after.

i) The marginal cdf F is absolutely continuous with respect to the Lebesgue measure with density f.

ii) The joint cdf $F_{X;W}$ is absolutely continuous with Lebesgue-integrable density $f_{X;W}$ such that for all $(x;w)$ $P(0;+1]\times W$, $w=(w_1;\dots;w_d)$,

$$f_{X;W}(x;w) := c_{X;W}(F(x);F_{W_1}(w_1);\dots;F_{W_d}(w_d))\, f(x) \prod_{j=1}^{d} f_{W_j}(w_j);$$

for some copula density $c_{X;W}: R_+^* \times R^d \to R$ and $f_{W_1};\dots;f_{W_d}$ the marginal densities of the distribution of W.

iii) $c_{X;W}$ bounded in a neighborhood of $\{1\}\times[0;1]^d$.

Then, for

$$\sigma_p^2 := \int_W \frac{1}{p(w)}\, c_{X;W}(1;F_{W_1}(w_1);\dots;F_{W_d}(w_d)) \prod_{j=1}^{d} f_{W_j}(w_j)\, dw_1 \dots dw_d$$

and provided that $K \to +1$ as $N \to +1$ so that $\sqrt{K}A(N/K) \to \lambda$ for some constant $\lambda \in R$, we have the convergence in distribution as $N \to +1$:

$$\sqrt{K}\left(H_{K;N}^{p} - \frac{1}{1-\gamma}\right) \xrightarrow{d} N\left(\frac{\lambda}{1-\rho};\ \sigma^2 \sigma_p^2\right). \tag{6.9}$$

As can be seen by examining the proof of this theorem in Section 6.6, the limit result in Proof 6.6.2 can be obtained using the following decomposition:

$$\sqrt{K}\left(H_{K;N}^{p} - \frac{1}{1-\gamma}\right) = \underbrace{\sqrt{K}\left(r_{K;N}\, H_{K;N}^{p} - H_{K;N}\right)}_{Q_N^{(1)}} + \underbrace{\sqrt{K}\left(H_{K;N} - \frac{1}{1-\gamma}\right)}_{Q_N^{(2)}} + \underbrace{\sqrt{K}\left(\frac{1}{r_{K;N}} - 1\right)}_{Q_N^{(3)}}; \tag{6.10}$$

where

$$r_{K;N} := r_{K;N}(T_N;p) := \frac{1}{K}\sum_{i=1}^{K} \frac{U_{N-i+1;N}}{p_{N-i+1;N}}. \tag{6.11}$$

These three quantities are studied independently under the hypotheses required in Theorem 6.8. First, we show that $r_{K;N}$ converges to 1 in probability as N tends to 1. Combined with Rényi's decomposition in log-spacings of the Hill estimator (refer for instance to Beirlant et al., 2004, Section 4.4), this establishes the asymptotic convergence, in probability, of $Q_N^{(1)}$ to 0. It also implies that $Q_N^{(2)}$ is equivalent to $\sqrt{K}(H_{K;N} - \gamma)$, a well-known quantity which tends to a Gaussian distribution with expectation $\lambda/(1-\rho)$ and variance σ^2 under the second order condition stipulated in Assumption 6.5 (De Haan and Ferreira, 2006, Theorem 3.2.5). As for $Q_N^{(3)}$, we calculate its expectation and variance conditionally on the full vector of observations $\{(X_i;W_i);\ 1 \le i \le N\}$, yielding expressions where the randomness induced by the

survey scheme has been controlled. Then, following in the lines of a Lindeberg-Feller theorem for independent and non-identically distributed variables (Feller, 1971, Theorem 3, p.262), we exhibit some sufficient conditions (namely i), ii) and iii) in Theorem 6.8) under which the conditional variance has a finite limit in probability relative to $t(X_i; W_i)$; $1 \leq i \leq N$. Provided that they are fulfilled, $Q_N^{(3)}$ converges weakly to a centered Gaussian distribution with variance $\sigma^2(\sigma_p^2 - 1)$. This lets us consider $Q_N^{(2)}$ and $Q_N^{(3)}$ as independent random variables (one depends on the data and the other on the survey scheme). Thus, the limit distribution of their sum is simply the sum of their limit distributions, thereby yielding Proof 6.6.2.

Remark 6.9 – On the asymptotic variance. Looking at the variance term in Proof 6.6.2 of Theorem 6.8, we see that the influence of the survey scheme is encapsulated by the multiplicative term $\sigma_p^2 \bullet 1$. Ideally, we would like to have at our disposal inclusion probabilities for which σ_p^2 is as close to 1 as possible. In that case, the Horvitz-Thompson version of the Hill estimator would perform as well as its population equivalent. For some chosen expected sample size $n^{\alpha} := E(n(S))$, they would solve the following optimization program:

$$\min_{p(w)} \int_{w} \frac{1}{p(w)} c_{X;W}(1; F_{W_1}(w_1); \ldots; F_{W_d}(w_d)) \prod_{j=1}^{d} f_{W_j}(w_j)\, dw_1 \ldots dw_d - 1$$

$$\text{subject to} \sum_{i=1}^{N} p(W_i) = n^{\alpha};$$

provided the ensuing sequence $p_1; \ldots; p_N$ satisfies Assumption 6.1.

6.3.2 Extension to rejective sampling schemes

We now show how the result stated in Theorem 6.8 can be extended to an important class of survey plans, namely rejective sampling schemes. For the sake of clarity, we first provide a brief description of the latter, refer to Hàjek (1964) and Berger (1998) for further details.

Fix $n \leq N$ and consider a vector $(\pi_1; \ldots; \pi_N)$ of first order inclusion probability. The rejective sampling, sometimes referred to as conditional Poisson sampling (CPS in short), exponential design without replacement or maximum entropy design (Tillé, 2006), is the sampling plan R_N which picks samples of fixed size $n(S) := n$ in order to maximize the entropy measure

$$H(R_N) = -\sum_{s \in P(U_N):\, \#s = n} R_N(s) \log R_N(s)$$

subject to the constraint stipulating that its vector of first order inclusion probabilities coincides with $(\pi_1; \ldots; \pi_N)$. It can be implemented in two steps, as follows.

1. Draw a sample S with a Poisson sampling plan (without replacement), with properly chosen first order inclusion probabilities $(p_1, \ldots, p_N)$. The representation is called canonical if $\sum_{i=1}^{N} p_i = n$. In that case the relationships between p_i and π_i, $1 \leq i \leq N$, are established in Hàjek (1964).
2. If $n(S) \neq n$, then reject it and go back to step one, otherwise stop.

The vector $(p_1, \ldots, p_N)$ must be chosen in a way that the resulting first order inclusion probabilities coincide with $\pi_1, \ldots, \pi_N$, by means of a dedicated optimization algorithm, see Tillé (2006). The corresponding probability distribution is given by: $\forall s \in \mathcal{P}(U_N)$,

$$R_N(s) = \frac{T_N^p(s)\, \mathbb{I}_{\{\#s = n\}}}{\sum_{s' \in \mathcal{P}(U_N):\, \#s' = n} T_N^p(s')} \prod_{i \in s} p_i \prod_{i \notin s} (1 - p_i)\, \mathbb{I}_{\{\#s = n\}}.$$

Refer to Hàjek (1964, p. 1496) for more details on the p_i's.

Turning now to the extension of the result stated in Theorem 6.8 for the Poisson survey scheme to the case of rejective sampling, we introduce the following quantities: $\forall K \leq N$,

$$D_{K,N}(R_N, T_N) := r_{K,N}(R_N, \pi)\, H_{K,N}^{\pi} - r_{K,N}(T_N, p)\, H_{K,N}^{p}$$

and

$$r_{K,N}(R_N, T_N) := r_{K,N}(R_N, \pi) - r_{K,N}(T_N, p),$$

where $H_{K,N}^{\pi}$ (respectively $r_{K,N}(R_N, \pi)$) refers to the Horvitz-Thompson version of the Hill estimator (resp. of $r_{K,N}$ in Equation (6.11)) under rejective sampling, and $H_{K,N}^{p}$ (resp. $r_{K,N}(T_N, p)$) to its Poisson counterpart. The corresponding inclusion probabilities are denoted by $\pi_1, \ldots, \pi_N$ and $p_1, \ldots, p_N$ respectively. The ensuing approach follows in the footsteps of Hàjek (1964) and relies more specifically on the results displayed in Theorem 5.1, p.1508. Let us start by defining the quantities

$$d_N = \sum_{i=1}^{N} p_i(1 - p_i) \quad \text{and} \quad \bar{p}_N = \frac{1}{d_N} \sum_{i=1}^{N} p_i^2(1 - p_i).$$

We assume that both R_N and T_N fulfill Assumption 6.1 for minoring constants $\pi_\star$ and $p_\star$ respectively and that the Poisson inclusion probabilities further satisfy the following condition.

Assumption 6.10

$$\limsup_{N \to +\infty} \frac{1}{N} \sum_{i=1}^{N} p_i(T_N) < 1.$$

Notice that, in this situation, $d_N = o(1/K)$ and $\bar{p}_N$ is bounded. In addition, as shown in Hàjek (1964) (see p.1510 therein), the decomposition below holds for all $i \in \{1, \ldots, N\}$:

$$p_i - \pi_i = \frac{\bar{p}_N - p_i}{d_N} + o(1/d_N)\, p_i(1 - \pi_i).$$

This roughly means that the inclusion probabilities of the rejective sampling scheme are very close to those of the underlying Poisson design from which it was built. So close in fact that $D_{K;N}(R_N;T_N)$ and $r_{K;N}(R_N;T_N)$ asymptotically vanish. Therefore, as revealed by the following result, Theorem 6.8 also holds when the sample is constructed with a rejective plan.

Theorem 6.11 – Limit distribution in the rejective survey case. Suppose that all the conditions required in Theorem 6.8 hold together with Assumption 6.10. Then, for

$$\sigma_p^2 := \int_{\mathbb{R}^d} \frac{1}{p(w)}\, c_{X;w}(1;F_{W_1}(w_1);\ldots;F_{W_d}(w_d)) \prod_{j=1}^{d} f_{W_j}(w_j)\, dw_1 \ldots dw_d$$

and provided that $K \to +\infty$ as $N \to +\infty$ so that $\sqrt{K}A(N/K) \to 0$ for some constant $\beta \in \mathbb{R}$, we have the convergence in distribution as $N \to +\infty$:

$$\sqrt{K}(H_{K;N} - \gamma) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\frac{\beta}{1-\gamma}; \gamma^2 \sigma_p^2\right). \tag{6.12}$$

The proof of this theorem is available in Section 6.6. Notice that the limit variance does not depend on the inclusion probabilities $\pi_1;\ldots;\pi_N$, but on those of the underlying Poisson design, $p_1;\ldots;p_N$ namely. In that sense, this result is very similar to those obtained in Chapter 5: the asymptotic properties of the rejective sampling scheme are intricately linked to the Poisson plan with which it is coupled.

6.4 practical issues and illustrative experiments

6.4.1 On the choice of an optimal k

All results presented in the previous section depend on some appropriate number K of largest observations in the population $X_1;\ldots;X_N$. Unfortunately, the estimated tail quantile $X_{N-K;N}$ from which $H_{K;N}^{\gamma}$ is computed may not be included in the sample. Hence, we need to choose a number k of largest values in the sample to which we may associate some K that respects the necessary conditions for consistency and asymptotic normality to hold ($K = K(N) \to +\infty$ and $\sqrt{K}A(N/K) \to \beta \neq 1$ as $N \to +\infty$). Recall that we defined π_N^{γ} in Equation (6.5), a non-injective random map that assigns an index k in the sample to any index K in the population so that $X_{n-k;n} = U_N^{\gamma}(N/K)$. Setting

$$\hat{R}(k) := (\pi_N^{\gamma})^{-}(k) := \left\lfloor N \prod_{i=1}^{n-k} \frac{1}{\pi_{i;n}} \right\rfloor;$$

where r.s is the ceiling function, it is straightforward to show that the limit results stated in the sections above remains true for $\widehat{R}(k) \, H_{k;n} \; \square \; \square$ as $N; k \to +1$ so that $\widehat{R}(k) \, A(N=\widehat{R}(k)) \to \square$ almost-surely. This result can be naturally used to ground the construction of asymptotic Gaussian confidence intervals. The only work left is to find a suitable estimator p_p^2 of $\square_p^2$ such that, by virtue of Slutsky's Lemma combined with Theorem 6.4, the quantity $\widehat{R}(k) \, H_{k;n}^{\square} \; \square \; \square / H_{k;n}^{\square} \, p_p$ would be asymptotically pivotal, distributed as a standard Gaussian random variable as $N; k \to +1$.

In practice, choosing an optimal threshold $X_{N \square K;N}$ is already complicated in the iid case. Many techniques have been proposed in the literature, often based on the minimization of the MSE (see Danielsson et al., 2001; Gomes and Oliveira, 2001; Goegebeur et al., 2008 and the references therein). Since they involve in general the estimation of the second order parameter $\square$, which goes beyond the scope of our analysis, we leave such considerations for future research. In the meantime, we propose to simply rely on heuristics such as the stability of the Horvitz-Thompson version of the Hill estimator around the appropriate k.

6.4.2 Numerical experiments

As a complement to the theoretical results established in the previous section, we provide here some illustrations based on simulations. In particular, we consider a model that does not fulfill condition ii) in Theorem 6.8, which requires the absolute continuity of $F_{X;W}$. The encouraging empirical results we obtain nonetheless give hope that this assumption may be relaxed. Such desirable extensions are left for future research.

6.4.2.1 Experiment setting

Simulations were based on the following model, chosen for its simplicity in terms of both computation and interpretation:

$$X = \frac{(1 \square F_W(W))^{\square} \square 1}{\square}; \; \square \circ 0;$$

$$W \; ; \; TN(\square; \square_W^2; w_{\square}; w^{\square});$$

where X is the variable of interest, W the auxiliary information with cdf F_W and $TN(\square; \square_W^2; w_{\square}; w^{\square})$ refers to the truncated Normal distribution over $[w_{\square}; w^{\square}]$, with expectation $\square$ and variance $\square_W^2$. Under such a representation, the distribution of X is a General Pareto with scale parameter 1 and EVI $\square$, i.e. $F(x) = 1 \square (1 + \square x)^{\square 1=\square}$. This is a well-known family of distributions, the second order properties of which are easily derived (De Haan and Ferreira, 2006, Section 3.2). In particular, we have

$\gamma = \rho\beta$ and $A(x) = x^{\rho}\beta$. Following De Haan and Ferreira (2006, p.80), the optimal number of largest observations in the population is

$$K^{\gamma} = K^{\gamma}(N) = \left[\frac{N^{2\rho}}{2} \beta^{3} (1+\rho)^{2} \right]^{1/(2\rho+1)},$$

where $\lfloor x \rfloor$ is the integer part of x. It follows that $\sqrt{K^{\gamma}} A(N/K^{\gamma}) \to 0$ as $N \to +\infty$. Concerning the joint distribution of X and W, it is straightforward to see that

$$F_{X,W}(x,w) = F(x) \wedge F_W(w),$$

which means that the copula linking both marginals is the well-known singular copula $M(u,v) := u \wedge v$, $(u,v) \in [0,1]^2$. Unfortunately, it is not derivable on it entire support and condition ii) in Theorem 6.8 is not fulfilled here. However, as we shall see in the next subsection, this does not impede tail estimation.

For a given population $\mathcal{U}_N$ of size N, where it is assumed that $\{W_i, i \in \mathcal{U}_N\}$ are independent realizations of W, inclusion probabilities of the Poisson sampling scheme are defined as

$$p_i = p(W_i) = n \frac{W_i}{\sum_{j=1}^{N} W_j}, \tag{6.13}$$

with $n = \lambda N$, $\lambda \in (0,1)$, the desired expected sample size (Hàjek, 1964, Section 6, p.1512); this is the same formula as in Section 5.4 of Chapter 5. Thus defined, $p(W) \in [p_\square, p^\square]$, where $p_\square = \lambda w_\square / \bar{w}$ and $p^\square = \lambda w^\square / \bar{w}$, which offers an easy way of ensuring Assumption 6.1 is fulfilled. Furthermore, given the formula used to compute X as a function of W, the more extreme the observations, the greater the probabilities of inclusion.

Numerical experiments were conducted on a set of populations with increasing sizes $N = 10^3$, 5×10^3, 10^4 and 5×10^4. Several scenarios were investigated depending on the EVI γ; they are summarized in Table 6.1. For each scenario, two sample sizes were considered: one small with $n = 0.1 \times N$ and one relatively large with $n = 0.5 \times N$. Parameters of the distribution of W were chosen to ensure that for all $i \in \mathcal{U}_N$, $p_i \in [0.01; 1]$. Specifically, we set $\beta = 1$, $\sigma_W^2 = 0.09$, $w_\square = 0.1$ and $w^\square = 2$, thereby implying that $(p_\square, p^\square) = (0.01, 0.02)$ when $n = 0.1 \times N$ and $(p_\square, p^\square) = (0.05, 1)$ when $n = 0.5 \times N$.

Table 6.1 – List of scenarios depending on γ and corresponding optimal $K^{\gamma}(N)$

		$K^{\gamma}(N)$			
Scenario	γ	$N = 10^3$	$N = 5 \times 10^3$	$N = 10^4$	$N = 5 \times 10^4$
S_1	1/2	11	26	37	83
S_2	1	125	368	584	1709
S_3	2	514	1863	3245	11760

For each scenario, we drew 1000 samples according to a rejective sampling scheme, following Algorithm 5.9 in Tillé (2006). The true inclusion probabilities, denoted by π_i, $1 \leq i \leq N$, were deduced from their Poisson equivalents defined in Equation (6.13) using a Monte-Carlo approximation technique, based on the repetition (10^5 times) of the basic algorithm stated in Section 6.3.2. Notice that since rejective sampling is a Poisson sampling conditioned upon its size, we have $(p_i = 1) \Rightarrow (\pi_i = 1)$.

The Horvitz-Thompson version of the Hill estimator was calculated using Equation (6.6) on each of the 1000 simulated samples. The ensuing results are presented herein-after.

6.4.2.2 Experiment results

Illustrations of the behavior of $H^{\pi}_{K;N}$ in a neighborhood of $K^{\pi}(N)$ as N grows are presented in Figure 6.1 for each scenario. As a complement, we display in Figure 6.2 the empirical estimator of $\gamma^2 \pi^2_p$ in a neighborhood of $K^{\pi}(N)$ for each scenario and each sample size as N increases; for large populations, this gives some indication as to the form of the variance of $\sqrt{K}\left(H^{\pi}_{K;N} - \gamma\right)$.

Since we only considered one fixed population, these results should be interpreted with caution: they only illustrate the behavior of $\sqrt{K}\left(H^{\pi}_{K;N} - H_{K;N}\right)$ given the full vector $(X_1; W_1); \ldots; (X_N; W_N)$. We can see on Figure 6.1 that the Horvitz-Thompson version of the Hill estimator behaves perfectly well, even if the condition ii) in Theorem 6.8 is not satisfied. In particular, both its mean and variance decrease with N, more quickly when $n = 0.5 \cdot N$ than when $n = 0.1 \cdot N$, and the distribution of the estimator appears to be symmetric around its classical version, which advocates normality. Scrutinizing Figure 6.2, we can see that the asymptotic variance of $\sqrt{K}\left(H^{\pi}_{K;N} - H_{K;N}\right)$ seems indeed to be finite, depending on both the sample size and γ^2 (the smaller the EVI, the smaller the variance). This gives hope that the existence of a joint density may not be necessary for the asymptotic normality of our estimator to hold.

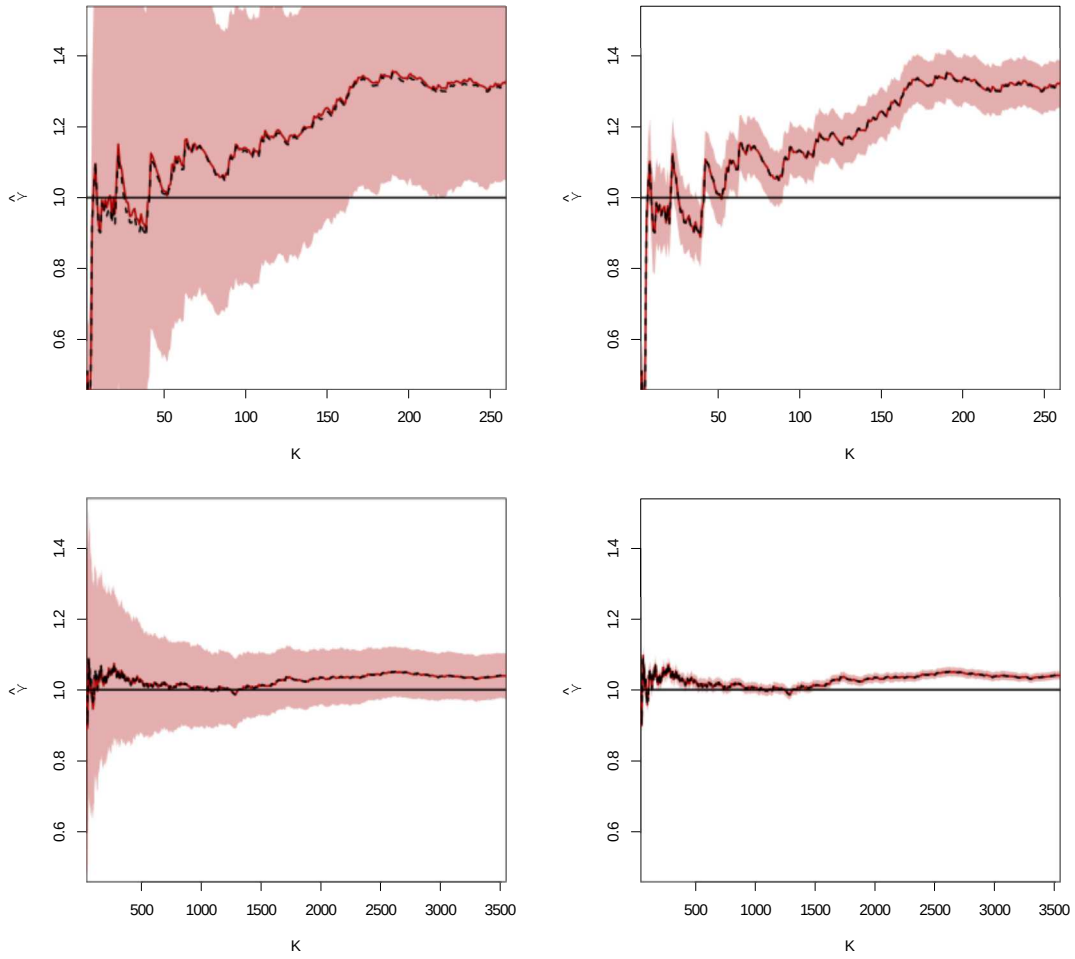


Figure 6.1 – Average values of $H_{K;N}$ (red line) and empirical 95% confidence band (pink area) computed on the 1000 simulated samples under scenario S_2 for $n = 0.1 \cdot N$ (left hand plots) and $n = 0.5 \cdot N$ (right hand plots), then compared to $H_{K;N}$ (black dotted line) for $N = 10^3$ (upper plots) and $N = 5 \cdot 10^4$ (lower plots)

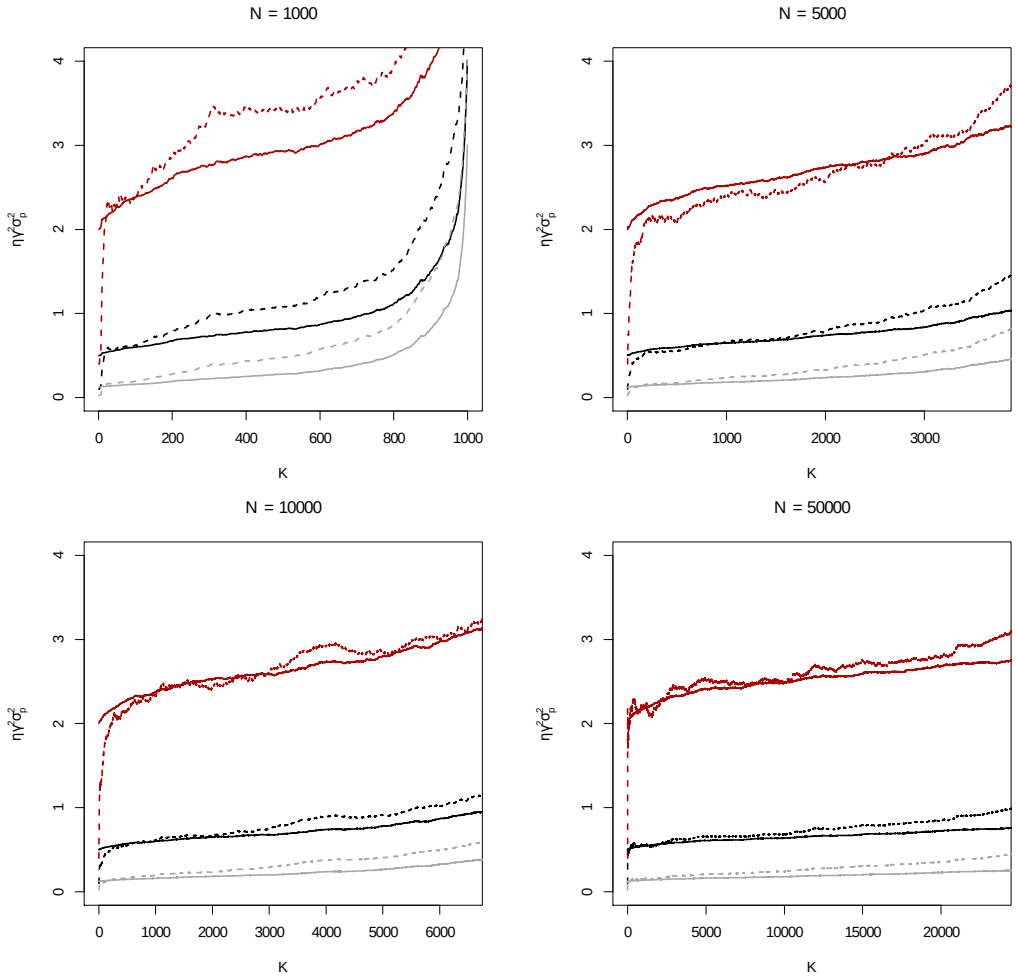


Figure 6.2 – Estimation of $\gamma^2 \sigma_p^2$ based on the 1000 simulated samples for $\rho = 0.1$ (dotted lines) and $\rho = 0.5$ (plain lines) under scenarios S_1 (grey lines), S_2 (black lines) and S_3 (red lines)

6.5 discussion

In an attempt to start adapting classical extreme value analysis to the case of survey data, we introduced in Section 6.2 a Horvitz-Thompson version of the widely celebrated Hill estimator of the extreme value index. After exhibiting some sufficient hypotheses on both the superpopulation model and the sampling scheme for the consistency of this novel statistic to hold in Section 6.2.2, we proved in Section 6.3 its asymptotic convergence to a Gaussian distribution when the survey design is of Poisson type. The exhibited rate of convergence appeared to be the same as the standard Hill estimator, namely $\sqrt{K}$, and the asymptotic variance was simply perturbed by a multiplicative constant depending solely on the sampling plan. In view of the empirical results presented in Section 6.4, hope is that the existence of a density copula linking those of the variable of interest and of the auxiliary information is not

necessary for the asymptotic normality to be true. This encourages further research to try and relax this assumption. Other improvements may be brought to these first results, for instance situations where the true inclusion probabilities are not available and replaced by an estimated version issued from post-calibration methods could be inspected. The assumptions made on the first and second order inclusion probabilities are also quite restrictive. Following in the lines of Boistard et al. (2012), higher order conditions could permit to get rid of Assumption 6.1. This remark is also true concerning the results obtained in Chapter 5. Sampling designs of other nature than the Poisson type may be considered as well, especially complex ones such as that used in the INCA2 database.

Though we could not directly apply our findings to dietary risk analysis, they could be of great interest in the context of big data management. Indeed, it is more and more frequent to meet databases that increase regularly (in finance, information about the markets is stocked every hour at least) and cannot be saved, thus analyzed, on a single computer. When accessing such huge files becomes a challenge, sampling is a natural solution. In this context, the superpopulation model and the asymptotic nature of our results are perfectly relevant. Moreover, the analyst has then complete control over the survey scheme they desire to adopt, which is typically rarely the case with institutional data. Hence, the Poisson and rejective schemes, which are not of frequent use in practice, are revealed as especially convenient for such types of analyses. With these potential assets in mind, we hope that this preliminary step towards the elaboration of a new extreme value theory for survey data will engender further research in the near future.

6.6 proofs and supplements

6.6.1 Consistency of the Horvitz-Thompson variant of the Hill estimator

We start by establishing the following intermediate results, in order to describe next the limit behavior of the Horvitz-Thompson tail empirical process.

First, we introduce the point measure:

$$\tilde{\alpha}_N := \frac{1}{K} \sum_{i=1}^{N} \frac{\pi_i}{\pi_i} \delta_{X_i / U(N/K)}.$$

Notice that the point measure α_N can be obtained from the latter by replacing the threshold $U(N/K)$ by the empirical counterpart $U_N(N/K)$.

Lemma 6.12 Under the assumptions of Theorem 6.4, as N, n and K tend to infinity so that $K/N \to 0$, we have:

$$\tilde{\alpha}_N \Rightarrow \alpha_{1/\gamma}, \tag{6.14}$$

where "$\Rightarrow$" denotes weak convergence in the space of positive Radon measures on $(0, +\infty]$.

Proof Consider first the tail empirical process

$$\alpha_N := \frac{1}{K} \sum_{i=1}^{N} \varepsilon_{X_i / U(N/K)}.$$

We shall prove that for any $t > 0$, as N, n and K tend to $+\infty$, provided K/N converges to 0,

$$D_N(t) := \tilde{\alpha}_N(t, +\infty] - \alpha_N(t, +\infty] \to 0 \text{ in } L_2. \tag{6.15}$$

Indeed, $\forall t > 0$, provided Assumption 6.1 and Assumption 6.2 hold, we have

$$\begin{aligned} E(D_N(t)) &= \frac{1}{K} \sum_{i=1}^{N} E\left(\frac{E\left[\varepsilon_i \, t(X_i, W_i)_{1 \le i \le N} \right]}{\varepsilon_i} - 1 \right) \mathbb{I}\{X_i > t U(N/K)\} \\ &= 0, \end{aligned}$$

together with

$$\begin{aligned} E\left[D_N(t)^2\right] &= \frac{1}{K^2} \sum_{i=1}^{N} E\left[E\left[\left(\frac{\varepsilon_i}{\varepsilon_i} - 1 \right)^2 t(X_i, W_i)_{1 \le i \le N} \right] \mathbb{I}\{X_i > t U(N/K)\} \right] \\ &\quad + \frac{2}{K^2} \prod_{1 \le i < j \le N} E\left[E\left[\left(\frac{\varepsilon_i}{\varepsilon_i} - 1 \right)\left(\frac{\varepsilon_j}{\varepsilon_j} - 1 \right) t(X_i, W_i)_{1 \le i \le N} \right] \right] \\ &\qquad \mathbb{I}\{X_i \wedge X_j > t U(N/K)\} \\ &= \frac{1}{K^2} \sum_{i=1}^{N} E\left[\frac{1}{\varepsilon_i} - 1 \right] \mathbb{I}\{X_i > t U(N/K)\} \\ &\quad + \frac{2}{K^2} \prod_{1 \le i < j \le N} E\left[\frac{\varepsilon_{i,j} - \varepsilon_i \varepsilon_j}{\varepsilon_i \varepsilon_j} \right] \mathbb{I}\{X_i \wedge X_j > t U(N/K)\} \\ &\le \left(\frac{1}{\varepsilon_n} - 1 \right) \frac{1}{K^2} \sum_{i=1}^{N} P(X_i > t U(N/K)) \\ &\quad + \left(\frac{\varepsilon_{2n}}{\varepsilon_n^2} - \frac{2}{K^2} \right) \prod_{1 \le i < j \le N} P\left[X_i \wedge X_j > t U(N/K) \right] =: \ell_{N,K}(t). \end{aligned}$$

Since $\bar{F}$ is supposed to be regularly varying with index $-1/\gamma$ and $K/N \to 0$, we have $P(X_i > t U(N/K)) \sim t^{-1/\gamma} K/N$ for all $i \in \{1, \ldots, N\}$ as N and K go to infinity. It follows that as $N, n, K \to +\infty$,

$$\ell_{N,K}(t) \sim \left(\frac{1}{\varepsilon_n} - 1 \right) t^{-1/\gamma} \frac{1}{K} + \left(\frac{\varepsilon_{2n}}{\varepsilon_n^2} \right) t^{-2/\gamma} \left(1 - \frac{1}{N} \right) \to 0.$$

Hence, the convergence in Equation (6.15) is proved and the desired convergence will then result from the fact that $\Gamma_N \circ \Gamma_{1/\gamma}$, see Resnick (2007, Theorem 4.1). □

We next prove the lemma below, claiming that the threshold $U_N^{\leftarrow}(N/K)$ and $U(N/K)$ are asymptotically equivalent in probability.

Lemma 6.14 Under the assumptions of Theorem 6.4, we have: as N, n and K tend to infinity,

$$\frac{U_N^{\leftarrow}(N/K)}{U(N/K)} \to 1 \text{ in probability.} \tag{6.16}$$

Proof This is a straightforward consequence of Lemma 6.12. Indeed, for all $\varepsilon > 0$, we have:

$$\begin{aligned}
P\left[\frac{U_N^{\leftarrow}(N/K)}{U(N/K)} \notin \left[1-\varepsilon, 1+\varepsilon\right]\right] &= P\left(U_N^{\leftarrow}(N/K) > (1+\varepsilon)\, U(N/K)\right) \\
&\quad + P\left(U_N^{\leftarrow}(N/K) < (1-\varepsilon)\, U(N/K)\right) \\
&\le P\left(\frac{1}{N}\sum_{i=1}^{N} \frac{\gamma_i}{\bar{\gamma}} \mathbb{1}_{X_i > ((1+\varepsilon)\, U(N/K);+\infty]} > \frac{K}{N}\right) \\
&\quad + P\left(\frac{1}{N}\sum_{i=1}^{N} \frac{\gamma_i}{\bar{\gamma}} \mathbb{1}_{X_i > ((1-\varepsilon)\, U(N/K);+\infty]} < \frac{K}{N}\right) \\
&\le P\left(\tilde{\Gamma}_N\left((1+\varepsilon;+\infty]\right) > 1\right) + P\left(\tilde{\Gamma}_N\left((1-\varepsilon;+\infty]\right) \le 1\right).
\end{aligned}$$

Therefore, by virtue of the lemma previously established, we asymptotically have: $\tilde{\Gamma}_N\left((1+\varepsilon;+\infty]\right) \to 1/(1+\varepsilon)^{1/\gamma} < 1$ and $\tilde{\Gamma}_N\left((1+\varepsilon;+\infty]\right) \to 1/(1-\varepsilon)^{1/\gamma} > 1$ in probability. Combined with the bound above, this proves the lemma. □

Equipped with these preliminary results, we may now tackle the proof of Theorem 6.4, which is recalled below for convenience.

Theorem – Consistency. Let $K = K(N)$ be a sequence of integers such that $K \to +\infty$ and $K/N \to 0$ as $N, n \to +\infty$. Provided that Assumption 6.1 and Assumption 6.2 are fulfilled, we then have, as N and n tend to $+\infty$:

$$H_{K,N}^{\leftarrow} \to \gamma \text{ in probability.} \tag{6.17}$$

Proof The consistency result can be established by following line by line the proof for the consistency of the Hill estimator in the iid situation given in Resnick (2007): by a continuous mapping theorem argument, one derives from Lemma 6.12 and Lemma 6.14 that the Horvitz-Thompson tail empirical process Γ_N converges in probability to $\Gamma_{1/\gamma}$ in the space of positive Radon measures on $(0;+\infty]$. Then, it classically suffices to integrate the tail measures against dt/t (cf. Equation (6.4) and Equation (6.6)) and apply the convergence previously mentioned. See Resnick (2007, Section 4.4.1) for further details. □

6.6.2 Limit distribution of $H^{p}_{K;N}$ in the Poisson survey case

Before handling Theorem 6.8, we start by establishing three intermediate results, which are introduced herein-after. The first lemma claims that the quantity $r_{K;N}$ defined in Equation (6.11) converges to 1 in probability.

Lemma 6.17 Let $\delta_1;\ldots;\delta_N$ and $p_1;\ldots;p_N$ be respectively the inclusion indicators and probabilities of a Poisson survey plan in some population $U_N := \{1;\ldots;N\}$. Then, provided Assumption 6.1 holds, for any $K \in \{1;\ldots;N\}$ such that $K := K(N) \xrightarrow[N\to\infty]{} +\infty$ we have

$$r_{K;N} := \frac{1}{K}\sum_{i=1}^{K} \frac{\delta_{N-i+1;N}}{p_{N-i+1;N}} \xrightarrow[N\to\infty]{\mathbb{P}} 1.$$

Proof Recall that under a Poisson sampling plan, all $\delta_1;\ldots;\delta_N$ are independent. We had set $\mathbb{E}\left[\delta_i \,|\, (X_i;W_i)\right] := p_i$ for all $i \in \{1;\ldots;N\}$, hence

$$\mathbb{E}(r_{K;N}) = \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}\left[\frac{\mathbb{E}\left[\delta_{N-i+1;N} \,|\, (X_1;W_1);\ldots;(X_N;W_N)\right]}{p_{N-i+1;N}}\right] = 1.$$

In addition, $\mathbb{V}\left[\delta_i \,|\, (X_i;W_i)\right] := p_i(1-p_i)$ for all $i \in \{1;\ldots;N\}$, therefore

$$\begin{aligned}\mathbb{V}(r_{K;N}) &= \frac{1}{K^2}\sum_{i=1}^{K} \mathbb{E}\left[\frac{\mathbb{V}\left[\delta_{N-i+1;N} \,|\, (X_1;W_1);\ldots;(X_N;W_N)\right]}{p^2_{N-i+1;N}}\right] \\ &\quad + \frac{1}{K^2}\sum_{i=1}^{K} \mathbb{V}\left[\frac{\mathbb{E}\left[\delta_{N-i+1;N} \,|\, (X_1;W_1);\ldots;(X_N;W_N)\right]}{p_{N-i+1;N}}\right] \\ &= \frac{1}{K^2}\sum_{i=1}^{K} \mathbb{E}\left[\frac{1}{p_{N-i+1;N}} - 1\right] \leq \frac{1}{K}.\end{aligned}$$

Under Assumption 6.1 it is clear that $\mathbb{V}(r_{K;N}) = O(\frac{1}{K})$ when $K \xrightarrow[N\to\infty]{} +\infty$. This concludes the proof. □

We now move to the quantity $\sqrt{K}\left(r_{K;N}\, H^{p}_{K;N} - H_{K;N}\right)$ appearing in Equation (6.10). The result below reveals that is vanishes asymptotically.

Lemma 6.19 Suppose that Assumption 6.5 is fulfilled by the underlying heavy-tailed model and that Assumption 6.1 is satisfied by the considered sequence of Poisson inclusion probabilities $p_1;\ldots;p_N$, $N \geq 1$. Assume also that $K \to +\infty$ as $N \to +\infty$ so that $\sqrt{K}A(N/K) \to \lambda$ for some constant $\lambda \in \mathbb{R}$. Then we have

$$\sqrt{K}\left(r_{K;N}\, H^{p}_{K;N} - H_{K;N}\right) \xrightarrow[N\to\infty]{\mathbb{P}} 0.$$

Proof Let us start by introducing the weighted versions of log-spacings in the population, given by

$$\forall \{1,\dots,N\},\quad \zeta_i := i\,(\log X_{N-i+1;N} - \log X_{N-i;N}).$$

Let $K \in \{1,\dots,N\}$, these random variables are intrinsically linked to both $H_{K;N}$ and $H^p_{K;N}$, given by Equation (6.3) and Equation (6.6) respectively. Indeed, they can be expressed as

$$H_{K;N} = \frac{1}{K}\sum_{i=1}^{K} \zeta_i$$

and

$$H^p_{K;N} = \left(\frac{1}{K}\sum_{j=1}^{K}\frac{\varepsilon_{N-j+1;N}}{p_{N-j+1;N}}\right)^{-1} \frac{1}{K}\sum_{i=1}^{K}\left(\frac{1}{i}\sum_{j=1}^{i}\frac{\varepsilon_{N-j+1;N}}{p_{N-j+1;N}}\right)\zeta_i.$$

Combining the two we immediately obtain

$$r_{K;N}\,H^p_{K;N} - H_{K;N} = \frac{1}{K}\sum_{i=1}^{K}\left(\frac{1}{i}\sum_{j=1}^{i}\frac{\varepsilon_{N-j+1;N}}{p_{N-j+1;N}} - 1\right)\zeta_i. \tag{6.18}$$

When Assumption 6.5 is fulfilled, it is possible to approximate the distribution of the ζ_i's corresponding to the $K+1$ largest values. Denoting by $E_1,\dots,E_K$ a collection of independent random variables with standard exponential distribution, the random variables ζ_i are approximately distributed as

$$\left(\frac{i}{K+1}\right)^{\gamma}\left(\frac{N+1}{K+1}\right)^{\gamma} E_i,\quad 1 \leqslant i \leqslant K. \tag{6.19}$$

This property is at the basis of most of the asymptotic analyzes that were led concerning $H_{K;N}$, see de Haan and Resnick (1998) for more details. As mentioned in Remark 6.6, alternative approaches taking advantage of the Glivenko-Cantelli and Donsker theorems in the formulation of the Hill estimator in Equation (6.3) were also developed, see Resnick (2007) for instance.

Given the decomposition in Equation (6.18), just like $r_{K;N}$ in Lemma 6.17 the expectation and variance of $r_{K;N}\,H^p_{K;N} - H_{K;N}$ are easily derived by conditioning upon the full vector of observations $(X_1;W_1),\dots,(X_N;W_N)$. In particular, it is straightforward to see that

$$\mathbb{E}\left[r_{K;N}\,H^p_{K;N} - H_{K;N}\right] = \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\left(\frac{1}{i}\sum_{j=1}^{i}\frac{\zeta_{N-j+1;N}}{p_{N-j+1;N}} - 1\right)\zeta_i\right] = 0.$$

Turning now to the variance, under Assumption 6.1, we have

$$\mathbb{V}\left[r_{K;N}\,H^p_{K;N} - H_{K;N}\right] = \frac{1}{K^2}\sum_{i=1}^{K}\mathbb{E}\left[\frac{\zeta_i^2}{i^2}\sum_{j=1}^{i}\frac{1}{p_{N-j+1;N}} - 1\right]^{\gamma}$$
$$\leqslant \frac{1}{K^2}\sum_{i=1}^{K}\mathbb{E}\left[\zeta_i^2\right]\gamma\, p_{\gamma-1} - 1.$$

To simplify notations, set $c = p_{\square}^{\square 1} \square 1$ and consider the variable $A_{K;N} = A^{\square \frac{N+1}{K+1} \square}$ involved in Equation (6.19). Using this particular equation, we are able to establish asymptotic properties of the right hand side in the inequality herein above: as $N;K \to +1$ we have

$$\begin{aligned}
\frac{1}{K^2}\sum_{i=1}^{K} E\left[\square_i^2 \square \frac{c}{i}\right] \square \frac{1}{K^2}\sum_{i=1}^{K} & \square + \left(\frac{i}{K+1}\right)^{\square\square} A_{K;N}\Bigg]^2 E\left[\square E_i^2 \square \frac{c}{i}\right] \\
& \square \frac{2c\square^2}{K^2}\sum_{i=1}^{K}\frac{1}{i} + \frac{2c}{K^2(K+1)^{\square 2\square}}A_{K;N}^2\sum_{i=1}^{K}\frac{1}{i^{2\square+1}} \\
& + \frac{4c\square}{K^2(K+1)^{\square\square}}A_{K;N}\sum_{i=1}^{K}\frac{1}{i^{\square+1}} \\
& \square \frac{\log K}{K^2}(2c\square^2) \square \frac{K^{\square 2\square}\square 1}{K^2(K+1)^{\square 2\square}}\frac{c}{\square}A_{K;N}^2 \\
& \square \frac{K^{\square\square}\square 1}{K^2(K+1)^{\square\square}}\frac{4c\square}{\square}A_{K;N} \\
& \square \frac{\log K}{K^2}(2c\square^2) \square \frac{1}{K^2}\frac{c}{\square}A_{K;N}^2 \square \frac{1}{K^2}\frac{4c\square}{\square}A_{K;N}.
\end{aligned}$$

Because $A \in PR_{\square}$ with $\square \dagger 0$, then $A_{K;N} \to 0$ as $N;K \to +1$, and we can conclude that $V\left[r_{K;N} H_{K;N}^{p} \square H_{K;N}^{\square}\right] = o(1=K)$. □

The last intermediate result concerns the quantity $Q_N^{(3)}$ in Equation (6.10). It claims that the latter converges weakly to a centered Normal distribution.

Lemma 6.21 Suppose that Assumption 6.1 is satisfied by the considered sequence of Poisson inclusion probabilities $p_1;\ldots;p_N$, $N \bullet 1$, constructed from some set of auxiliary variables as in Assumption 6.7. Further assume the conditions i) □ iii) introduced in Theorem 6.8. Then, for

$$\square_p^2 := \int_w^a \frac{1}{p(w)} c_{X;W}(1;F_{W_1}(w_1);\ldots;F_{W_d}(w_d)) \prod_{j=1}^{d} f_{W_j}(w_j)\,dw_1\ldots dw_d$$

and provided that $K \to +1$ as $N \to +1$ so that $K = o(N)$, we have the convergence in distribution as $N \to +1$:

$$\sqrt{K}(1\square r_{K;N}) \grave{O} N\left(0;\square_p^2 \square 1\right).$$

Proof This proof is based on the application of Feller (1971, Theorem 3, p.262) to the collection of random variables $\{Z_{i;N}(\square);\ 1 \leq i \leq N\}$ defined for all $i \in \{1;\ldots;N\}$ as follows:

$$Z_{i;N}(\square) := \sqrt{\frac{1}{K}}\left(1\square\frac{\square_i}{p_i}\right) I\{X_i \neq X_{N\square K;N}\};$$

with distribution P_i only depending on the survey scheme (they are conditioned upon the vectors $t(X_i; W_i);\ 1 \leq i \leq N$). Indeed, notice that we have

$$\sqrt{K}(1 - r_{K;N}) = \prod_{i=1}^{N} Z_{i;N}(\theta).$$

In order to apply this theorem, we first have to check the three conditions below.

(C_1) For all $i \in \{1; \ldots; N\}$, we have $E(Z_{i;N}(\theta)) = 0$ and $V(Z_{i;N}(\theta)) = \sigma_i^2 < 1$.

(C_2) There exists some real constant $\sigma^2 < 1$ such that $S_N^2 := \sum_{i=1}^{N} \sigma_i^2 \xrightarrow{P} \sigma^2$.

(C_3) For each $t > 0$, we have $\sum_{i=1}^{N} \int_{|z| \geq t S_N} z^2 P_i(dz) \xrightarrow{P} 0$.

Condition (C_1) Let us start by calculating the expectation of $Z_{i;N}(\theta)$ for all $i \in \{1; \ldots; N\}$. Because the $\theta_1; \ldots; \theta_N$ are independent random variables with respective Bernoulli distributions $B(p_1); \ldots; B(p_N)$, it is straightforward to see that

$$E(Z_{i;N}(\theta)) = \frac{1}{\sqrt{K}}\left(1 - E\left[\theta_i \frac{t(X_i; W_i)_{1 \leq i \leq N}}{p_i}\right] \mathbb{I}\{X_i > X_{N-K;N}\}\right) = 0.$$

As for the variance, we have

$$\begin{aligned} V(Z_{i;N}(\theta)) &= E\left[Z_{i;N}^2(\theta)\right] \\ &= \frac{1}{K} E\left[\left(1 - \frac{\theta_i}{p_i}\right)^2 t(X_i; W_i)_{1 \leq i \leq N} \mathbb{I}\{X_i > X_{N-K;N}\}\right] \\ &= \frac{1}{K} \frac{1 - p_i}{p_i} \mathbb{I}\{X_i > X_{N-K;N}\} =: \sigma_i^2 < 1. \end{aligned}$$

Therefore, condition (C_1) is fulfilled.

Condition (C_2) We now have to prove that S_N^2 converges in probability to a finite constant σ^2 as N tends to infinity and exhibit the required conditions for this property to hold. First observe that

$$S_N^2 := \sum_{i=1}^{N} \sigma_i^2 = \frac{1}{K} \sum_{i=1}^{N} \frac{1}{p_i} \mathbb{I}\{X_i > X_{N-K;N}\} - 1;$$

where $X_{N-K;N}$ is a consistent estimator of $U(N-K)$ (Resnick, 2007, Section 4.4.1, p.81). With this remark in mind, we will proceed in two steps and successively prove that there exists a real constant $\sigma^2 > 0$ such that

$$\tilde{S}_N^2 := \frac{1}{K} \sum_{i=1}^{N} \frac{1}{p_i} \mathbb{I}\{X_i > U(N-K)\} - 1 \xrightarrow{P} \sigma^2;$$

then that $S_N^2 - \tilde{S}_N^2 \xrightarrow{P} 0$. Since $S_N^2 = S_N^2 - \tilde{S}_N^2 + \tilde{S}_N^2$, this will yield the desired result.

Let us start with $\tilde{S}_N^2$. By virtue of the law of large numbers, as $N \to +1$ we have

$$\tilde{S}_N^2 + 1 = \frac{N}{K} \int_{w}^{a} \int_{U(N=K)}^{a} \frac{1}{p(w)} P_{X;W}(dx; dw) + o_P(1);$$

provided that this integral exists. Under Assumptions i) and ii) in Theorem 6.8, this yields

$$\tilde{S}_N^2 + 1 = \frac{N}{K} \int_{w}^{a} \int_{U(N=K)}^{a} \frac{1}{p(w)} c_{X;W}(F(x); F_{W_1}(w_1); \ldots; F_{W_d}(w_d)) \cdot f(x) \prod_{j=1}^{d} f_{W_j}(w_j)\, dx\, dw_1 \ldots dw_d + o_P(1).$$

Further set $u := \frac{N}{K} \bar{F}(x)$, then for $K := K(N) \to +1$ as $N \to +1$ and $K = o(N)$, under Assumption iii) we have:

$$\begin{aligned}\tilde{S}_N^2 + 1 &= \int_{w}^{a} \int_{0}^{a} \frac{1}{p(w)} c_{X;W}\left(1 - \frac{K}{N} u; F_{W_1}(w_1); \ldots; F_{W_d}(w_d)\right) \\ &\quad \cdot \prod_{j=1}^{d} f_{W_j}(w_j) du\, dw_1 \ldots dw_d + o_P(1) \\ &\xrightarrow[N \to 1]{P} \int_{w} \frac{1}{p(w)} c_{X;W}(1; F_{W_1}(w_1); \ldots; F_{W_d}(w_d)) \\ &\quad \cdot \prod_{j=1}^{d} f_{W_j}(w_j) dw_1 \ldots dw_d =: \sigma_p^2.\end{aligned}$$

Provided that all the aforementioned hypotheses hold, we can conclude that

$$\tilde{S}_N^2 \xrightarrow[N \to 1]{P} \sigma_p^2 \cdot 1.$$

There remains to control the quantity $\left(S_N^2 - \tilde{S}_N^2\right)$, which we denote by S_N for simplicity. For any fixed $N \in \mathbb{N}^*$ and $\eta > 0$, it can be decomposed as follows:

$$S_N := \frac{1}{K} \sum_{i=1}^{N} \frac{1}{p_i} (\mathbb{1}_{X_i > U(N=K)u} - \mathbb{1}_{X_i > X_{N-K;N}u}) = S_N^{(1)}(\eta) + S_N^{(2)}(\eta);$$

where

$$S_N^{(1)}(\eta) := S_N \left| \frac{X_{N-K;N}}{U(N=K)} - 1 \right| > \eta \Bigg|^*$$

and

$$S_N^{(2)}(\eta) = S_N \left| \frac{X_{N-K;N}}{U(N=K)} - 1 \right| \leq \eta \Bigg|^*.$$

We shall use this decomposition to prove that $S_N \xrightarrow[N \to 1]{P} 0$. Referring to Assumption 6.1, it is easy to see that $S_N^{(1)}(\eta) \xrightarrow[N \to 1]{P} 0$. Indeed,

$$\left| S_N^{(1)}(\eta) \right| \leq \frac{1}{p_-} \left| \frac{1}{K} \sum_{i=1}^{N} (\mathbb{1}_{X_i > U(N=K)u} - \mathbb{1}_{X_i > X_{N-K;N}u}) \right|$$

$$\mathbb{I}\left\{\frac{X_{N-K:N}}{U(N/K)} > 1+\varepsilon\right\} \leq \frac{1}{p_\varepsilon}\frac{1}{K}\sum_{i=1}^{N} \mathbb{I}\{X_i > U(N/K)u\} \mathbb{I}\left\{\frac{X_{N-K:N}}{U(N/K)} > 1+\varepsilon\right\}.$$

The law of large numbers ensures that as $N \to \infty$,

$$\frac{1}{K}\sum_{i=1}^{N} \mathbb{I}\{X_i > U(N/K)u\} = \frac{N}{K}\mathbb{P}(X_i > U(N/K)) \to 1 + o_P(1) \to 0.$$

In addition, we know that for all $\varepsilon > 0$,

$$\mathbb{I}\left\{\frac{X_{N-K:N}}{U(N/K)} > 1+\varepsilon\right\} \xrightarrow[N\to\infty]{\mathbb{P}} 0$$

(Resnick, 2007, Section 4.4.1, p.81), hence $S_N^{(1)}(\varepsilon) \xrightarrow[N\to\infty]{\mathbb{P}} 0$.

As for $S_N^{(2)}(\varepsilon)$, combining Assumption 6.1 with triangular inequalities provides an absolute bound:

$$\begin{aligned}
S_N^{(2)}(\varepsilon) &\leq \frac{1}{p_\varepsilon}\frac{1}{K}\sum_{i=1}^{N} \left|\mathbb{I}\{X_i > U(N/K)u\} - \mathbb{I}\{X_i > X_{N-K:N}u\}\right| \\
&\quad \mathbb{I}\left\{\frac{X_{N-K:N}}{U(N/K)} > 1-\varepsilon\right\} \\
&\leq \frac{1}{p_\varepsilon}\frac{1}{K}\sum_{i=1}^{N} \mathbb{I}\{(1-\varepsilon)U(N/K) < X_i \leq (1+\varepsilon)U(N/K)u\} \\
&\leq \frac{1}{p_\varepsilon}\frac{1}{K}\sum_{i=1}^{N} \mathbb{I}\{X_i > (1-\varepsilon)U(N/K)u\} - \mathbb{I}\{X_i > (1+\varepsilon)U(N/K)u\}.
\end{aligned}$$

Denote by $c_N(\varepsilon)$ the quantity in the right hand part of the last inequality, i.e.

$$c_N(\varepsilon) := \frac{1}{p_\varepsilon}\frac{1}{K}\sum_{i=1}^{N} \mathbb{I}\{X_i > (1-\varepsilon)U(N/K)u\} - \mathbb{I}\{X_i > (1+\varepsilon)U(N/K)u\}.$$

Applying the law of large numbers and recalling that $\bar{F} \in RV_{-1/\gamma}$, we get that as $N \to +\infty$,

$$\begin{aligned}
c_N(\varepsilon) &= \frac{1}{p_\varepsilon}\frac{N}{K}\left[(1-\varepsilon)^{-1/\gamma} - (1+\varepsilon)^{-1/\gamma}\right]\bar{F}(U(N/K)) + o_P(1) \\
&\xrightarrow{\mathbb{P}} \frac{1}{p_\varepsilon}\left[(1-\varepsilon)^{-1/\gamma} - (1+\varepsilon)^{-1/\gamma}\right] =: c_\varepsilon;
\end{aligned}$$

meaning that $c_N(\varepsilon) \xrightarrow[N\to\infty]{\mathbb{P}} c_\varepsilon$. Furthermore, notice that $c_\varepsilon \xrightarrow[\varepsilon\to 0]{} 0$, since

$$\begin{aligned}
|c_\varepsilon| &= \frac{1}{p_\varepsilon}\left[(1-\varepsilon)^{-1/\gamma} - (1+\varepsilon)^{-1/\gamma}\right] \\
&\leq \frac{1}{p_\varepsilon}\left[\left(1+\frac{\varepsilon}{\gamma}+o(\varepsilon)\right) - \left(1-\frac{\varepsilon}{\gamma}+o(\varepsilon)\right)\right]
\end{aligned}$$

$$\leq \frac{2\varepsilon}{p_{\varepsilon}\varepsilon} + o(\varepsilon) \xrightarrow[\varepsilon \to 0]{} 0.$$

We may now prove that $S_N \xrightarrow[N \to +\infty]{P} 0$ by going back to the definition.

Formally, we have shown that $|S_N| \leq |S_N^{(1)}(\varepsilon)| + c_N(\varepsilon) =: \delta_N(\varepsilon)$, for all $\varepsilon > 0$. We also know that $\delta_N(\varepsilon) \xrightarrow[N \to +\infty]{P} c_\varepsilon$ and $c_\varepsilon \xrightarrow[\varepsilon \to 0]{} 0$. Now for any fixed $\varepsilon > 0$, we have to verify that $P(|S_N| > 2\varepsilon) \xrightarrow[N \to +\infty]{} 0$. First choose $\varepsilon_0 > 0$ such that for all $0 < \varepsilon \leq \varepsilon_0$, $|c_\varepsilon| \leq \varepsilon$, then take some $\varepsilon > 0$. We need to prove that there exists $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$, $P(|S_N| > 2\varepsilon) \leq \varepsilon$. In order to construct this N_0, first fix any $\varepsilon > 0$ such that $\varepsilon \leq \varepsilon_0$. Since $\delta_N(\varepsilon) \xrightarrow[N \to +\infty]{P} c_\varepsilon$, there exists $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$, we have $P(|\delta_N(\varepsilon) - c_\varepsilon| > \varepsilon) \leq \varepsilon$. In parallel, since $\varepsilon \leq \varepsilon_0$, we have for all $N \in \mathbb{N}$:

$$|S_N| \leq |\delta_N(\varepsilon)| \leq |\delta_N(\varepsilon) - c_\varepsilon| + |c_\varepsilon| \leq |\delta_N - c_\varepsilon| + \varepsilon.$$

This implies that for all $N \geq N_0$,

$$P(|S_N| > 2\varepsilon) \leq P(|\delta_N(\varepsilon) - c_\varepsilon| + \varepsilon > 2\varepsilon) = P(|\delta_N(\varepsilon) - c_\varepsilon| > \varepsilon) \leq \varepsilon.$$

Since this is true for any $\varepsilon > 0$, this means that $S_N \xrightarrow[N \to +\infty]{P} 0$.

In fine, we can conclude that under all the hypotheses stated in Lemma 6.21,

$$S_N^2 \xrightarrow[N \to +\infty]{P} \sigma_p^2 > 1.$$

Ú Condition (C_3) Our last task consists in verifying that for any $t > 0$, the quantity below, denoted for simplicity by $Z_N(t)$, converges to 0 as $N \to +\infty$. By definition, we have

$$\begin{aligned} Z_N(t) &:= \sum_{i=1}^{n} \int_{|z| > t\, S_N} z^2 P_i(dz) \\ &= \sum_{i=1}^{n} E\left[Z_{i;N}^2(\varepsilon)\, \mathbb{I}\{|Z_{i;N}(\varepsilon)| > t\, S_N\}\, t(X_i; W_i)_{1 \leq i \leq N} \right] \\ &= \frac{1}{K} \sum_{i=1}^{n} \mathbb{I}\{X_i > X_{N-K;N}\} \\ &\quad E\left[\left(1 - \frac{\pi_i}{p_i}\right)^2 \left| \left(1 - \frac{\pi_i}{p_i}\right) \right| > \frac{t\, S_N}{K\, \mathbb{I}\{X_i > X_{N-K;N}\}} \right] t(X_i; W_i)_{1 \leq i \leq N} \Big]. \end{aligned}$$

Using Hölder's inequality, we obtain

$$\begin{aligned} Z_N(t) \leq \frac{1}{K} \sum_{i=1}^{n} \mathbb{I}\{X_i > X_{N-K;N}\} & E\left[\left(1 - \frac{\pi_i}{p_i}\right)^3 t(X_i; W_i)_{1 \leq i \leq N} \right]^{2/3} \\ & E\left[\left| \left(1 - \frac{\pi_i}{p_i}\right) \right| > \frac{t\, S_N}{K\, \mathbb{I}\{X_i > X_{N-K;N}\}} \right] t(X_i; W_i)_{1 \leq i \leq N} \Big]^{1/3}. \end{aligned}$$

Observe that under Assumption 6.1, we have

$$\mathbb{E}\left[\left(1-\frac{\varepsilon_i}{p_i}\right)^3 \mathbb{1}_{\{t(X_i;W_i)_{1\le i\le N}\}}\right] = \left(\frac{1}{p_i}-3\right)\left(\frac{1}{p_i}-2\right)^{2/3} \le 3\left(\frac{1}{p_\star}-1\right)^{2/3}.$$

Moreover, conditional on the vectors $t(X_i;W_i)$, $1 \le i \le N$, the random variable $\left(1-\frac{\varepsilon_i}{p_i}\right)$ equals either $(p_i-1)/p_i$ with probability p_i or 1 with probability $1-p_i$. Therefore, by virtue of Markov's inequality, we can further bound $Z_N(t)$ from above:

$$\begin{aligned} Z_N(t) &\le \frac{1}{K}\sum_{i=1}^{N} \mathbb{I}\{X_i > X_{N-K;N}\}\, 3\left(\frac{1}{p_\star}-1\right)^{2/3} \\ &\quad + \frac{1}{K}\mathbb{I}\{X_i > X_{N-K;N}\}\,\frac{2(1-p_i)}{t\,S_N}^{1/3} \\ &\le \frac{3^{2/3}}{K^{4/3}}\left(\frac{1}{p_\star}-1\right)^{2/3}\sum_{i=1}^{N}\mathbb{I}\{X_i > X_{N-K;N}\}\,\frac{2(1-p_i)}{t\,S_N}^{1/3}. \end{aligned}$$

Using again Assumption 6.1, this yields

$$\begin{aligned} Z_N(t) &\le \frac{2^{1/3}\cdot 3^{2/3}}{K^{4/3}}\left(\frac{1}{p_\star}-1\right)^{2/3}\left(\frac{1-p_\star}{t\,S_N}\right)^{1/3}\sum_{i=1}^{N}\mathbb{I}\{X_i > X_{N-K;N}\} \\ &\le \frac{2^{1/3}\cdot 3^{2/3}}{K^{1/3}}\left(\frac{1}{p_\star}-1\right)^{2/3}\left(\frac{1-p_\star}{t\,S_N}\right)^{1/3}; \end{aligned}$$

where we have shown that $S_N \xrightarrow{P} \infty$. Consequently, the right hand part of this last inequality tends to 0 in probability as N tends to infinity for any $t > 0$. Hence, condition (C_3) is fulfilled.

With all three conditions (C_1), (C_2) and (C_3) satisfied, by virtue of Feller (1971, Theorem 3, p.262) we finally have

$$\sqrt{K}\,(1-r_{K;N}) \underset{N\to+\infty}{\Rightarrow} \mathcal{N}\left(0;\sigma_p^2-1\right).$$

□

We are now fully equipped to prove Theorem 6.8, the statement of which is recalled below.

Theorem – Limit distribution in the Poisson survey case. Suppose that Assumption 6.5 is fulfilled by the underlying heavy-tailed model and that Assumption 6.1 is satisfied by the considered sequence of Poisson inclusion probabilities $p_1, \ldots, p_N$, $N \ge 1$, constructed from some set of auxiliary variables as in Assumption 6.7. Further assume the conditions introduced herein-after.

i) The marginal cdf F is absolutely continuous with respect to the Lebesgue measure with density f.

ii) The joint cdf $F_{X;W}$ is absolutely continuous with Lebesgue-integrable density $f_{X;W}$ such that for all $(x;w) \in (0;+\infty[\times W$, $w = (w_1;\dots;w_d)$,

$$f_{X;W}(x;w) := c_{X;W}(F(x);F_{W_1}(w_1);\dots;F_{W_d}(w_d))\, f(x) \prod_{j=1}^{d} f_{W_j}(w_j);$$

for some copula density $c_{X;W} : \mathbb{R}_+^* \times \mathbb{R}^d \to \mathbb{R}$ and $f_{W_1};\dots;f_{W_d}$ the marginal densities of the distribution of W.

iii) $c_{X;W}$ bounded in a neighborhood of $\{1\} \times [0;1]^d$.

Then, for

$$\sigma_p^2 := \int_W \frac{1}{p(w)}\, c_{X;W}(1;F_{W_1}(w_1);\dots;F_{W_d}(w_d)) \prod_{j=1}^{d} f_{W_j}(w_j)\, dw_1 \dots dw_d$$

and provided that $K \to +\infty$ as $N \to +\infty$ so that $\sqrt{K}A(N/K) \to \lambda$ for some constant $\lambda \in \mathbb{R}$, we have the convergence in distribution as $N \to +\infty$:

$$\sqrt{K}\left(H_{K;N}^{p} - \gamma^{p}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\frac{\lambda}{1-\rho}; \gamma^{2p}\sigma_p^2\right).$$

Proof Recall the decomposition in Equation (6.10):

$$\sqrt{K}\left(H_{K;N}^{p} - \gamma^{p}\right) = \underbrace{\frac{\sqrt{K}}{r_{K;N}}\, r_{K;N}^{p}\left(H_{K;N}^{p} - H_{K;N}\right)}_{Q_N^{(1)}} + \underbrace{\frac{\sqrt{K}}{r_{K;N}}\left(H_{K;N} - \gamma\right)}_{Q_N^{(2)}} + \underbrace{\sqrt{K}\left(\frac{1}{r_{K;N}} - 1\right)}_{Q_N^{(3)}}.$$

Combining Lemma 6.17 and Lemma 6.19, provided that Assumption 6.1 and Assumption 6.5 hold and that $K = K(N) \to +\infty$, $K = o(N)$ and $\sqrt{K}A(N/K) \to \lambda$ for some constant $\lambda \in \mathbb{R}$, we have

$$Q_N^{(1)} \xrightarrow[N\to\infty]{\mathbb{P}} 0.$$

Lemma 6.17 also ensures that under Assumption 6.1, $Q_N^{(2)}$ is equivalent to

$$\sqrt{K}\left(H_{K;N} - \gamma\right).$$

Referring for instance to De Haan and Ferreira (2006, Theorem 3.2.5), this entails that provided Assumption 6.5 holds and that $K = K(N) \to +\infty$, $K = o(N)$ and

$\sqrt{K}A(N/K) \to \lambda$ for some constant $\lambda \in \mathbb{R}$, we have the convergence in distribution as $N \to +\infty$:

$$Q_N^{(2)} \Rightarrow \mathcal{N}\left(\frac{\lambda}{1-\theta}, \sigma^2\right).$$

Finally, by virtue of Lemma 6.17 and Lemma 6.21, if Assumption 6.1, Assumption 6.7 and conditions i), ii), iii) in Theorem 6.8 hold together with $K = K(N) \to +\infty$ and $K = o(N)$, we have the convergence in distribution as $N \to +\infty$:

$$Q_N^{(3)} \Rightarrow \mathcal{N}\left(0, \sigma^2 \sigma_p^2 (1-\theta)\right).$$

Because the limit distribution of $Q_N^{(3)}$ was established conditionally on the set $\{(X_i, W_i),\ 1 \leq i \leq N\}$ and in probability relative to this full vector of observations, we can consider $Q_N^{(2)}$ and $Q_N^{(3)}$ as independent random variables (one depends on the data and the other on the survey scheme). The limit distribution of their sum is then the sum of their limit distributions. This concludes the proof. □

6.6.3 Limit distribution of $H_{K,N}^{(\alpha)}$ in the rejective survey case

We shall prove Theorem 6.11, the statement of which is recalled below.

Theorem – Limit distribution in the rejective survey case. Suppose that all the conditions required in Theorem 6.8 hold together with Assumption 6.10. Then, for

$$\sigma_p^2 := \int_{\mathbb{R}_+^d} \frac{1}{p(w)}\, c_{X,W}(1, F_{W_1}(w_1), \ldots, F_{W_d}(w_d)) \prod_{j=1}^{d} f_{W_j}(w_j)\, dw_1 \ldots dw_d$$

and provided that $K \to +\infty$ as $N \to +\infty$ so that $\sqrt{K}A(N/K) \to \lambda$ for some constant $\lambda \in \mathbb{R}$, we have the convergence in distribution as $N \to +\infty$:

$$\sqrt{K}\left(H_{K,N}^{(\alpha)} - \gamma\right) \Rightarrow \mathcal{N}\left(\frac{\lambda}{1-\theta}, \sigma^2 \sigma_p^2\right).$$

Proof We shall write $H_{K,N}^{(\alpha)}(R_N)$ when the Horvitz-Thompson version of the Hill estimator involves the inclusion variables $\pi_1, \ldots, \pi_N$ drawn under the sampling plan R_N and the probabilities of inclusion $\pi_1, \ldots, \pi_N$. Consider a Poisson scheme T_N with probabilities $p_1, \ldots, p_N$ and a Rejective scheme R_N with probabilities $\pi_1, \ldots, \pi_N$. Since Theorem 6.8 establishes the asymptotic convergence of $\sqrt{K}\left(H_{K,N}^{(\alpha)}(T_N) - \gamma\right)$ to a Normal distribution, we only have to control the quantity $\sqrt{K}\left(H_{K,N}^{(\alpha)}(R_N) - H_{K,N}^{(\alpha)}(T_N)\right)$. Using triangular inequalities, we obtain

$$\sqrt{K}\left|H_{K,N}^{(\alpha)}(R_N) - H_{K,N}^{(\alpha)}(T_N)\right| \leq \underbrace{\sqrt{K}\left|H_{K,N}^{(\alpha)}(R_N) - H_{K,N}^{(\alpha)}(T_N)\right|}_{Q_N^{(4)}} + \underbrace{\sqrt{K}\left|H_{K,N}^{(\alpha)}(T_N) - H_{K,N}^{(\alpha)}(T_N)\right|}_{Q_N^{(5)}}.$$

We shall successively prove that both $Q_N^{(4)}$ and $Q_N^{(5)}$ tend to 0 in probability as $N \to \infty$.

Let us start with $Q_N^{(4)}$. We use the bounded Lipschitz metric d_{BL} as defined in van der Vaart and Wellner (1996, p.73) and consider, conditionally on the full vector of observations $t(X_i; W_i); 1 \leq i \leq N u$,

$$d_{BL}\left(\bar{K} H_{K;N}^{\leftarrow}(R_N); \bar{K} H_{K;N}^{\leftarrow}(R_N)\right) := \sup_{b \in BL_1(R)} \left| E_{R_N}\left[b\left(\bar{K} H_{K;N}^{\leftarrow}(R_N)\right)\right] - E_{T_N}\left[b\left(\bar{K} H_{K;N}^{\leftarrow}(T_N)\right)\right] \right|;$$

where $BL_1(R)$ is the set of Lipschitz real functions on R bounded by 1. By virtue of the results in van der Vaart and Wellner (1996, p.73), if this distance tends to 0 as $N \to \infty$ then we have $Q_N^{(4)} \xrightarrow[N \to \infty]{P} 0$. Take $b \in BL_1(R)$, then

$$E_{R_N}\left[b\left(\bar{K} H_{K;N}^{\leftarrow}(R_N)\right)\right] - E_{T_N}\left[b\left(\bar{K} H_{K;N}^{\leftarrow}(T_N)\right)\right] =$$

$$\prod_{s \in PP(U_N)} b\left(\bar{K} H_{K;N}^{\leftarrow}(R_N(s))\right)^{R_N(s)} - \prod_{s \in PP(U_N)} b\left(\bar{K} H_{K;N}^{\leftarrow}(T_N(s))\right)^{T_N(s)}$$

$$\leq \prod_{s \in PP(U_N)} R_N(s) - T_N(s) \leq \prod_{s \in PP(U_N)} |R_N(s) - T_N(s)|;$$

since b is bounded by 1 and for a fixed sample s,

$$b\left(\bar{K} H_{K;N}^{\leftarrow}(R_N(s))\right)^{R_N(s)} = b\left(\bar{K} H_{K;N}^{\leftarrow}(T_N(s))\right)^{T_N(s)}.$$

As established in Berger (1998); Hàjek (1964), the right hand part of this last inequality converges to 0 as $N \to \infty$. Therefore, $Q_N^{(4)} \xrightarrow[N \to \infty]{P} 0$.

Let us now turn to $Q_N^{(5)}$. We have:

$$\begin{aligned} Q_N^{(5)} &= \left(\bar{K} H_{K;N}^{\leftarrow}(T_N) - H_{K;N}^{p}(T_N)\right) \\ &= \frac{\bar{K} D_{K;N}(R_N; T_N) + H_{K;N}^{p}(T_N)^{\bar{K}} r_{K;N}(R_N; T_N)}{r_{K;N}(R_N; T_N) + r_{K;N}(T_N; p)} \\ &\leq \frac{\bar{K} |D_{K;N}(R_N; T_N)| + H_{K;N}^{p}(T_N)^{\bar{K}} |r_{K;N}(R_N; T_N)|}{r_{K;N}(R_N; T_N) + r_{K;N}(T_N; p)}. \end{aligned}$$

We start by analyzing $D_{K;N}(R_N; T_N)$. Observe that it can be written as a function of d_N and $\bar{p}$, which have nice asymptotic properties:

$$\begin{aligned} D_{K;N}(R_N; T_N) &= \frac{1}{K} \sum_{i=1}^{K} \tau_{N-i+1;N} \left(\frac{1}{\tau_{N-i+1;N}} - \frac{1}{p_{N-i+1;N}}\right) \log\left(\frac{X_{N-i+1;N}}{X_{N-K;N}}\right) \\ &= \frac{1}{K} \sum_{i=1}^{K} \frac{\tau_{N-i+1;N}}{p_{N-i+1;N}} \frac{p_{N-i+1;N} - \tau_{N-i+1;N}}{\tau_{N-i+1;N}} \log\left(\frac{X_{N-i+1;N}}{X_{N-K;N}}\right) \\ &= \frac{1}{K} \sum_{i=1}^{K} \frac{\tau_{N-i+1;N}}{p_{N-i+1;N}} \left(\frac{\bar{p}_N - p_{N-i+1;N}}{d_N} + o(1/d_N)\right) \end{aligned}$$

$$\left(p_{N,i+1:N} \left(\frac{1}{\alpha_{N,i+1:N}} - 1 \right) \log \frac{X_{N,i+1:N}}{X_{N,K:N}} \right).$$

Set $\alpha_{\star} := (1 - p_{\star})^{\frac{1}{\alpha_{\star}}} - 1$, then under Assumption 6.1 we have $\forall K \leq N$,

$$\begin{aligned}
|D_{K,N}(R_N, T_N)| &\leq \frac{1}{K} \sum_{i=1}^{K} \frac{\alpha_{N,i+1:N}}{p_{N,i+1:N}} \left| \frac{\alpha_{\star}}{\alpha_{\star}} \right| + o(1/d_N) \left| \log \frac{X_{N,i+1:N}}{X_{N,K:N}} \right| \\
&\leq r_{K,N}(T_N; p) H^{p}_{K,N}(T_N) |\alpha_{\star} + o(1)| \frac{1}{d_N} \\
&\leq \left| r_{K,N}(T_N; p) H^{p}_{K,N}(T_N) - H_{K,N} \right| |\alpha_{\star} + o(1)| \frac{1}{d_N} \\
&\quad + \frac{H_{K,N}}{d_N} |\alpha_{\star} + o(1)|
\end{aligned}$$

Recall that under Assumption 6.10 we have $d_N = o(1/K)$ and that $H_{K,N}$ is a consistent estimator of γ (Mason, 1982). Therefore, by virtue of Lemma 6.19, we can conclude that $D_{K,N}(R_N, T_N) = o_P(1/\sqrt{K})$. Mimicking exactly this procedure and considering the same set of assumptions, we also obtain:

$$|r_{K,N}(R_N, T_N)| \leq r_{K,N}(T_N; p) |\alpha_{\star} + o(1)| \frac{1}{d_N};$$

leading to $r_{K,N}(R_N, T_N) = o_P(1/\sqrt{K})$ by virtue of Lemma 6.17. Combining these two results with Theorem 6.4 yields

$$\sqrt{K} \left(H_{\star,N}(T_N) - H^{p}_{K,N}(T_N) \right) \xrightarrow[N \to \infty]{P} 0;$$

provided that $K = K(N) \to +\infty$ as $N \to +\infty$ and that $K = o(N)$. This concludes the proof. □

BIBLIOGRAPHY

Afssa. INCA 2 (2006-2007), Etude Individuelle Nationale des Consommations Alimentaires 2. Report of the Individual and the National Study on Food Consumption., 2009. URL www.afssa.fr/Documents/PASER-Sy-INCA2EN.pdf.

N. Aires. Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto □ps sampling designs. Methodology and Computing in Applied Probability, 1(4):457–469, 1999.

O. Allais and J. Tressou. Using decomposed household food acquisitions as inputs of a kinetic dietary exposure model. Statistical Modelling, 9(1):27–50, 2009.

M.I.F. Alves, M.I. Gomes, and L. de Haan. A new class of semi-parametric estimators of the second order parameter. Portugal. Math., 60(2):193–214, 2003. ISSN 0032-5155.

Anses. Composition nutritionnelle des aliments, Table CIQUAL 2008, 2008. URL http://www.afssa.fr/TableCIQUAL/.

Anses. Etude de l'alimentation totale française 2 (EAT 2), 2011. URL http://www.anses.fr/cgi-bin/countdocs.cgi?Documents/PASER2006sa0361Ra1.pdf.

Anses. Rapport d'étude de l'Anses relatif à "Substitution du bisphénol A - L'identification des dangers des substituts potentiels au bisphénol A - Etat des lieux sur les alternatives au bisphénol A" en complément du rapport de l'Anses relatif à l'évaluation des risques liés au bisphénol A (BPA) pour la santé humaine et aux données toxicologiques et d'usage des bisphénols S, F, M, B, AP, AF, et BADGE, 25 March 2013. URL http://www.anses.fr/sites/default/files/documents/CHIM2009sa0331Ra-3.pdf.

M.A. Arcones and E. Giné. On the bootstrap of M-estimators and other statistical functionals. Exploring the Limits of Bootstrap, ed. by R. LePage and L. Billard, Wiley, pages 13–47, 1992.

P. Barbe and P. Bertail. The weighted bootstrap, volume 98. Springer Verlag, 1995.

S.M. Barlow, J.B. Greig, J.W. Bridges, A. Carere, A.J.M. Carpy, C.L. Galli, J. Kleiner, I. Knudsen, H.B.W.M. Koeter, L.S. Levy, et al. Hazard identification by methods of animal-based toxicology. Food and Chemical Toxicology, 40(2):145–191, 2002.

C. Béchaux, M. Zetlaoui, J. Tressou, J.C. Leblanc, F. Héraud, and A. Crépet. Identification of pesticide mixtures and connection between combined exposure and diet. Submitted for publication in Food and Chemical Toxicology, 2013.

J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. Statistics of extremes: theory and applications. John Wiley & Sons Inc, 2004. ISBN 0471976474.

J. Beirlant, T. de Wet, and Y. Goegebeur. A goodness-of-fit statistic for Pareto-type behaviour. J. Comput. Appl. Math., 186(1):99–116, 2006. ISSN 0377-0427.

J. Beirlant, F. Caeiro, and M.I. Gomes. An overview and open research topics in statistics of univariate extremes. Revstat, 10:1–31, 2012.

J.C. Bemis and R.F. Seegal. Polychlorinated biphenyls and methylmercury act synergistically to reduce rat brain dopamine content in vitro. Environ. Health Persp., 107 (11):879, 1999.

Y.G. Berger. Rate of convergence to normal distribution for the Horvitz-Thompson estimator. J. Stat. Plan. Inf, 67(2):209–226, 1998.

Y.G. Berger. Pak. J. Statist. 2011 Vol. 27 (4), 407-426 Asymptotic consistency under large entropy sampling designs with unequal probabilities. Pak. J. Statist, 27(4): 407–426, 2011.

P. Bertail and J. Tressou. Incomplete generalized U-statistics for food risk assessment. Biometrics, 62(1):66–74, 2006.

P. Bertail, S. Clémençon, J. Tressou, et al. A storage model with random release rate for modeling exposure to food contaminants. Mathematical biosciences and engineering: MBE, 5(1):35, 2008.

P. Bertail, S. Clémençon, and J. Tressou. Statistical analysis of a dynamic model for dietary contaminant exposure. Journal of Biological Dynamics, 4(2):212–234, 2010.

P. Bertail, E. Chautru, and S. Clémençon. Empirical processes in survey sampling. Submitted to the Scandinavian Journal of Statistics, 2013.

G. Biau and L. Bleakley. Statistical Inference on Graphs. Statistics & Decisions, 24: 209–232, 2006.

N.H. Bingham, C.M. Goldie, and J.L. Teugels. Regular variation. Encyclopedia of Mathematics and its applications. Cambridge Univ Press, Cambridge, 1987.

G. Blom. Some properties of incomplete U-statistics. Biometrika, 63(3):573–580, 1976.

H. Boistard, H.P. Lopuhaä, and A. Ruiz-Gazen. Approximation of rejective sampling inclusion probabilities and application to high order correlations. Electronic Journal of Statistics, 6:1967–1983, 2012.

M.O. Boldi and A.C. Davison. A mixture model for multivariate extremes. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2):217–229, 2007.

D. Bonnery. Propriétés asymptotiques de la distribution d'un échantillon dans le cas d'un plan de sondage informatif. PhD thesis, Université Rennes 1, 2011.

D. Bonnéry, J. Breidt, and F. Coquet. Propriétés asymptotiques de l'échantillon dans le cas d'un plan de sondage informatif. Submitted for publication, 2011.

P.E. Boon, M. Bonthuis, H. van der Voet, and J.D. van Klaveren. Comparison of different exposure assessment methods to estimate the long-term dietary exposure to dioxins and ochratoxin A. Food and Chemical Toxicology, 49(9):1979–1988, 2011.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of Classification: A Survey of Some Recent Advances. ESAIM: Probability and Statistics, 9:323–375, 2005.

G.E.P. Box and D.R. Cox. An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological), pages 211–252, 1964.

N.E. Breslow and J.A. Wellner. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. Scandinavian Journal of Statistics, 35:186–192, 2007.

N.E. Breslow and J.A. Wellner. A Z-theorem with estimated nuisance parameters and correction note for "Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression". Scandinavian Journal of Statistics, 35:186–192, 2008.

N.E. Breslow, T. Lumley, C. Ballantyne, L. Chambless, and M. Kulich. Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. Stat. Biosc., 1:32–49, 2009.

B.M. Brown and D.G. Kildea. Reduced U-statistics and the Hodges-Lehmann estimator. The Annals of Statistics, 6:828–835, 1978.

F. Caeiro and M.I. Gomes. Minimum-variance reduced-bias tail index and high quantile estimation. Revstat, 6(1):1–20, 2008.

J.J. Cai, L. de Haan, and C. Zhou. Bias correction in extreme value statistics with index around zero. Extremes, pages 1–29, 2011.

D.O. Carpenter, K. Arcaro, and D.C. Spink. Understanding the human health effects of chemical mixtures. Environ. Health Persp., 110(Suppl 1):25, 2002.

R.B. Cattell. The scree test for the number of factors. Multivar. Behav. Res., 1(2):245–276, 1966.

S. Clémençon. On U-processes and clustering performance. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems, pages 37–45, 2011.

S. Clémençon and N. Vayatis. Overlaying Classifiers: a practical approach to optimal scoring. Constructive Approximation, 32(3):619–648, 2010.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. The Annals of Statistics, 36(2):844–874, 2008.

W.G. Cochran. Sampling techniques. Wiley, NY, 1977.

U. Cormann and R.D. Reiss. Generalizing the Pareto to the log-Pareto model and statistical inference. Extremes, 12(1):93–105, 2009.

E. Counil, P. Verger, and J.L. Volatier. Fitness-for-purpose of dietary survey duration: A case-study with the assessment of exposure to ochratoxin A. Food and chemical toxicology, 44(4):499–509, 2006.

A. Crépet and J. Tressou. Bayesian nonparametric model with clustering individual co-exposure to pesticides found in the French diet. Bayesian analysis, 6(1):127–144, 2011.

A. Crépet, J. Tressou, V. Graillot, C. Béchaux, S. Pierlot, F. Héraud, and J.C. Leblanc. Accepted for publication in Environmental Research, 2012.

A. Crépet, F. Héraud, C. Béchaux, M.E. Gouze, S. Pierlot, A. Fastier, J.C. Leblanc, L. Le Hégarat, N. Takakura, V. Fessard, et al. The PERICLES research program: an integrated approach to characterize the combined effects of mixtures of pesticide residues to which the French population is exposed. Toxicology, 2013.

S.P. Crispim, J.H.M. de Vries, A. Geelen, O.W. Souverein, P.J.M. Hulshof, L. Lafay, A.S. Rousseau, I.T.L. Lillegaard, L.F. Andersen, I. Huybrechts, et al. Two non-consecutive 24 h recalls using EPIC-Soft software are sufficiently valid for comparing protein and potassium intake between five European centres–results from the European Food Consumption Validation (EFCOVAL) study. British Journal of Nutrition, 105(03):447–458, 2011.

J. Danielsson, L. De Haan, L. Peng, and C.G. De Vries. Using a bootstrap method to choose the sample fraction in tail index estimation. J. Multivariate Anal., 76(2): 226–248, 2001. ISSN 0047-259X.

B. Das and S.I. Resnick. Detecting a conditional extreme value model. Extremes, 14 (1):29–61, 2011.

B. Das, AbhimanyA. Mitra, and S.I. Resnick. Living on the multidimensional edge: seeking hidden risks using regular variation. Advances in Applied Probability, 45(1): 139–163, 2013.

E.J. De Boer, N. Slimani, P. van'T Veer, H. Boeing, M. Feinberg, C. Leclercq, E. Trolle, P. Amiano, L.F. Andersen, H. Freisling, et al. Rationale and methods of the European Food Consumption Validation (EFCOVAL) Project. European journal of clinical nutrition, 65:S1–S4, 2011.

W.J. de Boer, H. van der Voet, B.G.H. Bokkers, M.I. Bakker, and P.E. Boon. Comparison of two models for the estimation of usual intake addressing zero consumption and non-normality. Food Additives and Contaminants, 26(11):1433–1449, 2009.

L. de Haan. Extremes in higher dimensions: the model and some statistics. In In Proceedings of 45th session international statistics institute, (paper 26.3). The Hague: International Statistical Institute. Z. Zhang de, pages 317–337, 1985.

L. De Haan and A. Ferreira. Extreme value theory: an introduction. Springer Verlag, 2006.

L. de Haan and L. Peng. Comparison of tail index estimators. Statist. Neerlandica, 52: 60–70, 1998.

L. de Haan and S. Resnick. On asymptotic normality of the Hill estimator. Stochastic Models, 14:849–867, 1998.

L. de Haan and S. Stadtmüller. Generalized regular variation of second order. J. Austral. Math. Soc. Ser. A, 61:381–295, 1996.

A.L.M. Dekkers, J.H.J. Einmahl, and L. De Haan. A moment estimator for the index of an extreme-value distribution. Ann. Statist., 17(4):1833–1855, 1989.

J.C. Deville. Réplications d'échantillons, demi-échantillons, Jackknife, bootstrap dans les sondages. Economica, Ed. Droesbeke, Tassi, Fichet, 1987.

J.C. Deville and C.E. Särndal. Calibration estimators in survey sampling. JASA, 87: 376–382, 1992.

L. Devroye, L. Györfi, and G. Lugosi. A Probabilistic Theory of Pattern Recognition. Springer, 1996.

I.S. Dhillon, Y. Guan, and J. Kogan. Iterative clustering of high dimensional text data augmented by local search. In Data Mining, 2002. ICDM 2002. Proceedings. 2002 IEEE International Conference on, pages 131–138. IEEE, 2002.

D. Dietrich, L.D. Haan, and J. Hüsler. Testing extreme value conditions. Extremes, 5 (1):71–85, 2002.

K.W. Dodd, P.M. Guenther, L.S. Freedman, A.F. Subar, V. Kipnis, D. Midthune, J.A. Tooze, S.M. Krebs-Smith, et al. Statistical methods for estimating usual intake of nutrients and foods: a review of the theory. Journal of the American Dietetic Association, 106(10):1640–1650, 2006.

D.L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. AMS Math. Challenges Lecture, pages 1–32, 2000.

J.J. Droesbeke, B. Fichet, and P. Tassi. Les sondages. Economica, 1987.

R.M. Dudley. Nonlinear functionals of empirical measures and the bootstrap. In Probability in Banach Spaces 7, pages 63–82. Springer, 1990.

R.M. Dudley. Uniform Central Limit Theorems. Cambridge University Press, 1999.

E. Dybing, J. Doe, J. Groten, J. Kleiner, J. O'Brien, A.G. Renwick, J. Schlatter, P. Steinberg, A. Tritscher, R. Walker, et al. Hazard characterisation of chemicals in food and diet: dose response, mechanisms and extrapolation issues. Food and Chemical Toxicology, 40(2):237–282, 2002.

L. Edler, K. Poirier, M. Dourson, J. Kleiner, B. Mileson, H. Nordmann, A. Renwick, W. Slob, K. Walton, and G. Würtzen. Mathematical modelling and quantitative methods. Food and Chemical Toxicology, 40(2):283–326, 2002.

EFSA. Guidance of the Scientific Committee on a request from EFSA related to Uncertainties in Dietary Exposure Assessment. 14 December 2006. URL http://www.efsa.europa.eu/en/efsajournal/pub/438.htm.

J.H.J. Einmahl and J. Segers. Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. Ann. Statist., 37(5B):2953–2989, 2009.

J.H.J. Einmahl, L. de Haan, and V.I. Piterbarg. Nonparametric estimation of the spectral measure of an extreme value distribution. Ann. Statist., pages 1401–1423, 2001. ISSN 0090-5364.

P. Embrechts, C. Klüppelberg, and T. Mikosch. Modelling extremal events: for insurance and finance, volume 33. Springer, 2011.

E. Enqvist. On sampling from sets of random variables with application to incomplete U-statistics. PhD thesis, 1978.

Max Feinberg, Patrice Bertail, Jessica Tressou, Philippe Verger, et al. Analysis of food risks. Editions Tec & Doc, 2006.

W. Feller. An introduction to probability theory and its applications. Vol. II. . Second edition. John Wiley & Sons Inc., New York, 1971.

C. Fischer, A. Fredriksson, and P. Eriksson. Neonatal co-exposure to low doses of an ortho-PCB (PCB 153) and methyl mercury exacerbate defective developmental neurobehavior in mice. Toxicology, 244(2-3):157–165, 2008.

R.A. Fisher and L.H.C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In Mathematical Proceedings of the Cambridge Philosophical Society, volume 24, pages 180–190. Cambridge Univ Press, 1928.

P.T. Fletcher, C. Lu, S.M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Trans. Med. Imag., 23(8):995–1005, 2004.

J. Friedman, T. Hastie, and R. Tibshirani. The elements of statistical learning, volume 1. Springer Series in Statistics, 2001.

M. Gessaman. A consistent nonparametric multivariate density estimator based on statistically equivalent blocks. Ann. Math. Stat., 41:1344–1346, 1970.

R.D. Gill. Non- and semiparametric maximum likelihood estimators and the von Mises method. Scand. J. Statistics, 22:205–214, 1989.

R.D. Gill, Y. Vardi, and J.A. Wellner. Large sample theory of empirical distributions in biased sampling models. The Annals of Statistics, 16(3):1069–1112, 1988.

E. Giné and J. Zinn. Some limit theorems for empirical processes. The Annals of Probability, 12(4):929–989, 1984.

B.V. Gnedenko. Sur la distribution limite du terme maximum d'une série aléatoire. Ann. Math., 44(3):423–453, 1943.

Y. Goegebeur, J. Beirlant, and T. de Wet. Linking Pareto-tail kernel goodness-of-fit statistics with tail index at optimal threshold and second order estimation. Revstat, 6(1):51–69, 2008.

C.M. Goldie and R.L. Smith. Slow variation with remainder: theory and applications. Quart. J. Math. Oxford, 38(1):45–71, 1987.

M.I. Gomes and O. Oliveira. The bootstrap methodology in statistics of extremes - choice of the optimal sample fraction. Extremes, 4(4):331–358, 2001. ISSN 1386-1999.

C. Gourieroux. Théorie des sondages. Economica, 1981.

C. Gourieroux. Effets d'un sondage: cas du χ^2 et de la régression. Economica, Ed. Droesbeke, Tassi, Fichet, 1987.

S. Guillotte, F. Perron, and J. Segers. Non-parametric Bayesian inference on bivariate extremes. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73 (3):377–406, 2011.

J. Hàjek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. The Annals of Mathematical Statistics, 35(4):1491–1523, 1964.

F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. Robust statistics: the approach based on influence functions. 1986.

H.O. Hartley and J.N.K. Rao. Sampling with unequal probabilities and without replacement. Ann. Math. Statist., 33:350–374, 1962.

S. Haug, C. Klüppelberg, and G. Kuhn. Dimension reduction based on extreme dependence. Submitted for publication, 12, 2010.

J. Heffernan and S.I. Resnick. Hidden regular variation and the rank transform. Adv. in Appl. Probab., 37(2):393–414, 2005.

S. Hercberg, K. Castetbon, S. Czernichow, A. Malon, C. Mejean, E. Kesse, M. Touvier, and P. Galan. The Nutrinet-Santé Study: a web-based prospective study on the relationship between nutrition and health and determinants of dietary patterns and nutritional status. BMC public health, 10(1):242, 2010.

B.M. Hill. A simple general approach to inference about the tail of a distribution. Ann. Statist., 3(5):1163–1174, 1975.

W. Hoeffding. A class of statistics with asymptotically normal distribution. Ann. Math. Stat., 19:293–325, 1948.

J. Hoffmann-Jørgensen. Stochastic processes on Polish spaces. Aarhus Universitet, Denmark, 1991.

D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. JASA, 47:663–685, 1951.

S. Huckemann and H. Ziezold. Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. Adv. in Appl. Probab., 38(2): 299–319, 2006.

J. Hüsler and D. Li. On testing extreme value conditions. Extremes, 9(1):69–86, 2006.

O.A.Y. Jackson. An analysis of departures from the exponential distribution. J. R. Stat. Soc. Ser. B Methodol., 29(3):540–549, 1967.

S. Janson. The asymptotic distributions of Incomplete U-statistics. Z. Wahrsch. verw. Gebiete, 66:495–505, 1984.

P. Jenner, A.H.V. Schapira, and C.D. Marsden. New insights into the cause of Parkinson's disease. Neurology, 42(12):2241, 1992.

S. Jung, M. Foskey, and J.S. Marron. Principal arc analysis on direct product manifolds. Ann. Appl. Statist., 5(1):578–603, 2011.

S. Jung, I.L. Dryden, and J.S. Marron. Analysis of Principal Nested Spheres. Biometrika, 99(3):551–568, 2012.

M.C. Kennedy, V.J. Roelofs, C.W. Anderson, and J.D. Salazar. A hierarchical bayesian model for extreme pesticide residues. Food Chem. Toxicol., 49(1):222–232, 2011.

R. Kroes, D. Muller, J. Lambe, M.R.H. Lowik, J. van Klaveren, J. Kleiner, R. Massey, S. Mayer, I. Urieta, P. Verger, et al. Assessment of intake from the diet. Food and Chemical Toxicology, 40:327–385, 2002.

D.P. Kroese, T. Taimre, and Z.I. Botev. Handbook of Monte Carlo methods. Wiley, 2011.

M Ledoux and M. Talagrand. Probability in Banach spaces: isoperimetry and processes . Springer-Verlag, 1991.

A.J. Lee. U-statistics: Theory and practice. Marcel Dekker, Inc., New York, 1990.

S. Maddipatla, R. Sreenivasan, and V. Rasbagh. On sums of independent random variables whose distributions belong to the max domain of attraction of max stable laws. Extremes, 14:267–283, 2011. ISSN 1386-1999. URL http://dx.doi.org/10.1007/s10687-010-0109-3. 10.1007/ s10687-010-0109-3.

R. Maitra and I.P. Ramler. A k-mean-directions algorithm for fast clustering of data on the sphere. J. Comput. Graph. Statist., 19(2):377–396, 2010.

D.M. Mason. Laws of large numbers for sums of extreme values. Ann. Probab., 10: 756–764, 1982.

P. Massart. Strong approximation for multivariate empirical and related processes, via KMT constructions. Ann. Probab., 17(1):266–291, 1989.

C. McDiarmid. On the method of bounded differences, pages 148–188. Cambridge Univ. Press, 1989.

G. McLachlan and D. Peel. Finite mixture models, volume 299. Wiley-Interscience, 2000.

C. Neves and M.I.F. Alves. Testing extreme value conditions—an overview and recent approaches. Revstat, 6(1):83–100, 2008.

V. Nichèle et al. L'évolution des achats alimentaires: 30 ans d'enquêtes auprès des ménages en France. Cahiers de Nutrition et de Diététique, 43(3):123–130, 2008.

M.J. Paulo, H. Van der Voet, J.C. Wood, G.R. Marion, and J.D. Van Klaveren. Analysis of multivariate extreme intakes of food chemicals. Food Chem. Toxicol., 44(7):994–1005, 2006. ISSN 0278-6915.

D. Pollard. Convergence of stochastic processes. Springer-Verlag, New-York, 1984.

W. Polonik. Minimum volume sets and generalized quantile processes. Stochastic Processes and their Applications, 69(1):1–24, 1997.

O. Pons and E. de Turkheim. Von mises method, bootstrap and Hadamard differentiability for nonparametric general models. Statistics= A Journal of Theoretical and Applied Statistics, 22(2):205–214, 1991.

R.D. Reiss and M. Thomas. Statistical analysis of extreme values: from insurance, finance, hydrology, and other fields. Boston: Basel, 1997.

A.G. Renwick, S.M. Barlow, I. Hertz-Picciotto, A.R. Boobis, E. Dybing, L. Edler, G. Eisenbrand, J.B. Greig, J. Kleiner, J. Lambe, et al. Risk characterisation of chemicals in food and diet. Food and Chemical Toxicology, 41(9):1211–1271, 2003.

S.I. Resnick. Hidden regular variation, second order regular variation and asymptotic independence. Extremes, 5(4):303–336, 2002.

S.I. Resnick. Heavy-tail phenomena: probabilistic and statistical modeling. Springer Verlag, 2007. ISBN 0387242724.

S.I. Resnick. Multivariate regular variation on cones: application to extreme values, hidden regular variation and conditioned limit laws. Stochastics, 80(2-3):269–298, 2008.

C. Rigaux, S. Ancelet, F. Carlin, I. Albert, et al. Inferring an augmented Bayesian network to confront a complex quantitative microbial risk assessment model with durability studies: application to Bacillus Cereus on a courgette purée production chain. Risk Analysis, 2012.

P.M. Robinson. On the convergence of the Horvitz-Thompson estimator. Australian Journal of Statistics, 24(2):234–238, 1982.

P. Rosen. Asymptotic theory for successive sampling. AMS, 43:373–397, 1972.

M. Rueda, S. Martínez, H. Martínez, and A. Arcos. Estimation of the distribution function with calibration methods. Journal of statistical planning and inference, 137(2): 435–448, 2007.

A. Sabourin and P. Naveau. Dirichlet Mixture model for multivariate extremes. 2012.

T. Saegusa and J.A. Wellner. Weighted likelihood estimation under two-phase sampling. Preprint available at http://arxiv.org/abs/1112.4951v1, 2011.

C. Scarrott and A. MacDonald. A review of extreme value threshold estimation and uncertainty quantification. REVSTAT–Statistical Journal, 10(1):33–60, 2012.

C. Scott and R. Nowak. A Neyman-Pearson approach to statistical learning. IEEE Trans. Inf. Theory, 51(8):3806–3819, 2005.

C. Scott and R. Nowak. Learning Minimum Volume Sets. Journal of Machine Learning Research, 7:665–704, 2006.

G. Shorack and J.A. Wellner. Empirical processes with applications to statistics. Wiley, 1986.

V. Sirot, T. Guérin, J.L. Volatier, and J.C. Leblanc. Dietary exposure and biomarkers of arsenic in consumers of fish and shellfish from France. Science of the Total Environment, 407(6):1875–1885, 2009.

I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. J. Machine Learning Research, 6:211–232, 2005.

A. Stephenson. Simulating multivariate extreme value distributions of logistic type. Extremes, 6(1):49–59, 2003.

J.L. Teugels and G. Vanroelen. Box-Cox transformations and heavy-tailed distributions. Journal of Applied Probability, 41:213–227, 2004.

Y. Tillé. Utilisation d'informations auxiliaires dans les enquêtes par sondage. Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa, 23(3):491–505, 1999.

Y. Tillé. Sampling algorithms. Springer Series in Statistics, 2006.

J.A. Tooze, D. Midthune, K.W. Dodd, L.S. Freedman, S.M. Krebs-Smith, A.F. Subar, P.M. Guenther, R.J. Carroll, and V. Kipnis. A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution. Journal of the American Dietetic Association, 106(10):1575–1587, 2006.

J.A. Tooze, V. Kipnis, D.W. Buckman, R.J. Carroll, L.S. Freedman, P.M. Guenther, S.M. Krebs-Smith, A.F. Subar, and K.W. Dodd. A mixed-effects model approach for estimating the distribution of usual intake of nutrients: The NCI method. Statist. Med., 2010.

J. Tressou. Nonparametric modeling of the left censorship of analytical data in food risk assessment. J. Amer. Statist. Assoc., 101(476):1377–1386, 2006.

J. Tressou, A. Crépet, P. Bertail, M.H. Feinberg, and J.C. Leblanc. Probabilistic exposure assessment to food chemicals based on extreme value theory. Application to heavy metals from fish and sea products. Food Chem. Toxicol., 42(8):1349–1358, 2004a. ISSN 0278-6915.

J. Tressou, J.C. Leblanc, M.H. Feinberg, and P. Bertail. Statistical methodology to evaluate food exposure to a contaminant and influence of sanitary limits: application to Ochratoxin A. Regul. Toxicol. Pharm., 40(3):252–263, 2004b.

A. Tsybakov. On nonparametric estimation of density level sets. Annals of Statistics, 25:948–969, 1997.

Office of Pesticide Programs U.S. Environmental Protection Agency. Draft Guidance for Performing Aggregate Exposure and Risk Assessment. 1 February 1999.

S. van de Geer. Empirical processes in M-estimation. Cambridge University Press, 2000.

A.W. Van der Vaart. Asymptotic statistics, volume 3. Cambridge university press, 2000.

A.W. van der Vaart and J.A. Wellner. Weak convergence and empirical processes. Springer, 1996.

J.D. van Klaveren, P.W. Goedhart, D. Wapperom, and H. van der Voet. A European tool for usual intake distribution estimation in relation to data collection by EFSA. 2012. URL http://www.efsa.europa.eu/fr/supporting/doc/300e.pdf.

R. Vert and J.P. Vert. Consistency and convergence rates of one-class SVM and related algorithms. J. Machine Learning Research, 17:817–854, 2006.

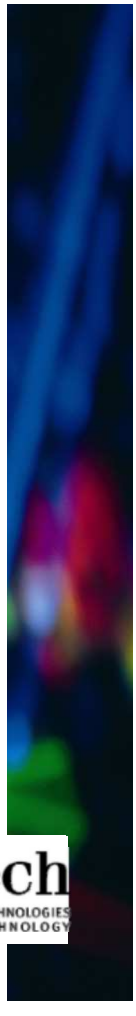
J.L. Wadsworth, J.A. Tawn, and P. Jonathan. Accounting for choice of measurement scale in extreme value modeling. The Annals of Applied Statistics, 4(3):1558–1578, 2010.

P. Weihe, P. Grandjean, F. Debes, and R. White. Health implications for Faroe Islanders of heavy metals and PCBs from pilot whales. Sci. Total Environ., 186(1-2): 141–148, 1996.

WHO. Joint FAO/ WHO Workshop on Methodology for Exposure Assessment of Contaminants and Toxins in Food. 7 - 8 June 2000. URL http://www.who.int/fsf.

RESUME : Véritable carrefour de problématiques économiques, biologiques, sociologiques, culturelles et sanitaires, l'alimentation suscite de nombreuses polémiques. Dans un contexte où les échanges mondiaux facilitent le transport de denrées alimentaires produites dans des conditions environnementales diverses, où la consommation de masse encourage les stratégies visant à réduire les coûts et maximiser le volume de production (OGM, pesticides, etc.) il devient nécessaire de quantifier les risques sanitaires que de tels procédés engendrent. Notre intérêt se place ici sur l'étude de l'exposition chronique, de l'ordre de l'année, à un ensemble de contaminants dont la nocivité à long terme est d'ores et déjà établie. Les dangers et bénéfices de l'alimentation ne se restreignant pas à l'ingestion ou non de substances toxiques, nous ajoutons à nos objectifs l'étude de certains apports nutritionnels. Nos travaux se centrent ainsi autour de trois axes principaux. Dans un premier temps, nous nous intéressons à l'analyse statistique des très fortes expositions chroniques à une ou plusieurs substances chimiques, en nous basant principalement sur des résultats issus de la théorie des valeurs extrêmes. Nous adaptons ensuite des méthodes d'apprentissage statistique de type ensembles de volume minimum pour l'identification de paniers de consommation réalisant un compromis entre risque toxicologique et bénéfice nutritionnel. Enfin, nous étudions les propriétés asymptotiques d'un certain nombre d'estimateurs permettant d'évaluer les caractéristiques de l'exposition, qui prennent en compte le plan de sondage utilisé pour collecter les données.

MOTS-CLEFS : Analyse des risques alimentaires - Apports nutritionnels de long terme - Théorie des valeurs extrêmes - Mesure spectrale - Théorie des sondages - Processus empiriques - Estimation de l'indice de valeurs extrêmes - Ensembles de volume minimum - U-statistiques - Risque-bénéfice



TELECOM ParisTech

EDiTE

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY