Sélection de modèles à l'aide des chemins de régularisation pour l'objectivation mono et multi-prestations. Application à l'agrément de conduite

Jean-Francois Germain

Thèse présentée pour obtenir le grade de
docteur de Télécom ParisTech
Spécialité : Traitement statistique du signal

Jury

Patrice BERTAIL	Président
Jean-Michel POGGI	Rapporteur
Jean-Michel MARIN	Rapporteur
François ROUEFF	Directeur
Eric MOULINES	Directeur
Antoine SAINT-MARCOUX	Membre invité

Sélection de modèles à l'aide des chemins de régularisation pour l'objectivation mono et multi-prestations. Application à l'agrément de conduite.

Jean-François GERMAIN
<germain@telecom-paristech.fr>

Paris, le 2 octobre 2008

Remerciements

Je tiens tout d'abord à remercier l'ensemble des membres du jury pour m'avoir permis de leur présenter mes travaux.

Merci à Monsieur Patrice Bertail pour avoir accepté de le présider.

Merci à Messieurs Jean-Michel Marin et Jean-Michel Poggi pour avoir émis des rapports favorables à l'aboutissement de ce travail de trois années.

Un merci tout particulier à Monsieur Eric Moulines, pour avoir lancé mes travaux de recherche, tant dans le domaine académique que sur les possibles applications dans le domaine industriel.

Je remercie également chaleureusement Monsieur François Roueff pour avoir pris le relai de mon encadrement académique. Merci pour sa rigueur scientifique, pour l'abondance de ses propositions et pour sa grande patience.

Merci à eux deux d'avoir dirigé, avec sympathie, mes recherches au cours de cette thèse.

Enfin, je remercie tout particulièrement Monsieur Antoine Saint-Marcoux, non seulement pour avoir accepté l'invitation à participer à mon jury de thèse, mais également pour son encadrement côté industriel, empreint d'une grande disponibilité. Je garderai en mémoire nos longues, nombreuses et pour le moins passionnantes discussions que nous avons pu avoir durant cette thèse.

Je tiens évidemment à remercier Madame Nadine Ansaldi ainsi que Monsieur Marc Albertelli qui se sont assurés, avant Antoine, des retombées industrielles des travaux de recherche menées avec le laboratoire LTCI. J'ai pu nouer, avec chacun de ces trois tuteurs successifs, des relations épanouissantes dans lesquelles le professionnalisme n'élude jamais l'amicale bienveillance.

Ma pensée va maintenant à tous les collègues qui ont partagé ces trois années avec moi. Merci aux collègues de Renault de la direction de la recherche, des études avancées et des matériaux et des techniques *automobiles* avanées (je raccourcis) pour leur disponibilité et leur bonne humeur permanente malgré le stress inhérent au monde de l'entreprise. C'est une expérience très enrichissante pour moi

que d'avoir évolué pendant trois ans dans un groupe aux compétences diverses et poussées, où chacun est néanmoins tout disposé à apprendre continuellement des choses nouvelles.

J'ai bien sûr une pensée toute particulière pour Claire, Julien, Marine et Rani, mes compagnons de galère sur ce fleuve de la thèse, tantôt trop opaque, tantôt bien trop clair.

Merci aux collègues de Télécom Paris, puis de Télécom ParisTech et plus particulièrement l'équipe STAT du laboratoire LTCI pour leur disponibilité et leur sagacité scientifique à toute épreuve. C'est une expérience unique et très stimulante de travailler au sein d'une équipe où l'excellence de chacun de ses membres se vérifie quotidiennement.

J'ai bien sûr une pensée toute particulière pour Julien, Zaïd, Tabea, Sarah, Natalia, Marine, Cyril, Nancy et Jean-Louis mes compagnons de galère sur ce fleuve de la thèse, tantôt trop violent, tantôt bien trop calme.

Ce ne serait pas honnête de ma part si j'oubliais dans ces remerciements ceux qui m'ont accompagné, réconforté, encouragé, supporté, ou simplement écouté au cours de ces trois années, parfois sans comprendre une once de ce que je leur racontais. Ce sont mes amis les plus proches, Catherine (*first and best*), Marine, Julien, Nathalie, Pauline, Vincent, Silvain, Agnès, Choupi, Marc mais aussi ma famille, en particulier mes chers parents qui m'ont toujours suivi et encouragé. Désolé à tous de vous avoir seriné avec acharnement avec mes travaux auxquels vous ne compreniez parfois pas grand chose. Merci pour votre infinie patience. Merci à tous du plus profond de mon cœur.

Enfin, merci au lecteur anonyme de cette thèse, en espérant qu'il y trouvera ce qu'il est venu y chercher. Qu'il veuille bien me pardonner les éventuelles coquilles tenaces qui subsisteraient encore dans ce document.


ET À QUOI ÇA SERT ?

Table des matières

Chapitre 1

Introduction

1.1 Méthodes classiques d'analyse discriminante

Le cadre principal de nos travaux de recherche concerne l'analyse discriminante, et plus particulièrement la régression logistique. On propose en introduction un rapide panorama des différentes méthodes classiques d'analyse discriminante.

Pour fixer un vocabulaire commun, on parlera dans la suite de variables explicatives et de variable à expliquer (ou de manière équivalente, de variable réponse). La différence principale entre régression et analyse discriminante tient à la nature de la variable réponse. Dans le cadre de la régression, la réponse est une variable continue, pouvant prendre une infinité de valeurs. En analyse discriminante, la variable réponse ne peut prendre qu'un nombre fini de valeurs. On parle alors de classes.

On peut considérer principalement deux buts dans l'analyse discriminante.

- Le premier est de proposer le meilleur modèle permettant de discriminer la variable réponse. Le sens du mot *meilleur* dépend de ce que l'on cherche à étudier. Il s'agit par exemple de trouver une frontière, dans l'espace des variables explicatives, qui sépare les différentes classes de la variable réponse. Dans le cas où la variable à expliquer n'a que deux classes et si l'on postule un modèle de discrimination linéaire, alors on cherche la droite séparatrice, qui classe *au mieux* les observations, en mettant d'un côté les observations appartenant à la première classe et de l'autre côté, les autres observations.
- Le deuxième but de l'analyse discriminante n'est pas complètement déconnecté du premier. Dans ce que l'on vient de voir, on tente de séparer au mieux les données qui sont à notre disposition. La surface séparatrice a été choisie pour obtenir par exemple le moins d'erreurs de classement. Supposons que de nouvelles observations sont maintenant disponibles. Le modèle que l'on

vient de construire nous permet de *prédire* la valeur de la variable réponse. Le second but est celui de la prédiction : on peut vouloir obtenir un modèle qui se trompera le moins possible quand on le teste avec de nouvelles observations. Ainsi, il s'agit, à partir d'un ensemble de données connu - que l'on appelle ensemble d'apprentissage - de construire un modèle que l'on évalue sur un jeu de données qui n'a pas servi à l'apprentissage.

On verra que chacune des méthodes classiques d'analyse discriminantes décrites ci-après sert plutôt l'un ou plutôt l'autre de ces deux buts. La plupart de ces méthodes poursuit en réalité les deux buts de manière simultanée.

Pour chacune des méthodes, on cherche à savoir quel critère est optimisé, comment la surface séparatrice est construite, d'où vient la règle d'affectation, notamment si celle-ci découle directement ou non du calcul de la surface séparatrice et enfin comment est gérée la présence de plus de deux classes pour la variable réponse.

Les méthodes retenues sont les suivantes :

- Analyse discriminante arborescente par moindres écarts
- Support Vector Machine - Séparateur Vaste Marge (SVM)
- Analyse Factorielle Discriminante (AFD)
- Arbre de régression et de discrimination (CART - Classification And Regression Tree)
- Régression logistique
- Régression multiple
- Moindres carrés partiels (PLS - Partial Least Squares)

1.1.1 Notations

On suppose que l'on travaille en présence d'un échantillon constitué de :

- n réalisations d'une variable aléatoire $\boldsymbol{Y}$ à valeurs dans $\mathcal{Y}$. On suppose que $\mathcal{Y}$ est un ensemble discret, éventuellement ordonné.
- n réalisations $(\boldsymbol{x_i}) \in \mathbb{R}^p$ de variables aléatoires. Ces réalisations peuvent également être vues comme p vecteurs colonnes $[\boldsymbol{X_1}, ..., \boldsymbol{X}_p]$ de la matrice $\boldsymbol{X}$.

1.1.2 Méthodes de discrimination avec règle d'affectation liée au critère optimisé

On dit qu'une méthode d'analyse discriminante est *directe* quand elle propose directement un moyen pour construire une règle d'affectation. La règle d'affectation permet d'attribuer à $\boldsymbol{Y}$ une valeur y^{new} calculée à partir d'une nouvelle observation $\boldsymbol{x}^{new}$ des variables explicatives $[\boldsymbol{X_1}, ..., \boldsymbol{X}_p]$. La qualité de la modélisation

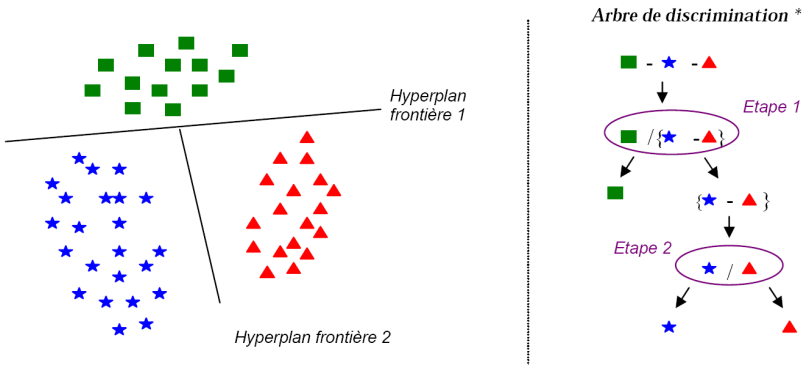

FIG. 1.1 – Analyse discriminante à trois modalités, avec regroupement de modalité

est évaluée par les performances en prédiction de la règle d'affectation.

La règle d'affectation des deux méthodes de discrimination exposées ci-après se déduit directement du critère que l'on optimise dans la construction du modèle.

1.1.2.1 Analyse arborescente par moindres écarts

On commence ce tour d'horizon des méthodes d'analyse discriminante par une méthode qui est loin d'être une méthode classique. Cependant on s'y intéresse car, comme on le verra plus loin, c'est cette méthode qui a été utilisée jusqu'alors et qui est implémentée au sein de RENAULT pour répondre à la problématique que l'on appelle par la suite l'« Objectivation des prestations ». Cette méthode est décrite très en détail dans la thèse de N. Ansaldi [1]. C'est cette méthode d'analyse arborescente par moindres écarts que nous proposons de remplacer par une nouvelle méthodologie. La description de cette nouvelle méthodologie fait l'objet du chapitre 2 et l'implémentation pratique à proprement parler est décrite dans la partie 2.7.

Analyse arborescente L'outil est conçu pour traiter une réponse qualitative $\boldsymbol{Y}$ qui peut prendre un nombre de valeurs discrètes éventuellement plus grand que 2. Lorsque les valeurs possibles sont au nombre de deux, on parle d'analyse binaire. C'est un cas plus simple auquel on cherche à se ramener. Nous expliquons, dans ce paragraphe, comment est réalisée l'analyse arborescente. On cherche tout d'abord comment apparier les différentes modalités de la variable réponse pour traiter les groupes de modalités deux à deux (cf fig. 1.1). Il s'agit d'une technique dite par *regroupement de modalités*.

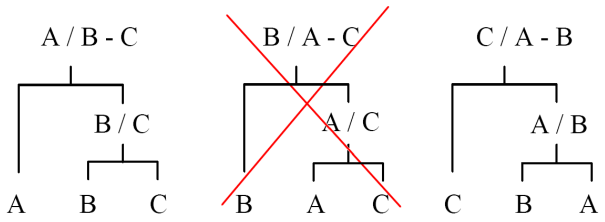

FIG. 1.2 – Analyse discriminante arborescente pour une réponse à trois modalités ordonnées. Le deuxième arbre proposé n'est pas valide.

Par exemple, dans le cas de trois réponses A, B et C, les différents regroupements à envisager sont naturellement ceux présentés fig. 1.2. Comme Y étant ordonné, certains regroupements sont interdits, comme ici le deuxième, si on postule l'ordre $A > B > C$. En effet, il n'y a pas de sens à regrouper les individus de modalités A et C pour les opposer aux individus de modalité B. La notation« A/B » dans la figure 1.2 indique que l'on sépare la modalité A de la modalité B tandis que la notation « A-B » est la modalité issue de la réunion des modalités A et B.

La combinatoire augmente avec le nombre de modalités à considérer. Une réponse à quatre modalités possibles implique le choix entre cinq arbres distincts. L'analyse discriminante arborescente nécessite de parcourir de manière exhaustive tous les arbres de regroupements possibles. Cette étape est très coûteuse en temps de calculs.

Moindres écarts La recherche de l'arbre de regroupement optimal se fait en même temps que la recherche de la meilleure séparation entre deux modalités (ou groupes de modalités). Nous nous intéressons maintenant plus spécifiquement à la discrimination. On se place dans le cas binaire.
L'analyse discriminante mise en œuvre est une analyse par moindres écarts. Cette technique est introduite par Freed et Glover [16] et est inspirée du critère de discrimination binaire MSD (Minimize the Sum of the Deviation) [26].

La méthode cherche à séparer les données à l'aide d'un hyperplan. Le critère optimisé dans le calcul de l'hyperplan séparateur optimal fournit un compromis qui donne le taux de bons classements le plus élevé possible (et par le fait, un taux d'erreurs de classement faible) tout en assurant à la fois une distance à la frontière de séparation maximale et un nombre minimal de variables explicatives présentes dans le modèle, le tout par un jeu de fonctions de pénalisation.

La règle d'affectation est intrinsèquement liée au critère optimisé pour obtenir le modèle. En effet, il suffit de déterminer les positions des nouvelles observations

par rapport à l'hyperplan séparateur pour décider à quelle modalité elles appartiennent. On parle alors de méthode géométrique.

Cette méthode est utilisée en premier lieu pour décrire des données. Mais il est possible de l'utiliser comme méthode de prédiction, grâce à la règle d'affectation construite.

1.1.2.2 Support Vector Machine - SVM

SVM est une méthode finalement assez proche de la méthode par moindres écarts évoquée précédemment. Le concept des SVM est introduit par Vapnik dans [50]. Tout d'abord on peut noter qu'il n'existe pas dans la littérature de méthode SVM globale, de complexité algorithmique raisonnable, permettant le traitement d'une variable $\boldsymbol{Y}$ pouvant prendre plus de deux valeurs. Le cas d'une variable réponse à k modalités, $k > 2$, est également traitée par regroupement de modalités. Dans le cas de la méthode SVM, le regroupement de modalités se fait classiquement en « un contre tous ».

On se place dans le cas à deux modalités. Là aussi, on cherche l'équation d'un hyperplan séparateur. L'hyperplan recherché est celui qui maximise la marge entre les deux nuages. La marge est définie par la distance euclidienne entre les enveloppes convexes des nuages, dans le cas où les nuages sont parfaitement séparables. La méthode peut être facilement généralisée au cas non séparable, par une technique de relâchement de contraintes. En réalité, l'hyperplan *s'appuie* sur des observations particulières qui sont appelées « vecteurs supports », d'où le nom de la méthode.

La règle d'affectation est là encore intrinsèquement liée au critère optimisé pour obtenir le modèle. On procède exactement de la même manière que dans la méthode précédente : ce sont les positions des nouvelles observations par rapport à l'hyperplan séparateur qui déterminent leur modalité.

C'est une méthode géométrique du fait de la règle d'affectation par positionnement par rapport à un hyperplan. De même, cette méthode est d'abord utilisée pour décrire des données. Mais il est possible de l'utiliser comme méthode de prédiction, via la règle d'affectation.

La méthode offre la possibilité de coller aussi près des données que l'on veut grâce à l'introduction de noyaux. L'utilisation de ces noyaux permet de projeter le problème dans un espace de dimension plus grande - voire beaucoup plus grande - afin d'y séparer linéairement les données. L'inconvénient majeur de coller trop près aux données est évidemment la perte en généralisation du modèle obtenu. Le risque est d'apprendre par cœur les données. On parle alors de sur-apprentissage. Dans l'espace initial, la frontière obtenue par l'introduction des noyaux n'est plus un hyperplan mais une surface.

1.1.3 Méthodes de discrimination avec règle d'affectation non liée au critère optimisé

Les trois méthodes qui suivent fournissent également une règle d'affectation explicite, mais celle-ci ne découle pas du critère que l'on optimise pour construire le modèle.

1.1.3.1 Analyse Factorielle Discriminante - AFD

L'analyse factorielle discriminante (voir [8] pour des détails sur cette méthode) cherche la transformation de l'espace des variables explicatives qui maximise la variance inter-classe tout en minimisant la variance intra-classe. Il s'agit de compacter au maximum les observations d'une même modalité (minimisation de la variance intra-classe) et d'écarter les nuages de modalités différentes le plus possible les uns des autres (maximisation de la variance inter-classe).

La règle d'affectation n'a pas de lien direct avec le critère optimisé. Elle est purement géométrique. En effet, une nouvelle observation se voit affecter la modalité du nuage dont le centre de gravité est le plus proche de la nouvelle observation.

L'AFD est utilisée en premier lieu pour décrire les données : on s'attache davantage à la représentation lisible des données plutôt qu'à la prédiction de la modalité d'observations futures.

On peut noter enfin que cette méthode est assez sensible aux observations aberrantes, du fait de l'utilisation de moyennes pour déterminer les centres de gravité des nuages de même modalité.

1.1.3.2 Arbre de régression et de discrimination (CART)

Cette dernière méthode de discrimination directe propose une approche séquentielle qui intègre le problème délicat de la sélection de variables. Le concept de cette méthode a été introduit par Breiman et al. en 1984 (voir [6] pour les détails). La méthode est arborescente. On cherche à partitionner l'échantillon initial par divisions successives jusqu'à obtenir un arbre élagué, optimal au sens du minimum d'erreurs de classement. Cet arbre optimal est appelé *arbre de décision*.

En présence d'une nouvelle observation, on parcourt l'arbre de décision, nœud après nœud, en partant de la racine jusqu'à ce que l'observation arrive dans un nœud terminal. Chaque nœud contient une règle de décision du type : si $x_j > seuil$ alors descendre dans le nœud fils 1, sinon descendre dans le nœud fils 2. Un nœud terminal est étiqueté par la modalité la plus représentée dans ce nœud terminal. Une même modalité peut donc étiqueter plusieurs nœuds terminaux. Le modèle obtenu est donc un ensemble de règles de décision.

Le critère optimisé est la réduction de l'impureté des sous-échantillons descendants. Il n'y a effectivement pas de rapport avec la règle d'affectation. On note que celle-ci se rapproche d'une règle probabiliste.

La représentation graphique qui en découle donne une partition de l'espace des variables explicatives résultant de cette succession de divisions parallèles aux axes.

Alors que les méthodes précédentes sont plutôt tournées vers une description des données, la méthode CART est davantage tournée vers la prédiction. Certes on cherche à décrire les observations de l'ensemble d'apprentissage, mais l'accent est tout particulièrement mis sur l'obtention de règles de décisions afin d'affecter une modalité aux observations futures.

On peut noter que la méthode est peu robuste. En effet, l'ajout et/ou le retrait d'une variable explicative et/ou d'une observation peut modifier complètement la forme du modèle final. Des techniques existent pour rendre la méthode plus robuste. Il s'agit du Bagging ou du Boosting. L'inconvénient est que l'on n'a plus accès à une interprétation simple des règles de décisions obtenues. On perd en description ce que l'on gagne en prédiction. De plus, graphiquement, on obtient un pavage irrégulier de formes dont les côtés sont parallèles aux axes.

1.1.3.3 Régression logistique

La régression logistique (introduite par Luce en 1959 [31]) propose de traiter la variable réponse qualitative de manière astucieuse. Elle s'intéresse à la probabilité qu'une observation soit d'une certaine modalité. Cela revient à transposer le problème et à chercher les expressions des probabilités d'appartenance à chacune des modalités.

La méthode est basée sur le maximum de vraisemblance par rapport aux données d'apprentissage. La régression logistique fera l'objet d'un développement particulier, dans le paragraphe 1.2.1.

La règle d'affectation n'est pas imposée par la méthode. En cela, la régression logistique n'est pas une méthode directe. Cependant, il existe une façon de procéder que l'on peut appeler *canonique* : une fois réalisé l'apprentissage des probabilités d'appartenance, l'affectation d'une observation se calcule par comparaison de ces probabilités. La nouvelle observation est alors affectée à la classe la plus probable. On peut toutefois procéder de manière moins tranchée. Par exemple, si l'on se place dans le cas de deux classes, « bon » et « mauvais », on peut décider qu'une observation sera classée comme « bonne » si la probabilité pour cette observation d'être « bonne » est supérieure à 0.9, et non 0.5, comme c'est le cas pour l'affectation canonique dans le cas binaire.

On peut également représenter les résultats à l'aide d'hyperplans séparateurs. Ces hyperplans séparateurs sont alors les surfaces d'équi-probabilité.

Le traitement de la multi-modalité est également possible directement. Rien n'impose de revenir à une analyse « un contre tous » ni à réaliser des regroupements de modalité.

Là encore, même si le premier intérêt de cette méthode n'est pas de décrire les données - car on cherche à estimer les probabilités d'appartenance qui servent à prédire l'affectation de futures observations - on récupère néanmoins un modèle explicatif, dans le sens où il affecte des probabilités d'appartenance à des observations dont on connaît déjà la modalité.

1.1.4 Méthodes de discrimination sans règle d'affectation

Une méthode de discrimination indirecte est une méthode qui ne fournit pas directement de règle d'affectation pour de nouvelles observations. La modalité que l'on affecte à une nouvelle observation peut résulter par exemple d'un calcul, que l'on compare souvent à un seuil. En ce sens, la régression logistique est une méthode indirecte. Mais, comme on vient de le voir, il existe une construction *canonique* de cette règle d'affectation.

1.1.4.1 Régression multiple

La régression multiple est une méthode extrêmement classique et bien connue. En dimension un, il s'agit de trouver la droite qui passe *au plus près* d'un nuage de points, au sens des moindres carrés des écarts. Dans le cas général, dans un espace de dimension p, on cherche l'hyperplan qui régresse au mieux les observations.

Le passage à la discrimination n'est valable que pour une variable réponse qualitative binaire. Quel que soit le système de codage de la variable réponse, on obtient une équation de régression. Il suffit ensuite de fixer un seuil pour obtenir la discrimination entre les modalités.

Le choix de fixer le seuil à l'exact milieu des isobarycentres des deux modalités revient à la même règle d'affectation que pour l'AFD. Par contre, le critère optimisé n'est pas le même. Dans le cas de la régression multiple, on ne s'intéresse pas aux variances des observations intra ou inter-classe mais aux carrés des écarts à l'hyperplan de régression.

Le traitement d'une variable réponse qualitative à plus de deux modalités par régression multiple est fortement déconseillé du fait que les modalités ont besoin d'être recodées par des valeurs numériques. Le fait de coder en numérique implique qu'il existe un ordre entre les modalités de la variable réponse, ce qui n'est pas toujours le cas. De plus, le choix du codage a une forte influence sur le résultat de la méthode.

Notons également que cette méthode est très sensible aux valeurs aberrantes.

1.1.4.2 Moindres carrés partiels - PLS

On attribue à Wold l'origine de cette méthode [51]. Un tutorial sur la régression PLS est donné dans [17].

Les méthodes PLS sont des heuristiques. La régression PLS [45] propose une méthode de traitement d'une variable réponse qualitative et l'approche PLS [44] proposer une méthode pour traiter plusieurs variables réponses qualitatives.

Les deux heuristiques proposent la construction itérative de composantes factorielles qui améliorent successivement la qualité de l'explication. Comme c'est le cas en analyse des composantes principales (ACP), on améliore l'explication de la variable réponse en cherchant successivement les composantes PLS des résidus courants.

Ces méthodes ont pour but de décrire les données. Le champ d'application le plus courant de ces méthodes est le marketing. Le modèle obtenu n'est pas conçu en vue de prédire l'affectation de nouvelles observations. La construction d'une règle d'affectation n'est pas décrite dans la méthode et n'apparaît pas à première vue comme quelque chose de simple, notamment en multimodalité et/ou en multi-réponses.

Les méthodes PLS ont toutes la même philosophie : le critère optimisé est à mi-chemin entre description des données et explication de la variable réponse. Cependant du fait du caractère intrinsèquement non prédictif de la méthode, on la considère plutôt comme une approche exploratoire.

1.1.5 Méthode de discrimination sur variables fonctionnelles

On terminera cette partie de l'introduction avec une méthode d'analyse discriminante qui traite des variables fonctionnelles. Cette méthode développée par Poggi et Tuleau dans [37] entre dans le cadre des travaux sur l'objectivation des prestations menée chez Renault (voir partie 2 pour la description du contexte industriel), c'est pourquoi l'on s'y intéresse dans cette introduction aux méthodes statistiques d'analyse discriminante. Les résultats obtenus par les auteurs constituent actuellement la base d'une réflexion sur la mise en place d'une solution logicielle permettant de traiter des variables discriminantes fonctionnelles comme données d'entrée. Cette solution logicielle pourrait vraisemblablement faire l'objet chez Renault d'un module supplémentaire au sein du logiciel interne (PRESTool) de traitement des prestations. La méthode d'analyse discriminante sur variables vectorielles actuellement implémentée dans l'outil PRESTool est basée sur une analyse discriminante par moindres écarts, méthode que nous avons décrite dans le paragraphe 1.1.2.1. On pourra également se reporter au paragraphe 2.7.1 pour des détails sur les évolutions que nous avons apportées à l'outil logiciel interne. On

note également que al méthode proposée par les auteurs réalise une sélection des variables pertinentes, tout comme c'est déjà le cas dans la méthode implémentée chez Renault.

Les variables d'entrée ne sont plus des critères physiques issus de courbes de mesure, mais les courbes elles-mêmes. Pour tout $i \in \{1, \ldots, n\}$ et $j \in \{1, \ldots, J\}$, on note $X_i^j(t)$ le $j^{\text{ème}}$ signal mesuré lors de l'essai i. Le vecteur des réponses est noté Y.

La démarche adoptée est la suivante. Il s'agit de construire une fonction de prédiction F de la réponse Y à partir des variables fonctionnelles. Les auteurs proposent de sélectionner un nombre restreint de variables pour expliquer cette réponse puis, pour chacune d'elle, de ne retenir qu'un petit nombre de descripteurs, notés C^{j_k}, la décrivant. On obtient alors un vecteur de prédiction de Y :

$$\hat{Y} = F(C^{j_1}, \ldots, C^{j_K})$$

avec $K \ll J$, par exemple de l'ordre de 5.

Les trois étapes proposées dans la méthodologie sont les suivantes :

- mise en forme des signaux ;
- description de chacune des variables sur une base d'ondelettes commune et compression de la taille des variables fonctionelles ;
- sélection des variables puis des critères pertinents par applications successives de la méthode CART.

On rappelle que la méthode CART a été briévement introduite dans le paragraphe 1.1.3.2 de cette même introduction.

Mise en forme des signaux. Cette étape consiste tout d'abord à tronquer les signaux, à l'aide d'une connaissance extérieure donnant les instants de début et de fin de la zone d'intérêt pour chacune des variables fonctionnelles.

Les signaux subissent ensuite un débruitage adaptatif en espace, en l'occurence ici une décomposition sur une base d'ondelettes. L'ondelette utilisée est l'ondelette de Daubechies presque symétrique d'ordre 4, avec un niveau de décomposition entre 3 et 5 (seuillage « universel » de Donoho et Johnstone [13]).

La dernière étape dans la mise en forme des signaux consiste en la synchronisation des signaux, c'est-à-dire à ramener tous les signaux $\left(X_i^j(t)\right)_{1 \leq i \leq n}$ sur une même grille temporelle, de m points régulièrement espacé sur $[0, 1]$. Cette synchronisation passe par un recalage linéaire en temps puis par une interpolation linéaire des signaux. Les nouvelles abscisses des signaux s'interprêtent alors comme des proportions de la durée de l'essai écoulée.

Compression des signaux. Pour réaliser cette compression des données, on décompose les variables fonctionnelles sur une base d'ondelette. L'ondelette utilisée ici est également l'ondelette de Daubechies presque symétrique d'ordre 4.
Pour chaque variable fonctionnelle X^j, le niveau de décomposition retenu est choisi de la manière suivante : pour un signal d'origine $X_i^j(t)$, on reconstruit le signal $A_i^j(t)$ à partir des coefficients d'approximation du nieau p ; on définit ensuite l'erreur pour la variable fonctionnelle numéro j comme suit :

$$EQ_j(p) = \sum_{i=1}^{n} \|X_i^j(t) - A_i^j(t)\|^2$$

Le niveau de décomposition retenu K_j est celui pour lequel la fonction $p \mapsto EQ_j(p)$ présente un changement de pente « significatif ». Une unité est également retranchée à titre conservatoire.

Sélection des variables et critères pertinents par CART. Pour tout j, on construit, avec la méthode CART, l'arbre de classification A^j qui explique la réponse Y grâve aux coefficients d'ondelettes $\left(C^j\right)_{1\leq j\leq K_j}$ retenus.
On sélectionne parmi ces coefficients ceux qui une importance des variables, au sens de Breiman et al. [6], supérieure à 50%. On note ces critères $\left(\widetilde{C}^j\right)_{1\leq j\leq K_j}$.
On ordonne ces paquets de coefficients $\left(\widetilde{C}^j\right)_{1\leq j\leq K_j}$ au moyen de l'erreur de classification de l'arbre A^j.
On construit alors les modèles emboîtés M^j de la manière suivante :

$$M^j = \cup_{l\leq j}\{\widetilde{C}^l\}$$

On calcule pour chaque modèle M^j l'erreur commise lors de la classification et on appelle M^{j_0} le modèle qui minimise cette erreur.
Les variables fonctionnelles pertinentes sont celles retenues dans le modèle M^{j_0}.
Par la méthode CART, on construit l'arbre de classification qui explique la réponse Y grâce aux critères du modèle M^{j_0}, à savoir $\cup_{j<=j_0}\widetilde{C}^j$.
Les critères pertinents sont ceux de plus grande importance, au sens de l'importance des variables de Breiman.

1.2 Sélection de modèle via les chemins de régularisation

1.2.1 Régression logistique

On se place dans le cadre des modèles linéaires généralisés en soulignant que les applications que l'on sera amené à traiter sont des cas d'analyse discriminante.

On précise l'échantillon sur lequel on travaille. Supposons que l'on a :

- n réalisations $(y_i) \in \mathcal{Y}^n$ d'une variable aléatoire que l'on cherche à expliquer et/ou à prédire. Cette variable est observée. On note cette variable $\boldsymbol{Y}$,
- n réalisations $(\boldsymbol{x_i}) \in \mathbb{R}^p$ de variables aléatoires. Ces variables explicatives peuvent être vues comme une variable de dimension $p \geq 1$. Cette variable est également observée et on la note $\boldsymbol{X}$,
- n réalisations (ε_i) de variables aléatoires non observées et que l'on note $\boldsymbol{\varepsilon}$.

Le lien entre les réalisations de ces variables est le suivant :

$$y_i = h(\boldsymbol{x_i}, \varepsilon_i)$$

où h est appelée fonction de décision. C'est une fonction déterministe, à savoir qu'à $\boldsymbol{x_i}$ et ε_i donnés, la valeur de la variable aléatoire $\boldsymbol{Y}$ est entièrement déterminée et vaut y_i.

Les (ε_i) ne sont pas observés ; de fait, on ne peut pas prédire parfaitement $\boldsymbol{Y}$. A la place, on se propose d'estimer la probabilité que $\boldsymbol{Y}$ prenne une certaine valeur. A l'aide de la fonction de décision h, la probabilité que $\boldsymbol{Y}$ ait une valeur donnée y_i est simplement la probabilité que ε_i soit tel que la fonction de décision rende comme résultat y_i :

$$\mathbb{P}\left(\boldsymbol{Y} = y_i | \boldsymbol{X} = \boldsymbol{x_i}\right) = \mathbb{P}\left(\varepsilon_i \text{ t.q. } h(\boldsymbol{x_i}, \varepsilon_i) = y_i\right)$$

Considérons la fonction indicatrice $\mathbb{I}_{\{h(\boldsymbol{x_i}, \varepsilon_i) = y_i\}}$ qui renvoit 1 si l'affirmation entre accolades est vraie et 0 sinon. Cela signifie que : si $\mathbb{I}_{\{.\}} = 1$ alors la valeur prise par ε_i, en combinaison avec $\boldsymbol{x_i}$, donne à $\boldsymbol{Y}$ la valeur y_i ; et si $\mathbb{I}_{\{.\}} = 0$ alors la valeur prise par ε_i, en combinaison avec $\boldsymbol{x_i}$, donne à $\boldsymbol{Y}$ une autre valeur que y_i. Donc la probabilité que $\boldsymbol{Y}$ ait comme valeur y_i est simplement l'espérance de cette fonction indicatrice, quand ε_i parcourt l'ensemble des valeurs possibles.

$$\begin{aligned} \mathbb{P}\left(\boldsymbol{Y} = y_i | \boldsymbol{X} = \boldsymbol{x_i}\right) &= \mathbb{E}\left(\mathbb{I}_{\{h(\boldsymbol{x_i}, \varepsilon_i) = y_i\}} = 1\right) \\ &= \int \mathbb{I}_{\{h(\boldsymbol{x_i}, u) = y_i\}} f(u) \partial u \end{aligned} \tag{1.1}$$

où f est la densité des $\boldsymbol{\varepsilon}$.

Remarque :

- Dans le cas de la régression logistique binaire, la variable aléatoire $\boldsymbol{Y}$ ne peut prendre que deux valeurs, par exemple faire une action ($y_i = 1$) ou ne pas la faire ($y_i = 0$). La fonction de décision est spécifiée de la manière suivante : on définit l'*utilité* associée à chacune des deux décisions possibles pour $\boldsymbol{Y}$.

Cette utilité U est reliée aux variables explicatives par $U = \boldsymbol{X}^T\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. On choisit de prendre la décision de faire l'action si l'utilité est positive. La probabilité de faire l'action, sachant les variables qui sont observées, est alors :

$$\mathbb{P}\left(y_i = 1 | \boldsymbol{X} = \boldsymbol{x_i}\right) = \int \mathbb{I}_{\{\boldsymbol{x_i}^T\boldsymbol{\beta}+u>0\}} f(u)\partial u$$

Supposons que $\boldsymbol{\varepsilon}$ soit distribué de manière logistique, *i.e.* de densité :

$$f(u) = \frac{e^{-u}}{(1+e^{-u})^2}$$

et de fonction de répartition :

$$F(u) = \frac{1}{1+e^{-u}}$$

Il vient que la probabilité de faire l'action s'écrit :

$$\begin{aligned}
\mathbb{P}\left(y_i = 1 | \boldsymbol{X} = \boldsymbol{x_i}\right) &= \int \mathbb{I}_{\{\boldsymbol{x_i}^T\boldsymbol{\beta}+u>0\}} f(u)\partial u \\
&= \int \mathbb{I}_{\{u>-\boldsymbol{x_i}^T\boldsymbol{\beta}\}} f(u)\partial u \\
&= \int_{u=-\boldsymbol{x_i}^T\boldsymbol{\beta}}^{+\infty} f(u)\partial u \\
&= 1 - F(-\boldsymbol{x_i}^T\boldsymbol{\beta}) = 1 - \frac{1}{1+e^{\boldsymbol{x_i}^T\boldsymbol{\beta}}} \\
&= \frac{e^{\boldsymbol{x_i}^T\boldsymbol{\beta}}}{1+e^{\boldsymbol{x_i}^T\boldsymbol{\beta}}}
\end{aligned}$$

De manière plus générale, supposons que les (y_i) suivent une distribution connue, de paramètre à estimer θ. (θ n'est pas nécessairement un vecteur monodimensionnel.) Si l'on note $f_\theta(.)$ la densité de probabilité associée à cette distribution, la fonction de vraisemblance a alors la forme suivante :

$$L_\theta(\boldsymbol{x_1}, ..., \boldsymbol{x_n}, y_1, ..., y_n) = L_n(\theta) = \prod_{i=1}^{n} f_\theta\left(\boldsymbol{x_i}, y_i\right)$$

Maximiser la vraisemblance, c'est maximiser les probabilités des réalisations observées. On cherche ainsi le vecteur de paramètres θ qui maximise la vraisemblance.

Dans le cas particulier de la discrimination, si on appelle $\mathcal{Y}$ l'ensemble des valeurs que peut prendre la variable $\boldsymbol{Y}$, la fonction de vraisemblance s'écrit :

$$L_n(\theta) = \prod_{i=1}^{n} \prod_{y_i \in \mathcal{Y}} \mathbb{P}_\theta \left(\boldsymbol{Y} = y_i | \boldsymbol{X_i} = \boldsymbol{x_i} \right)$$

Dans le cas de la régression logistique binaire, les y_i ne peuvent prendre que deux valeurs, par exemple 0 et 1. La vraisemblance prend alors la forme suivante :

$$\begin{aligned} L_n^{logit}(\theta) &= \prod_{i=1}^{n} \left[\mathbb{P} \left(y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x_i} \right) \right]^{y_i} \left[\mathbb{P} \left(y_i = 0 | \boldsymbol{X}_i = \boldsymbol{x_i} \right) \right]^{1-y_i} \\ &= \prod_{i=1}^{n} \left[\mathbb{P} \left(y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x_i} \right) \right]^{y_i} \left[1 - \mathbb{P} \left(y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x_i} \right) \right]^{1-y_i} \end{aligned}$$

En prenant l'opposé du logarithme de la vraisemblance, on obtient :

$$-\log L_n^{logit}(\theta) = \sum_{i=1}^{n} -y_i \log \frac{\mathbb{P} \left(y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x_i} \right)}{1 - \mathbb{P} \left(y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x_i} \right)} - \log \left(1 - \mathbb{P} \left(y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x_i} \right) \right)$$

Si l'on se place dans le cadre des modèles linéaires généralisés, on fait alors la première hypothèse suivante :

$$\frac{\mathbb{P} \left(y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x_i} \right)}{1 - \mathbb{P} \left(y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x_i} \right)} = \theta_i \tag{1.2}$$

La vraisemblance dans le cas de la famille exponentielle prend alors la forme générale suivante :

$$\begin{aligned} L(\boldsymbol{Y}, \theta, \phi) &= \prod_{i=1}^{n} \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right\} \\ &= \exp \left\{ \sum_{i=1}^{n} \left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right) \right\} \end{aligned} \tag{1.3}$$

Dans le cas de la régression logistique, on fait l'hypothèse supplémentaire que θ_i est linéaire en les variables explicatives : $\theta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. On en tire directement la forme que prend le modèle logistique :

$$\mathbb{P} \left(y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x_i} \right) = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}} \tag{1.4}$$

Pour récapituler, on dira que le but de la régression logistique est de trouver le vecteur de paramètres qui maximise la fonction de vraisemblance :

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \left(L_n^{logit}(\boldsymbol{\beta}) \right)$$

où

$$L_n^{logit}(\boldsymbol{\beta}) = \exp\left\{ \sum_{i=1}^{n} y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - \log\left(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right) \right\}$$

Cette expression est typique de la famille exponentielle, dont l'expression est donnée par l'équation (1.3), avec ici :

- $\theta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$,
- a est la fonction constante égale à 1,
- c est la fonction nulle et
- $b(\theta_i) = -\log\left(1 + e^{\theta_i}\right)$.

1.2.2 Sélection de modèle

Les données sont constituées de variables explicatives et d'une variable réponse, que l'on cherche à expliquer. Un modèle est un sous-ensemble de variables explicatives. Comme on dispose de p variables explicatives, il y a 2^p modèles possibles. La sélection de modèle consiste en un choix de certaines variables explicatives. Ce choix est motivé par la recherche du meilleur modèle ainsi obtenu. La définition du *meilleur modèle* dépend de la problématique traitée.

Classiquement, le meilleur modèle est choisi pour être le modèle le plus prédictif, c'est-à-dire celui qui assure le plus faible taux d'erreurs de prédiction. Cependant, une bonne explication des données peut être un objectif en soi, comme c'est le cas pour les applications pratiques en objectivation des prestations (voir la définition dans le chapitre 2). Une fois les données correctement expliquées, il est évidemment important de fournir une fonction de prédiction, afin de pouvoir valider la modélisation.

Notre procédure propose, comme on le verra plus en détails dans le chapitre 2, d'établir une hiérarchie entre les variables explicatives. On se sert ensuite de cette hiérarchie pour constituer une suite de modèles emboîtés. Cette procédure tend donc à sélectionner les variables explicatives qui ont une part significative dans l'explication de la variable réponse. En cela, on répond aux attentes premières d'explication de la réponse.

Quant à la fonction de prédiction, on propose, parmi cette suite de modèles, de sélectionner le *meilleur* modèle. On obtient ainsi une modélisation de la réponse et l'on peut, via ce modèle, prédire la modalité d'une nouvelle observation. Cependant, on note que la vraisemblance des modèles (ou un autre critère d'attache aux

données) croît à mesure que l'on considère davantage de variables explicatives, cela sans nécessairement réduire l'erreur de prédiction. Les méthodes par pénalisation proposent une solution à ce phénomène qui consiste à remplacer le critère de vraisemblance par un critère de vraisemblance pénalisée. En effet, la vraisemblance pénalisée n'atteint pas nécessairement son maximum avec le modèle comportant le plus grand nombre de variables explicatives.

Le choix de la pénalisation est un problème difficile. On peut citer les travaux de Birgé et Massart sur la recherche de fonctions de pénalisation pour lesquelles les estimateurs pénalisés correspondants vérifient des inégalités d'oracle (voir [3],[4] pour les détails). Il n'en reste pas moins que les fonctions de pénalisation proposées par Birgé et Massart introduisent des constantes et que le choix pratique de ces constantes est un problème difficile.

1.2.2.1 Régression Ridge et Lasso

Il est assez classique de prendre pour fonction de pénalisation une fonction de la taille des paramètres. On peut citer le problème de la *régression Ridge* qui fournit un estimateur des moindres carrés pénalisés par la somme des carrés des paramètres.

$$\widehat{\boldsymbol{\beta}}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \boldsymbol{x_i}^T \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} ,$$

si l'on suppose $\boldsymbol{Y}$ centré. λ est appelée constante de régularisation.

Cette expression est équivalente à l'expression suivante :

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{ridge} &= \arg\min_{\beta} \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \left(y_i - \boldsymbol{x_i}^T \boldsymbol{\beta} \right)^2 , \\ &\text{tel que } \textstyle\sum_{j=1}^{p} \beta_j^2 \leq s, s \geq 0 \end{aligned} \tag{1.5}$$

où il existe un lien de dépendance entre s et λ.

Si l'on suppose que l'on est en présence de nombreuses variables corrélées entre elles, l'identification des paramètres de la régression linéaire est difficile. Le problème est mal déterminé et donne une grande variance. Un coefficient positif exceptionnellement grand sur une variable peut être contre-balancé par un coefficient négatif aussi grand sur une variable corrélée avec la première. La contrainte sur la taille des paramètres (équation (1.5)) empêche l'apparition de ce phénomène.

Plutôt que de s'intéresser à la pénalisation L_2, comme c'est le cas dans la régression Ridge, on considère la pénalisation L_1, ce pour de meilleures performances en terme de parcimonie [22, page 72]. Un modèle est parcimonieux quand le vecteur de paramètres $\boldsymbol{\beta}$ n'a seulement que quelques composantes non nulles.

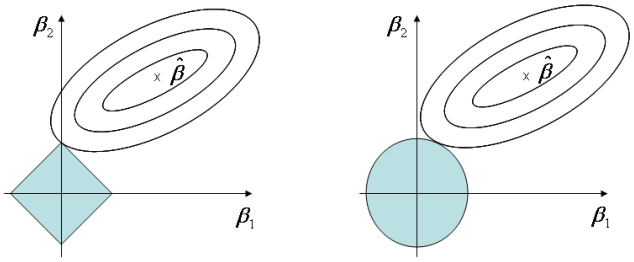

FIG. 1.3 – Illustration du Lasso (à gauche) et de la régression Ridge (à droite). Les zones bleues sont les régions de contrainte où $|\beta_1| + |\beta_2| \leq t$ et $\beta_1^2 + \beta_2^2 \leq t^2$ respectivement. $\widehat{\beta}$ est le point où l'erreur quadratique est nulle. Les ellipses sont les courbes d'iso-erreur quadratique.

On reproduit en fig.1.3 la figure 3.12 de l'ouvrage de Hastie et al. [22] (illustration que l'on trouve initialement dans l'article [46], fig. 2).

Le Lasso (pour Least Absolute Schrinkage and Selection Operator) est le nom donné par Tibshirani (voir [46]), pour décrire cette procédure basée sur les moindres carrés pénalisés L_1, qu'il étudie dans un contexte de sélection de variables. Sur la fig.1.3, sont représentés le Lasso et la régression Ridge dans le cas à deux variables. Sur la figure, $\widehat{\beta}$ est le β pour lequel l'erreur quadratique est nulle. Ce point n'est pas réalisable, car il n'est pas dans la zone de contrainte. Il faut donc s'éloigner de $\widehat{\beta}$ pour chercher une solution β qui appartienne à la zone de contrainte. Pour une erreur quadratique donnée $||X\beta - Y||^2 = c$, l'ensemble des β est une ellipse. La recherche d'un point réalisable consiste à trouver la plus petite ellipse qui intersecte la zone de contrainte. On fait donc croître l'ellipse jusqu'à rencontrer cette zone de contrainte. Contrairement à la boule unité de la distance L_2, qui est un disque, la boule unité de la distance L_1 a une forme de diamant. Si la rencontre a lieu sur un des coins, alors il y a un coefficient β_j égal à zéro. Quand $p > 2$, le diamant devient un rhomboïde. Avec ses nombreux coins, il y a beaucoup plus de chances d'avoir des coefficients à zéro.

Une autre façon de présenter cette propriété se trouve notamment dans [5]. On reproduit en fig.1.4 les histogrammes des β_j après estimation via le Lasso (en haut) et via la régression Ridge (en bas). La présence d'un point angulaire dans la fonction de pénalisation dans la procédure du Lasso « attire » les coefficients β_j vers zéro (on peut penser à l'algorithme de descente du gradient, par exemple).

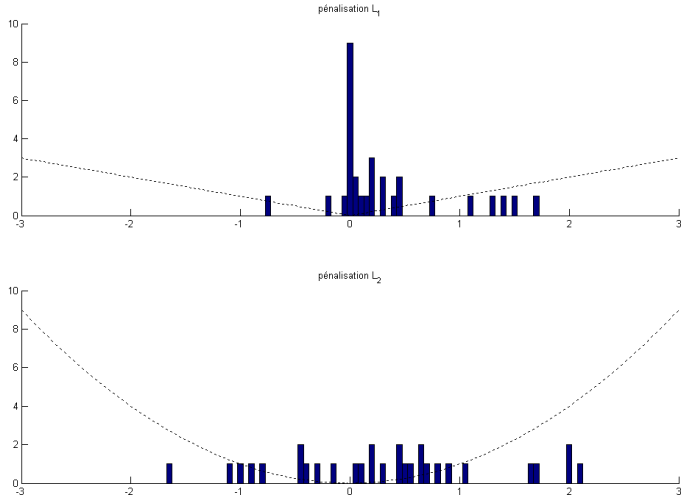

FIG. 1.4 – Histogrammes des coefficients de régression dans le cas du Lasso (en haut) et la régression Ridge (en bas).

1.2.2.2 Généralisations

Les techniques de moindres carrés pénalisés de type régression Ridge ou Lasso se généralisent en considérant des critères à minimiser en $\boldsymbol{\beta}$ de la forme :

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|^q$$

pour $q \geq 0$. De telles procédures sont appelées régression *Bridge* par Frank et Friedman ([15]). La valeur $q = 0$ correspond à la sélection discrète du sous-ensemble de variables explicatives. $q = 1$ correspond au problème du Lasso et $q = 2$ correspond à la régression Ridge. On note que le cas $q = 1$ (Lasso) est la plus petite valeur de q pour laquelle la zone de contrainte est convexe. Les zones de contrainte non convexes introduisent de sérieuses difficultés algorithmiques dans l'optimisation du critère.

Dans la suite, on note la fonction de pénalisation par $J(\boldsymbol{\beta})$. On choisit de considérer une pénalisation L_1 car les modèles que l'on cherche à estimer dans les cas pratiques sont parcimonieux.

$$J(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_1 \tag{1.6}$$

C'est de cette manière que l'on définit par la suite la taille d'un modèle.

Les techniques de maximum de vraisemblance pénalisé L_1 sont appelées procédures de type Lasso. On s'intéresse dans la suite au cas particulier de la régression logistique dans sa version avec pénalisation L_1.

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ -\log L_n^{logit}(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

La valeur de λ règle le compromis entre la vraisemblance et la taille du modèle. Le choix de λ, qui est un problème non résolu, fixe la taille du modèle estimé.

Pour aider au choix du λ optimal, on se propose d'utiliser la technique des chemins de régularisation.

1.2.3 Chemins de régularisation

Pour ne pas avoir à choisir la valeur de la constante de régularisation, on fait varier cette valeur sur l'ensemble des valeurs possibles et on étudie l'ensemble des solutions obtenues.

Au lieu de chercher d'abord à déterminer la valeur du meilleur compromis λ, puis de chercher $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{-\log L_n(\boldsymbol{\beta}) + \lambda J(\boldsymbol{\beta})\}$, on calcule directement l'ensemble :

$$\left\{ \widehat{\boldsymbol{\beta}}(\lambda), \lambda \in [0, +\infty[\text{ où } \widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \{-\log L_n(\boldsymbol{\beta}) + \lambda J(\boldsymbol{\beta})\} \right\}$$

De l'étude de cet ensemble de solution (que l'on nomme *chemin de régularisation* ou encore *chemin de solutions*), on espère pouvoir trouver quelle est la valeur optimale à donner au compromis λ.

Dans le cas du Lasso, Efron et al. ont proposé un algorithme de calcul du chemin de régularisation (voir [14] pour les détails). Cet algorithme se déroule de la façon suivante. On recherche la variable explicative la plus corrélée avec le résidu courant, qui est à la première étape confondu avec la variable à expliquer. On ne fait pas entrer cette variable explicative dans le modèle par pas de ε, comme c'est le cas dans la régression « Stagewise », mais directement avec le coefficient au-delà duquel cette variable n'est plus la seule à être la plus corrélée avec le résidu. On fait alors entrer dans le modèle cette deuxième variable explicative. De même que pour la première variable explicative, on fait entrer la deuxième variable directement avec le coefficient au-delà duquel ces deux variables ne sont plus les deux seules variables explicatives les plus corrélées avec le résidu courant. On procède

ainsi jusqu'à ce que le modèle contienne toutes les variables explicatives. Les variables qui sont dans le modèle se trouvent toutes corrélées de la même façon avec le résidu. Efron et al. parlent d'*équicorrélation*. Cet algorithme est appelé algorithme LARS, pour *Least Angle Regression*, régression de moindre angle, du fait des propriétés de l'équicorrélation. Une propriété intéressante de cet algorithme est que le calcul de la totalité du chemin de régularisation est de même complexité que le calcul de l'estimateur pour un λ donné ([14],[35]). Cela tient au fait que, dans le cas de la régression linéaire, chaque étape peut être résolue analytiquement.

Les chemins de régularisation dans le cas de la régression ont d'autres propriétés remarquables, notamment celle d'être affines par morceaux. Cela a pour conséquence que, pour connaître l'intégralité du chemin, il suffit de connaître les valeurs des $\widehat{\boldsymbol{\beta}}(\lambda)$ en un nombre fini de λ (si p est fini). Rosset et Zhu ont montré dans [40] que cette propriété se conserve sous certaines conditions sur les fonctions de coût et de pénalisation. Leur résultat s'énonce sous les hypothèses que L est une fonction de coût convexe et non-négative de $\mathbb{R}^n$ dans $\mathbb{R}^n$. J est une fonction de pénalisation supposée convexe et également non-négative de $\mathbb{R}^n$ dans $\mathbb{R}$ avec la propriété que $J(0) = 0$. Alors le chemin de régularisation :

$$\left\{ \widehat{\boldsymbol{\beta}}(\lambda), \lambda \in [0, +\infty[\text{ où } \widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{ L(\boldsymbol{Y}, \boldsymbol{X}^T\boldsymbol{\beta}) + \lambda J(\boldsymbol{\beta}) \right\} \right\}$$

est affine par morceaux si :

- L est quadratique par morceaux comme fonction de $\boldsymbol{\beta}$ et
- J est affine par morceau comme fonction de $\boldsymbol{\beta}$.

Le cas de la régression logistique, qui nous intéresse dans la suite du document, ne vérifie pas ces propriétés. La condition énoncée par Rosset et Zhu est nécessaire mais pas suffisante. Ainsi on ne peut rien conclure quant à la linéarité par morceaux du chemin de régularisation résultant. Dans la pratique, on vérifie qu'il n'est pas affine par morceaux.

Sans cette propriété, on ne peut pas espérer connaître le chemin de régularisation en intégralité avec seulement quelques points. Dans le cadre plus général des modèles linéaires généralisés, qui regroupent notamment la régression linéaire et la régression logistique, Park et Hastie proposent un algorithme de suivi du chemin de régularisation [36]. Cet algorithme est initialisé en partant du vecteur de régresseur nul (qui correspond à un λ grand). L'auteur procède par étapes successives :

- décrémentation de λ d'une certaine valeur ;
- étape de prédiction : approximation linéaire du nouveau $\widetilde{\boldsymbol{\beta}}$;
- étape de correction : optimisation ponctuelle de $\widehat{\boldsymbol{\beta}}$ en initialisant l'optimisation à $\widetilde{\boldsymbol{\beta}}$.

L'algorithme est d'autant plus performant que l'on sait décrémenter λ d'une valeur bien choisie. Des stratégies intuitives et flexibles sont proposées dans [36] pour

déterminer la décrémentation appropriée. Il est également dit que le choix du pas de la décrémentation est critique en ce qui concerne le contrôle de la précision du chemin. Cette approche ne dispense pas de calculs d'optimisation qui peuvent s'avérer nombreux.

L'algorithme que nous proposons dans le chapitre 2 ne cherche pas à calculer l'intégralité du chemin de régularisation. En effet, seuls nous intéresse les λ intéressants, c'est-à-dire ceux pour lesquels l'ensemble des variables actives change. On appelle *variable active* une variable explicative dont la composante $\widehat{\beta}_j$ du vecteur de paramètres du modèle est non nulle. On dit également qu'une variable active est *dans* le modèle quand son coefficient est non nul. Par analogie, on dira qu'une variable *entre* dans le modèle quand son coefficient devient non nul.
Dans notre approche, nous sommes également amenés à réaliser des calculs d'optimisation. Cependant, leur nombre est limité, comme on le verra dans le chapitre 2, par une recherche par dichotomie des λ intéressants.

1.3 Résultats théoriques autour des modèles parcimonieux et du Lasso

Depuis les travaux de Tibshirani sur le Lasso [46], l'étude des chemins de régularisation suscite un vif intérêt dans la communauté statistique. Parmi ces nombreuses recherches, nous nous intéresserons particulièrement dans le chapitre 3 à l'étude du comportement asymptotique des chemins de régularisation.

On se propose, dans cette introduction, de situer nos travaux par rapports aux travaux existants menés autour de cette problématique. Pour ce tour d'horizon des travaux récents, on ne rentrera pas dans les détails, mais on citera l'essentiel des résultats. On cite tout d'abord les travaux de Knight et Fu, travaux qui ont motivé nos recherches. Dans [29], il est montré que la procédure du Lasso est consistente en sélection et qu'il existe un théorème central limite. Nos travaux étendent ces résultats, c'est pourquoi l'on ne rappellera pas, dans cette introduction, le détail de leurs théorèmes. On trouvera notamment, dans la partie 3.1, l'application de nos résultats à ce cas particulier du Lasso. Nous renforçons les résultats de Knight et Fu aussi bien concernant la consistence : on montre que celle-ci est uniforme en λ, où λ est la constante de régularisation, que concernant le théorème de la limite centrale : on obtient une version fonctionnelle du TLC, c'est-à-dire pour lequel l'estimateur est vu comme une fonction de la constante de régularisation.

On verra, dans le paragraphe 1.3.1, que la définition de la consistence en sélection peut être élargie comparativement à la consistence telle que l'étudient Knight et Fu. C'est cette consistence en signe que Zhao et Yu obtiennent dans [52] pour la procédure du Lasso. Greenshtein et Ritov ont également travaillé sur la consistence

de la procédure du Lasso. Dans [20], ils se sont particulièrement intéressés à la consistence en risque, notion qu'ils nomment « persistence ». Le paragraphe 1.3.2 de cette introduction revient sur ces résultats de persistence. On cite également les travaux de Candès et Tao ainsi que ceux de Bunea, Tsybakov et Wegkamp. Dans [9], Candès et Tao introduisent un estimateur appelé « Dantzig selector ». Cet estimateur minimise la norme L_1 des paramètres du modèle à laquelle ils ajoutent une pénalisation. Il s'agit d'une nouvelle manière de poser le problème. Dans le paragraphe 1.3.4, on rappelle que pour ce nouvel estimateur, si le vrai modèle est parcimonieux alors la distance euclidienne entre leur estimateur et les vrais paramètres est bornée presque sûrement, avec une grande probabilité. Ce sont des techniques de majorations d'erreurs similaires qui sont utilisées dans les travaux de Bunea et al. En effet, dans [7], les auteurs montrent que sous certaines hypothèses sur la parcimonie du modèle, l'estimateur qu'ils proposent (et qui généralise le cas du Lasso) obéit, avec une grande probabilité, à des majorations du risque quadratique. Ils proposent également un contrôle avec une grande probabilité de la distance L_1 entre leur estimateur et les vrais paramètres du modèle. On expose, dans le paragraphe 1.3.3, ces principaux résultats.

1.3.1 Consistence en estimation, consistence en sélection

Le choix de la constante de régularisation est un problème difficile et encore ouvert. Leng et al. [30] montrent qu'à p fixé et quand la matrice des variables explicatives est orthogonale, lorsque l'on pilote le choix de la constante de régularisation par la recherche de l'estimateur donnant la plus grande prédictibilité, alors cette procédure d'estimation n'est pas consistente en sélection. De plus, les travaux de Osborne et al. [34] sur l'utilisation du Lasso pour la détection de nœuds de splines en régression, ont montré que, plus généralement, en présence d'une variable parasite très fortement corrélée avec une variable explicative du vrai modèle, le Lasso peut ne pas être capable de la distinguer des vraies variables explicatives, et ce quelle que soit la quantité de données disponibles et quelle que soit la valeur de la régularisation.

Zhao et Yu établissent, dans [52], une condition (qu'ils nomment *Irrepresentable Condition*, et que l'on traduit par *condition d'irreprésentabilité*) visant à établir un résultat de consistence en sélection. La condition d'irreprésentabilité dépend essentiellement de la covariance des variables explicatives. Cette condition s'avère nécessaire et presque sûrement suffisante. En d'autres termes, le Lasso est consistent en signe (qui implique la consistence en sélection) si et *presque* seulement si les variables explicatives qui ne sont pas dans le vrai modèle sont « irreprésentables » par les variables explicatives qui sont dans le vrai modèle.

Si les n observations des p les variables explicatives sont notées matricielle-

ment par $\boldsymbol{X_n} \in \mathcal{M}_{n,p}$ et les n observations de la variable à expliquer sont notées dans le vecteur Y_n, alors l'estimateur $\widehat{\boldsymbol{\beta}}^n = \left(\widehat{\beta}_1^n, ..., \widehat{\beta}_p^n\right)^T$ du Lasso est défini par :

$$\widehat{\boldsymbol{\beta}}^n(\lambda) = \arg\min_{\boldsymbol{\beta}} ||Y_n - \boldsymbol{X_n}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1 \tag{1.7}$$

On rappelle tout d'abord la différence entre la consistence en estimation et la consistence en sélection. La consistence en estimation impose :

$$\widehat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}^n \rightarrow_p 0, \text{ quand } n \rightarrow \infty$$

tandis que la consistence en sélection impose :

$$\mathbb{P}(\{i : \widehat{\boldsymbol{\beta}}_i^n \neq 0\} = \{i : \boldsymbol{\beta}_i^n \neq 0\}) \rightarrow 1, \text{ quand } n \rightarrow \infty$$

Pour pouvoir prouver le caractère nécessaire de la condition d'irreprésentabilité, Zhao et Yu définissent la *consistence en signe*.

On dit que $\widehat{\beta}^n$ est de même signe que β^n si

$$\text{sign}(\widehat{\boldsymbol{\beta}}^n) = \text{sign}(\boldsymbol{\beta}^n)$$

où $\text{sign}(.)$ est défini comme suit :

$$\text{sign}(x) = \begin{cases} 1 & \text{si } x > 0 \\ 0 & \text{si } x = 0 \\ -1 & \text{si } x < 0 \end{cases}$$

La consistence en sélection, qui n'impose la correspondance que pour les zéros, et non pas pour les signes, est donc plus faible que la consistence en signe.

Définition 1.1 (Forte consistence en signe du Lasso)
Le Lasso est fortement consistent en signe s'il existe λ_n, fonction de n et indépendant de $\boldsymbol{X_n}$ ou Y_n, tel que :

$$\lim_{n\rightarrow\infty} \mathbb{P}\left(\text{sign}(\widehat{\boldsymbol{\beta}}^n(\lambda_n)) = \text{sign}(\boldsymbol{\beta}^n)\right) = 1$$

Définition 1.2 (Consistence générale en signe du Lasso)
Le Lasso est généralement consistent en signe si :

$$\lim_{n\rightarrow\infty} \mathbb{P}\left(\exists\lambda \geq 0, \text{sign}(\widehat{\boldsymbol{\beta}}^n(\lambda)) = \text{sign}(\boldsymbol{\beta}^n)\right) = 1$$

La consistence forte en signe signifie que l'on peut choisir au préalable la valeur de λ pour laquelle on a la consistence en sélection pour le Lasso. Tandis que la consitence générale signifie juste que, pour une réalisation, il existe un λ qui conduit à sélectionner le vrai modèle. La consistence forte implique naturellement la consistence générale.

Les deux types de consistences en signe sont quasiment équivalentes sous la condition d'irreprésentabilité.

On suppose, sans perte de généralité que les q premières composantes de $\boldsymbol{\beta}^n$ sont les composantes non nulles du modèles parciminieux. Les $p-q+1$ dernières composantes sont donc supposées nulles. On pose $\boldsymbol{\beta}^n_{(1)} = \left(\boldsymbol{\beta}^n_1, ..., \boldsymbol{\beta}^n_q\right)^T$ et $\boldsymbol{\beta}^n_{(2)} = \left(\boldsymbol{\beta}^n_{q+1}, ..., \boldsymbol{\beta}^n_p\right)^T$. De même, on sépare les q premières colonnes de $\boldsymbol{X_n}$, que l'on note $\boldsymbol{X_n}(1)$ et les $p-q+1$ dernières, que l'on note $\boldsymbol{X_n}(2)$.

On note maintenant $C^n = \frac{1}{n}{\boldsymbol{X_n}}^T\boldsymbol{X_n}$. Cette matrice peut également s'écrire en bloc, de la manière suivante :

$$C^n = \begin{pmatrix} C^n_{11} & C^n_{12} \\ C^n_{21} & C^n_{22} \end{pmatrix}$$

où $C^n_{ij} = \frac{1}{n}\boldsymbol{X_n}(i)^T\boldsymbol{X_n}(j)$.

On suppose que C^n_{11} est inversible et que l'on a une des deux conditions suivantes :

Condition d'irreprésentabilité forte

Il existe un vecteur η constant positif tel que :

$$|C^n_{21}(C^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^n_{(1)})| \leq \mathbf{1} - \eta$$

où $\mathbf{1}$ est un vecteur de 1 de dimension $p-q$. L'inégalité doit être comprise élément par élément.

Condition d'irreprésentabilité faible

$$|C^n_{21}(C^n_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}^n_{(1)})| < \mathbf{1}$$

L'inégalité doit être comprise élément par élément.

On fait les deux hypothèses de régularité suivantes sur la matrice $\boldsymbol{X_n}$.
Premièrement :

$$C^n \rightarrow_{n\rightarrow\infty} C \tag{1.8}$$

où C est une matrice définie positive. Et secondement :

$$\frac{1}{n}\max_{1\leq\leq n}\left(\boldsymbol{x_i^n}\right)^T\boldsymbol{x_i^n} \rightarrow_{n\rightarrow\infty} 0 \tag{1.9}$$

où $\boldsymbol{x_i^n}$ est la $i^{\text{ème}}$ ligne de la matrice $\boldsymbol{X_n}$.

Théorème 1.1 (Zhao et Yu [52]) *Pour p et q fixés, pour $\boldsymbol{\beta}^n = \boldsymbol{\beta}$, sous les conditions (1.8) et (1.9), le Lasso est fortement consistent en signe si la condition d'irreprésentabilité forte est vérifiée.*

Si la condition d'irreprésentabilité est vérifiée, alors pour toute suite λ_n qui vérifie $\lambda_n/n \to 0$ et $\lambda_n/n^{\frac{1+c}{2}} \to \infty$ avec $0 \leq c < 1$, on a :

$$\mathbb{P}\left(\operatorname{sign}(\widehat{\boldsymbol{\beta}}^n(\lambda_n)) = \operatorname{sign}(\boldsymbol{\beta}^n)\right) = 1 - o\left(e^{-n^c}\right)$$

En plus de ce théorème, on sait par Knight et Fu (voir [29] et partie 3.1 de ce document), que pour $\lambda_n = o(n)$, le Lasso est également consistent en estimation et qu'il existe un théorème de la limite centrale. Ainsi la condition de consistence forte en signe permet d'avoir simultanément la consistence en sélection et en estimation du Lasso.

Théorème 1.2 (Zhao et Yu [52]) *Pour p et q fixés, pour $\boldsymbol{\beta}^n = \boldsymbol{\beta}$, sous les conditions (1.8) et (1.9), le Lasso possède la propriété de consistence générale en signe seulement si il existe N tel que la condition d'irreprésentabilité faible est vérifiée pour $n > N$.*

Ainsi la condition d'irreprésentabilité forte implique la consistence forte en signe qui implique elle-même la consistence générale en signe qui, elle, implique finalement la condition d'irreprésentabilité faible.

On note que la différence entre la condition d'irreprésentabilité faible et forte disparaît dans le cas où p et $\boldsymbol{\beta}^n$ sont fixés ($\boldsymbol{\beta}^n = \boldsymbol{\beta}$) et où les $\boldsymbol{x_i^n}$ sont, par exemple, des observations i.i.d. de matrice de covariance C. Alors les deux conditions sont équivalentes à $|C_{21}(C_{11})^{-1}\operatorname{sign}(\boldsymbol{\beta}_{(1)})| \leq \mathbf{1}$, presque sûrement.

On peut en déduire que la condition d'irreprésentabilité est presque nécessaire et suffisante à la fois pour la consistence forte et pour la consistence générale en signe.

De manière similaire, sous des hypothèses de régularité additionnelles sur le terme de bruit, en supposant que p tend également vers l'infini mais « pas trop vite », il est démontré que là encore, la condition d'irreprésentabilité forte implique la consistence forte en signe pour le Lasso.

La condition définie par Zhao et Yu est proche de la notion de « cohérence mutuelle » définie par Donoho et al. dans [12]. Dans l'article de Donoho et al., la cohérence mutuelle est définie de la manière suivante :

Définition 1.3 (Cohérence mutuelle) *La matrice des variables explicatives est ici notée $\boldsymbol{\Phi}$. Supposons que les colonnes de $\boldsymbol{\Phi}$ sont normalisées unitairement pour la*

norme L_2. La matrice de Gram $\boldsymbol{G}$ s'écrit $\boldsymbol{G} = \boldsymbol{\Phi}'\boldsymbol{\Phi}$. Si $G(i,j)$ est le terme général de cette matrice, alors la cohérence mutuelle est :

$$M = M(\boldsymbol{\Phi}) = \max_{1 \leq i,j \leq p, i \neq j} |G(i,j)|$$

Donoho et al. cherchent des conditions sous lesquelles il est possible de reconstruire exactement des modèles parcimonieux. Pour cela, ils invoquent une propriété de cohérence mutuelle quand celle-ci est suffisamment petite.

1.3.2 Persistence du Lasso en grande dimension

Le contexte de ce paragraphe est celui considéré par Greenshtein et Ritov [20], à savoir que l'on considère, pour $i = 1, ..., n$, le vecteur aléatoire $Z^i = (Y^i, X_1^i, ..., X_p^i)$. $\mathcal{F}^n$ est l'ensemble des distributions de vecteurs i.i.d. $Z^i, i = 1, ..., n$ de dimensions $p + 1$. On suppose qu'il y a beaucoup plus de variables explicatives que d'observations : $p = n^\alpha$, $\alpha > 1$.

On souhaite prédire Y par $\sum \beta_j X_j$ où $(\beta_1, ..., \beta_p) \in B_n \subseteq \mathbb{R}^p$. Pour le cas particulier où Z^i suit une distribution F multi-normale, alors on estime $\boldsymbol{Y} = \sum \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$ où $\boldsymbol{\alpha}$ est le meilleur prédicteur linéaire au sens du risque.

On se donne la possibilité de contraindre les ensembles B^n soit par un nombre maximal de composantes non nulles, soit par une boule L_1, c'est-à-dire que l'on considère des sous-ensembles $(\beta_1, ..., \beta_p) \subseteq \mathbb{R}^p$ vérifiant soit la contrainte $||\boldsymbol{\beta}||_0 = \sum_{j=1}^p \mathbb{I}(\beta_j \neq 0) < k$, $k \in \mathbb{N}$, soit la contrainte $||\boldsymbol{\beta}||_1 = \sum_{j=1}^p |\beta_j| < b$, $b \in \mathbb{R}^+$.

La question est la suivante : jusqu'où peut-on contraindre l'ensemble B^n sachant que l'on veut pouvoir encore sélectionner empiriquement un modèle dont le risque est proche de celui du meilleur modèle de l'ensemble ? Sous certaines hypothèses sur F, la loi de Z^i, Greenshtein et Ritov donnent dans [20] des bornes ajustées pour des contraintes sur k et sur b.

Greenshtein et Ritov montrent que si le modèle optimal est d'une certaine manière parcimonieux, alors il n'est asymptotiquement pas dommageable d'introduire beaucoup plus de variables explicatives que d'observations.

Un autre résultat important est que les procédures de type Lasso, *i.e.* les procédures d'optimisation sous une contrainte L_1, semblent être efficaces pour la sélection de modèle en grande dimension, toujours sous des hypothèses de parcimonie.

On note par :

$$L_F(\beta) = \mathbb{E}_F \left[Y - \sum_{i=1}^{p} \beta_j X_j \right]^2$$

le risque quadratique, sous la loi F de Z^i. Pour un ensemble de modèles B^n donné, ainsi qu'une distribution F_n donnée, on définit ainsi le modèle optimal par :

$$\beta^*_{F_n} = \arg \min_{\beta \in B^n} L_{F_n}(\beta)$$

On note de manière identique la procédure et le prédicteur obtenu par cette procédure, à savoir $\widehat{\beta}^n$.

Définition 1.4 (Persistence) *Étant donné une suite d'ensembles de modèles B^n, la suite des procédures d'estimation $\widehat{\beta}^n$ est dite persistente si pour tout suite $F_n \in \mathcal{F}^n$:*

$$L_{F_n}(\widehat{\beta}^n) - L_{F_n}(\beta^*_{F_n}) \to_p 0$$

On note que la persistence peut être vue comme une *consistence en risque*, ce qui est approprié dans l'optique de sélectionner le modèle de plus faible erreur de prédiction.

Considérons maintenant que l'on a n observations $(Z^1, ..., Z^n)$. On appelle $\widehat{F}$ leur distribution empirique et on définit le risque empirique par :

$$L_{\widehat{F}}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left(Y^i - \sum_{i=1}^{p} \beta_j X^i_j \right)^2$$

On considère les procédures de sélection de modèles suivantes : pour un $b \in \mathbb{R}$ donné, on a :

$$\widehat{\beta}^n = \arg \min_{\{\beta : ||\beta||_1 \leq b(n)\}} L_{\widehat{F}}(\beta)$$

où b dépend en réalité de n.

Sont alors considérés les deux types d'ensembles de modèles suivants :

1. B^n_k est l'ensemble des vecteurs $(\beta_1, \ldots, \beta_p)_{\beta_j \in \mathbb{R}, \forall j \in \{1,\ldots,p\}}$ ayant au plus $k = k(n)$ composantes non nulles. Cette première manière de contraindre les ensembles B^n peut être utilisée pour des problèmes de sélection de variables. En effet, cela revient à choisir k variables explicatives parmi cet ensemble de p variables possibles.
2. B^n_b est l'ensemble des vecteurs $(\beta_1, \ldots, \beta_p)_{\beta_j \in \mathbb{R}, \forall j \in \{1,\ldots,p\}}$ ayant une norme L_1 plus petite ou égale à $b = b(n)$. Cette seconde manière de contraindre les ensembles B^n est en lien avec les procédures de type Lasso.

Greenshtein et Ritov montrent dans [20] que, dans le cas où Z^i suit une loi normale multidimensionnelle, $b(n)$ est de l'ordre de $o((n/\log(n))^{1/2})$. Ils montrent également, toujours dans le cas d'une loi normale multidimensionnelle, que les procédures d'estimation sont persistentes pour $k(n) = o((n/\log(n))$.

Ces résultats donnent des indications sur la qualité des prédicteurs en terme de persistence. Il est difficile d'en tirer des préconisations pour l'interprétation dans la pratique des chemins de régularisation. En effet, dans le cas où $b(n)$ est grand, alors il y a de fortes chances que le meilleur prédicteur non-contraint soit dans $B^n_{b(n)}$. Et la persistence est alors triviale.
Dans le cas où $b(n)$ est petit, il se peut que l'on perde la propriété de persistence. Mais il faut noter que l'erreur, dans la définition de la persistence, est calculée entre la procédure et le meilleur prédicteur *contraint*. Et il se peut que dans le même temps l'erreur entre la procédure et le meilleur prédicteur non-contraint soit, elle, très grande (de même que l'erreur entre le prédicteur contraint et le prédicteur non-contraint peut être considérable). Dans ce cas, il n'est peut-être pas judicieux de vouloir imposer que la procédure soit persistente.

1.3.3 Inégalités d'oracle

On considère un ensemble de paires aléatoires $(X_1, Y_1), ..., (X_n, Y_n)$ telles que $(\boldsymbol{X}, \boldsymbol{Y}) \in (\mathcal{X}, \mathbb{R})$. Classiquement, le modèle de régression est le suivant :

$$Y_i = \beta_1 X_{1i} + ... + \beta_p X_{pi} + \varepsilon_i, i \in \{1, ..., n\}$$

et on rappelle que la procédure du Lasso consiste à trouver l'estimateur suivant :

$$\widehat{\boldsymbol{\beta}}_{Lasso} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\beta_1 X_{1i} + ... + \beta_p X_{pi})^2 + \lambda |\boldsymbol{\beta}|_1 \right\} \tag{1.10}$$

Considérons le modèle de régression non-paramétrique suivant : $\boldsymbol{Y} = f(\boldsymbol{X}) + \boldsymbol{\varepsilon}$. On note donc $f(\boldsymbol{X}) = \mathbb{E}(\boldsymbol{Y}|\boldsymbol{X})$ la fonction de régression, qui est inconnue, et $\mathcal{F}_p$ une famille finie de fonctions, appelée encore dictionnaire fini de fonctions à valeurs réelles f_j définies sur $\mathcal{X}$. On note μ la mesure de probabilité de $\boldsymbol{X}$.

En fonction de ce que l'on veut étudier, le dictionnaire $\mathcal{F}_p$ peut être de différente nature. $\mathcal{F}_p$ peut être :

1. une famille de fonctions de base utilisées pour approximer f dans un modèle de régression non-paramétrique ;
2. un vecteur p-dimensionnel de variables aléatoires $(f_1(\boldsymbol{X}), ..., f_p(\boldsymbol{X}))$, comme c'est le cas en régression linéaire ;
3. une famille de p estimateurs de f donnés.

On s'intéresse en particulier ici au troisième cas. Il s'agit d'un problème d'agrégation. Dans ce contexte, il est classique de rendre la fonction de pénalisation dépendante des données : $\mathrm{pen}(\boldsymbol{\beta})$. Le paramètre λ de régularisation dans la pénalisation utilisée dans (1.10) varie avec $j \in \{1, ..., p\}$. Pour tout $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ élément de $\mathbb{R}^p$, on définit alors $f_{\boldsymbol{\beta}}(x)$ comme $\sum_{j=1}^{p} \beta_j f_j(x)$.

L'estimateur des moindres carrés pénalisés de $\boldsymbol{\beta}$ devient :

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^{n}\{Y_i - f_{\boldsymbol{\beta}}(X_i)\}^2 + \mathrm{pen}(\boldsymbol{\beta}) \right\} \tag{1.11}$$

où Bunea et al. proposent, dans [7], de prendre :

$$\mathrm{pen}(\boldsymbol{\beta}) = 2\sum_{j=1}^{p} r_{n,p}||f_j||_n \tag{1.12}$$

avec $||f_j||_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n} f_j^2(X_i)}$, la norme L_2 empirique de la fonction f_j. L'estimateur correspondant est :

$$\widehat{f} = \sum_{j=1}^{p} \widehat{\beta}_j f_j$$

On prend les notations de Bunea et al. ([7]) en ce qui concerne la parcimonie du modèle $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}) = \sum_{j=1}^{p} \mathbb{I}_{\{\beta_j\neq 0\}} = \mathrm{Card}(J(\boldsymbol{\beta}))$$

où $\mathbb{I}$ est la fonction indicatrice et $J(\boldsymbol{\beta}) = \{j \in \{1, ..., p\} : \beta_j \neq 0\}$.
Ainsi $p(\boldsymbol{\beta})$ mesure la parcimonie du modèle : plus $p(\boldsymbol{\beta})$ est petit, plus le modèle est parcimonieux.

Prenons le cas simple de la régression linéaire : $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}) = f(\boldsymbol{X}) = \boldsymbol{X}^T\boldsymbol{\beta_0}$, où $\boldsymbol{\beta_0}$ a des coefficients non-nuls uniquement pour $j \in J(\boldsymbol{\beta_0})$. Bunea et al. montrent que, pour $\widehat{\boldsymbol{\beta}}$ donné par (1.11) et à la condition de prendre $r_{n,p} = A\sqrt{\log(p)/n}$ dans le terme de pénalisation de (1.12), alors $|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$ est borné par $p(\boldsymbol{\beta_0})/\sqrt{n}$, à des constantes et des logarithmes connus près. Cela signifie que l'estimateur $\widehat{\boldsymbol{\beta}}$ du vecteur de coefficients $\boldsymbol{\beta_0}$ s'adapte à la parcimonie du problème : l'erreur d'approximation diminue quand la parcimonie du vecteur $\boldsymbol{\beta_0}$ augmente.

Comme il n'est généralement pas possible de représenter exactement f par une combinaison linéaire d'éléments donnés f_j, on suppose que l'on peut contrôler le carré de la distance de f à $f_{\boldsymbol{\beta^*}}$ - pour un $\boldsymbol{\beta^*} \in \mathbb{R}^p$ donné - par $p(\boldsymbol{\beta^*})/n$, à des facteurs logarithmiques près. Les auteurs nomment ce cas *parcimonie faible*.

Soit C_f une constante ne dépendant que de la fonction f. On définit l'ensemble oracle de la manière suivante :

$$B = \{\boldsymbol{\beta} \in \mathbb{R}^p : ||f_{\boldsymbol{\beta}} - f||^2 \leq C_f r_{n,p}^2 p(\boldsymbol{\beta})\}$$

La notation $||.||^2$ est la norme dans l'ensemble des fonctions de carré intégrable pour la mesure μ :

$$||g||^2 = \int_{\mathcal{X}} g^2(x)\mu(dx)$$

Si B est non vide, alors on dit que f possède la *propriété de parcimonie faible, relativement au dictionnaire* $\{f_1, ..., f_p\}$.

On obtient donc, sous l'hypothèse de parcimonie faible :

$$\boldsymbol{\beta^*} = \arg\min\{||f_{\boldsymbol{\beta}} - f||^2 : \boldsymbol{\beta} \in \mathbb{R}^p, p(\boldsymbol{\beta}) = k^*\}$$

où $k^* = \min_{\boldsymbol{\beta} \in B} p(\boldsymbol{\beta})$.

Toutes les quantités $\boldsymbol{\beta^*}$, k^* et $f_{\boldsymbol{\beta^*}}$ peuvent être considérées comme des oracles.

Afin d'énoncer le principal résultat de [7], on fait un certain nombre d'hypothèses. La première hypothèse concerne le terme d'erreur : $\varepsilon_i = Y_i - f(X_i)$. On rappelle que $f(\boldsymbol{X}) = \mathbb{E}(\boldsymbol{Y}|\boldsymbol{X})$.

Hypothèse 1.1 *Les variables aléatoires $X_1, ..., X_n$ sont indépendantes, identiquement distribuées et ont pour mesure de probabilité μ. Les variables aléatoires ε_i sont identiquement distribuées avec :*

$$\mathbb{E}[\varepsilon_i | X_1, ..., X_n] = 0$$

et :

$$\mathbb{E}[\exp(|\varepsilon_i|)|X_1, ..., X_n] \leq b$$

pour $b > 0$ donné et $i \in \{1, ..., n\}$.

On pose maintenant des conditions sur la fonction f ainsi que sur les fonctions du dictionnaires f_j. On définit la norme infinie pour une fonction g bornée sur $\mathcal{X}$ par : $||g||_\infty = \sup_{x \in \mathcal{X}} |g(x)|$.

Hypothèse 1.2

(a) *Il existe $0 < L < \infty$ tel que $||f_j||_\infty \leq L, \forall j \in \{1, ..., p\}$.*

(b) *Il existe $c_0 > 0$ tel que $||f_j|| \geq c_0, \forall j \in \{1, ..., p\}$.*

(c) *Il existe $L_* < \infty$ tel que $||f||_\infty \leq L_*$.*

Remarques :

- (a) implique qu'il existe $L_0 < \infty$ tel que $\mathbb{E}[f_i^2(\boldsymbol{X}) f_j^2 \boldsymbol{X}] \leq L_0, \forall (i, j) \in \{1, ..., p\}^2$.
- (a) et (c) impliquent que, pour tout $\boldsymbol{\beta} \in \mathbb{R}^p$, il existe une constante positive $L(\boldsymbol{\beta})$ telle que $||f - f_{\boldsymbol{\beta}}||_\infty = L(\boldsymbol{\beta})$.

On fait une dernière hypothèse sur la matrice $\boldsymbol{X}$. Les résultats dépendent fortement du comportement de la matrice carrée Ψ_p de taille p, définie par :

$$\Psi_p = \left(\mathbb{E}[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X})]\right)_{1 \leq j,j' \leq p} = \left(\int f_j(x) f_{j'}(x) \mu(dx)\right)_{1 \leq j,j' \leq p}$$

Hypothèse 1.3 *Pour tout $p \geq 2$, il existe $\kappa_p > 0$ tel que $\Psi_p - \kappa_p \operatorname{diag}(\Psi_p)$ est semi-définie positive.*

On note que $0 < \kappa_p \leq 1$. On note également que l'hypothèse (1.3) et le point (b) de l'hypothèse (1.2) impliquent que la matrice Ψ_p est définie positive, dont la plus petite valeur propre est donc minorée par $c_0 \kappa_p$.

On peut maintenant énoncer le résultat suivant, qui est valable dès que $n \geq 1$, $p \geq 2$ et $r_{n,p} > 0$.

Théorème 1.3 (Bunea et al. [7]) *Supposons que les hypothèses (1.1), (1.2) et (1.3) sont vérifiées. Alors pour tout $\boldsymbol{\beta} \in B$, on a :*

$$\mathbb{P}\left\{||\widehat{f} - f||^2 \leq B_1 \kappa_p^{-1} r_{n,p}^2\right\} \geq 1 - \pi_{n,p}(\boldsymbol{\beta})$$

et

$$\mathbb{P}\left\{||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_1 \leq B_2 \kappa_p^{-1} r_{n,p}\right\} \geq 1 - \pi_{n,p}(\boldsymbol{\beta})$$

où $B_1 > 0$ et $B_2 > 0$ sont deux constantes qui dépendent seulement de c_0 et de C_f et

$$\begin{aligned} \pi_{n,p} \quad \leq \quad & 10p^2 \exp\left(-c_1 n \min\left\{r_{n,p}^2, \frac{r_{n,p}}{L}, \frac{1}{L^2}, \frac{\kappa_p^2}{L_0 p^2(\boldsymbol{\beta})}, \frac{\kappa_p}{L^2 p(\boldsymbol{\beta})}\right\}\right) \\ & + \exp\left(-c_2 \frac{p(\boldsymbol{\beta})}{L^2(\boldsymbol{\beta})} n r_{n,p}^2\right) \end{aligned}$$

avec c_1 et c_2 deux constantes positives qui dépendent seulement de c_0, C_f et b. On a noté également $L(\boldsymbol{\beta}) = ||f - f_{\boldsymbol{\beta}}||_\infty$.

1.3.4 « Dantzig Selector »

Nous présentons dans ce paragraphe les principaux résultats développés par Candès et Tao dans leur article sur le "Dantzig selector". Ce nouvel estimateur introduit dans [10] traite du cas où le nombre de variables explicatives p est très grand devant le nombre d'observations n. On pourra trouver dans [2] une analyse comparée du Dantzig selector et du Lasso.

On se place dans le cadre de la régression linéaire. Les observations sont notées $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, où $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{X}$ est la matrice d'expériences avec potentiellement beaucoup plus de colonnes que de lignes ($n \ll p$). Les vecteurs colonnes de $\boldsymbol{X}$ sont normés unitairement. Les ϵ_i sont i.i.d. $\mathcal{N}(0, \sigma^2)$.

Candès et Tao définissent le Dantzig selector comme suit :

$$\boldsymbol{\beta}^{DS} = \arg\min_{\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^p} ||\widetilde{\boldsymbol{\beta}}||_1 \text{ t.q. } ||\boldsymbol{X'r}||_\infty = \sup_{1 \leq j \leq p} |\left(\boldsymbol{X'r}\right)_j| \leq \lambda_p \sigma \tag{1.13}$$

où $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ est le résidu et λ_p le paramètre de régularisation.
La contrainte impose que $\boldsymbol{\beta}^{DS}$ soit un modèle qui donne des résidus (corrélés) en dessous du niveau de bruit. Le fait de formuler la contrainte en termes de résidus corrélés, plutôt que directement sur les résidus seuls, entraîne que la région réalisable est invariante par transformation orthonormale des données. En effet, si l'on applique une transformation orthonormale $\boldsymbol{U}$ aux données, alors $\boldsymbol{X}$ et $\boldsymbol{y}$ devient respectivement $\boldsymbol{UX}$ et $\boldsymbol{Uy}$. On a alors :

$$||(\boldsymbol{UX})'(\boldsymbol{UX}\boldsymbol{\beta}^{DS}) - \boldsymbol{Uy}||_\infty = ||(\boldsymbol{X})'(\boldsymbol{X}\boldsymbol{\beta}^{DS}) - \boldsymbol{y}||_\infty$$

et la région des $\boldsymbol{\beta}$ vérifiant la contrainte est inchangée, ce qui n'est pas le cas si l'on formule la contrainte avec $\sup_{1 \leq j \leq p} |\boldsymbol{r}_j| \leq \lambda_p \sigma$.

1.3.4.1 Hypothèses

Le vecteur $\boldsymbol{\beta}$ est supposé suffisamment parcimonieux (ne serait-ce que pour s'assurer que le modèle est identifiable). On suppose que seules S composantes sont non nulles ($\boldsymbol{\beta}$ est S-parcimonieux). L'autre hypothèse concerne la matrice $\boldsymbol{X}$. On suppose que celle-ci obéit à un principe d'incertitude uniforme (appelé UUP pour Uniform Uncertainty Principle). Sous cette condition UUP et en l'absence de bruit, on peut retrouver exactement $\boldsymbol{\beta}$ par :

$$\arg\min_{\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^p} ||\widetilde{\boldsymbol{\beta}}||_1 \text{ t.q. } \boldsymbol{X}\widetilde{\boldsymbol{\beta}} = \boldsymbol{y} \tag{1.14}$$

Le principe UUP est basé sur une notion d'isométrie restreinte à laquelle doit obéir la matrice $\boldsymbol{X}$. On note $\boldsymbol{X}_T$ les sous-matrices de $\boldsymbol{X}$ avec $|T|$ colonnes où $T \in \mathcal{P}(\{1, ..., p\})$.

Définition 1.5 (Constante d'isométrie S-restreinte) *Soit δ_S la plus petite valeur vérifiant :*

$$(1 - \delta_S)\,||c||_2^2 \leq ||\boldsymbol{X}_T c||_2^2 \leq (1 + \delta_S)\,||c||_2^2,$$

pour tous sous-ensembles T de cardinal $|T| \leq S$ et pour tous jeux de paramètres $(c_j)_{j\in T}$. On dit alors que δ_S est la constante d'isométrie S-restreinte pour la matrice $\boldsymbol{X}$.

Cette propriété signifie que tout sous-ensemble de colonnes de $\boldsymbol{X}$ se comporte approximativement comme un système orthonormal.

Candès et Tao ont précédemment montré, dans [9], que si $\delta_S + \delta_{2S} + \delta_{3S} < 1$ alors on peut retrouver exactement $\boldsymbol{\beta}$ par (1.14), toujours pour un modèle sans bruit.

Définition 1.6 (Constante d'orthogonalité {S,S'}-restreinte) *Soient S et S' deux entiers tels que $S + S' \leq p$.*
$\theta_{S,S'}$ est définie comme étant la plus petite valeur qui vérifie :

$$| < \boldsymbol{X}_T c, \boldsymbol{X}_{T'} c' > | \leq \theta_{S,S'} ||c||_2^2 ||c'||_2^2,$$

pour tous sous-ensembles disjoints T, T' de cardinal $|T| \leq S$ et $|T'| \leq S'$ et pour tous jeux de paramètres $(c_j)_{j\in T}$ et $\left(c'_j\right)_{j\in T'}$.
On dit alors que $\theta_{S,S'}$ est la constante d'orthogonalité {S,S'}-restreinte pour la matrice $\boldsymbol{X}$.

Cette propriété signifie que des sous-ensembles disjoints de variables engendrent des sous-espaces qui sont quasiment orthogonaux.

Candès et Tao ont montré, toujours dans [9], que si $\delta_S + \theta_{S,S} + \theta_{S,2S} < 1$ alors on peut retrouver exactement $\boldsymbol{\beta}$ par (1.14). En remarquant que $\theta_{S,S'} \leq \delta_{S+S'}$ pour $S' \geq S$, on voit que cette condition est un peu moins contraignante que la condition ne faisant intervenir que des constantes d'isométrie restreinte.

On souligne que le problème du Dantzig selector est convexe et qu'il peut se récrire sous la forme du problème linéaire contraint suivant :

$$\min \sum_{j=1}^{p} u_j \text{ t.q. } -\boldsymbol{u} \leq \boldsymbol{\beta} \leq \boldsymbol{u} \text{ et } -\lambda_p \sigma \mathbf{1} \leq \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) \leq \lambda_p \sigma \mathbf{1}$$

où $\boldsymbol{u}, \widetilde{\boldsymbol{\beta}} \in \mathbb{R}^p$ sont les variables d'optimisation et $\mathbf{1}$ est un vecteur de 1 de dimension p.

1.3.4.2 Théorèmes

Théorème 1.4 *Supposons que $\boldsymbol{\beta} \in \mathbb{R}^p$ est un vecteur S-parcimonieux et que l'on a : $\delta_{2S} + \theta_{S,2S} < 1$. On prend $\lambda_p = \sqrt{2 \log p}$ dans (1.13). Alors $\boldsymbol{\beta}^{DS}$ vérifie, avec*

une grande probabilité :

$$\left\|\boldsymbol{\beta}^{DS} - \boldsymbol{\beta}\right\|_2^2 \leq C_1^2 \lambda_p^2 S\sigma^2$$

La probabilité que cette inégalité soit vérifiée est plus grande que $1-(\sqrt{\pi \log p})^{-1}$.

La valeur de la constante C_1 est $C_1 = 4/(1 - \delta_{2S} - \theta_{S,2S})$. Si l'on considère le cas plus général où $\lambda_p = \sqrt{2(1+a)\log p}$, avec $a \geq 0$, alors la probabilité que l'inégalité du théorème soit vérifiée est plus grande que $1 - (\sqrt{\pi \log p}.p^a)^{-1}$.

Comme on a supposé que $\lambda_p = 2\log p$, on a donc $||\boldsymbol{\beta}^{DS} - \boldsymbol{\beta}|| \leq C_1^2 (2\log p) S\sigma^2$. On peut noter que, sans le facteur logarithmique, une telle borne n'est pas améliorable. En effet, si l'on suppose que l'on a accès à un oracle qui nous dit quel est le vrai modèle : $T_0 = j : \beta_j \neq 0$, alors la projection des moindres carrés nous donne tout de suite :

$$\boldsymbol{\beta}^*_{T0} = \left(\boldsymbol{X}_{T_0}^T \boldsymbol{X}_{T_0}\right)^{-1} \boldsymbol{X}_{T_0}^T \boldsymbol{y}$$

où $\boldsymbol{\beta}^*_{T0}$ est la restriction de $\boldsymbol{\beta}^*$ au sous-ensemble T_0. $\boldsymbol{\beta}^*$ vaut 0 en dehors de T_0. On a alors :

$$\boldsymbol{\beta}^* = \boldsymbol{\beta} + \left(\boldsymbol{X}_{T_0}^T \boldsymbol{X}_{T_0}\right)^{-1} \boldsymbol{X}_{T_0}^T \boldsymbol{\epsilon}$$

et donc

$$\mathbb{E}||\boldsymbol{\beta}^* - \boldsymbol{\beta}||_2^2 = \mathbb{E}||\left(\boldsymbol{X}_{T_0}^T \boldsymbol{X}_{T_0}\right)^{-1} \boldsymbol{X}_{T_0}^T \boldsymbol{\epsilon}||_2^2 - \sigma^2 Tr\left(\left(\boldsymbol{X}_{T_0}^T \boldsymbol{X}_{T_0}\right)^{-1}\right)$$

Comme les valeurs propres de $\boldsymbol{X}_{T_0}^T \boldsymbol{X}_{T_0}$ appartiennent à $[1-\delta_S, 1+\delta_S]$, on a :

$$\mathbb{E}||\boldsymbol{\beta}^* - \boldsymbol{\beta}||_2^2 \geq \frac{1}{1+\delta_S} S\sigma^2$$

Le facteur logarithmique est le prix à payer savoir à l'avance quelles composantes de $\boldsymbol{\beta}$ sont nulles.

Il s'avère que le théorème 1.4 peut sembler un peu simple. En effet, si $|\beta_j| \ll \sigma, \forall j$ alors $\boldsymbol{\beta}^{DS} = \mathbf{0}$ et $||\boldsymbol{\beta}^{DS} - \boldsymbol{\beta}||_2^2 = \sum_{j=1}^p |\beta_j|^2$, quantité qui est de toutes façons plus petite que $S\sigma^2$.

D'où le second théorème.

Théorème 1.5 *Soit $t > 0$ un réel. Supposons que $\boldsymbol{\beta} \in \mathbb{R}^p$ est S-parcimonieux et que l'on a $\delta_{2S} + \theta_{S,2S} < 1 - t$. On prend $\lambda_p = (1+t^{-1})\sqrt{2\log p}$ dans (1.13). Alors $\boldsymbol{\beta}^{DS}$ vérifie, avec une grande probabilité :*

$$\left\|\boldsymbol{\beta}^{DS} - \boldsymbol{\beta}\right\|_2^2 \leq C_2^2 \lambda_p^2 \left(\sigma^2 + \sum_{j=1}^p \min(\beta_j^2, \sigma^2)\right)$$

La probabilité que cette inégalité soit vérifiée est plus grande que $1-(\sqrt{\pi \log p})^{-1}$.

avec C_2 qui ne dépend que de $\delta_{2S} + \theta_{S,2S}$. Pour information, $C_2 \leq 16$. En généralisant à $\lambda_p = (\sqrt{1+a} + t^{-1})\sqrt{2\log p}$, la probabilité que l'inégalité du théorême soit vérifiée est plus grande que $1 - (\sqrt{\pi \log p}.p^a)^{-1}$.

La condition $\delta_{2S} + \theta_{S,2S} < 1 - t$ impose une borne inférieure sur les valeurs propres des sous-matrices et prévient les situations de multicolinéarité possibles entre les sous-modèles en compétition. En effet, soit $\boldsymbol{X}_{T \cup T'}$ une matrice déficiente de $2S$ colonnes avec $|T| = |T'| = S$. Sa plus petite valeur propre est $0 = 1 - \delta_{2S}$. Alors

$$\exists \boldsymbol{h} : \boldsymbol{X}\boldsymbol{h} = 0, \text{ avec } \boldsymbol{h} = \boldsymbol{\beta} - \boldsymbol{\beta'}$$

avec $\boldsymbol{\beta}$ (resp. $\boldsymbol{\beta'}$) ayant des composantes non nulles en dehors de T (resp. T'). Alors on a $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta'}$ et le modèle n'est donc pas identifiable. D'où la condition $\delta_{2S} < 1$. (Imposer $\delta_{2S} + \theta_{S,2S} < 1$ ou $< 1 - t$ est à peine plus contraignant.)

On cite également la variante de cet estimateur appelée "Gauss-Dantzig selector".

1.3.4.3 Gauss-Dantzig selector

La procédure d'estimation se fait en deux phases :

1. On estime $I = \{j : \beta_j \neq 0\}$ par $I^{DS} = \{j : \beta_j^{DS} \neq 0\}$ où $\boldsymbol{\beta}^{DS}$ est calculé par (1.13) (ou de manière plus générale, par $I^{DS} = \{j : |\beta_j^{DS}| > \alpha\sigma\}$ pour un certain $\alpha \geq 0$)
2. On construit l'estimateur :

$$\boldsymbol{\beta}^{DS}_{I^{DS}} = \left(\boldsymbol{X}^T_{I^{DS}} \boldsymbol{X}_{I^{DS}}\right)^{-1} \boldsymbol{X}^T_{I^{DS}} \boldsymbol{y}$$

les autres composantes de $\boldsymbol{\beta}^{DS}$ étant affectées à zéro.

Cet estimateur semble donner de bons résultats dans la pratique.
Cette procédure de construction de l'estimateur en deux étapes est proche de celle que l'on adopte dans le chapitre 2. En effet, après avoir construit le chemin de régularisation et sélectionné le meilleur modèle du point de vue du BIC, on construit alors l'estimateur final comme dans la deuxième étape du Gauss-Dantzig selector, à savoir que l'on construit l'estimateur de régression logistique qui ne fait intervenir que les variables sélectionnées, calculé sans la pénalisation.

1.4 Description des chapitres

Nous décrivons brièvement ici ce que contient chacun des chapitres du présent document.

Le chapitre 2 expose le contexte des travaux menés sur la problématique industrielle dite d'« objectivation de la prestation Accroc-Croquement ». On définit tous ces termes dans le chapitre 2. On y décrit également avec précision le jeu de données réel que nous avons traité. On présente ensuite la nouvelle méthodogie d'analyse discriminante que l'on a appliqué à ce cas réel. Cette méthodologie est une approche en deux étapes. La première étape consiste en la construction du chemin de régularisation et la seconde consiste à sélectionner, parmi la suite des modèles issus du chemin de régularisation, le meilleur modèle, au sens du critère BIC. Cette méthodologie a fait l'objet d'une publication dans le journal du Bentley College "Case Studies in Business, Industry and Government Statistics" (CSBIGS). L'article intitulé "Pampering the Client : Calibration Vehicle Parts to Satisfy Customers" est reproduit en annexe 6.1 de ce document. Les résultats y sont détaillés, c'est pourquoi on ne donne qu'un bref rappel de ceux-ci dans le chapitre 2. En revanche, nous développons ici la discussion qui est seulement entamée dans l'article. Enfin, nous décrivons comment cette méthodologie a fait l'objet d'une industrialisation en interne chez RENAULT.

Ces techniques de chemins de régularisation sont encore peu utilisées et leur potentiel demande encore un certain nombre de travaux d'exploration. Le chapitre 3 renvoit à l'article que l'on cite en annexe 6.3. La version de l'article reproduit en annexe n'est pas une version finale. La contribution apportée par cet article sont des résultats de consistence uniforme, ainsi que de théorème de la limite centrale, dans le cas général d'un M-estimateur pénalisé, sous des hypothèses faibles sur le crit'ere d'attache aux données et sur la fonction de pénalisation. On utilise en particulier l'approche de Pollard [38] pour obtenir des conditions très générales sur la partie d'attache aux données. Le théorème 4 de l'article cité en annexe 6.3 montre que si les hypothèses de Pollard sont vérifiées alors l'estimateur considéré reste consistent uniformément en λ et il existe encore un TLC (fonctionnel) quand on ajoute une pénalisation au contraste. Les exemples fournis par Pollard nous donne ainsi accès à une large classe d'estimateurs pour lesquels des résultats asymptotiques sont valables en contexte pénalisé. La partie 3.2 de ce document montre que les hypothèses faites sur la fonction de pénalisation sont vérifiées pour une classe de fonctions de pénalisation classiques, qui contient les pénalisations L_1 et L_2 qui sont les plus utilisées.

Observant que l'approche décrite dans le chapitre 2 traite du problème de l'objectivation d'une prestation unique, on propose, dans le chapitre 4, d'élargir la problématique à l'étude simultanée de plusieurs prestations. On parlera de traitement « multi-prestations », par opposition à ce que l'on a fait jusqu'à présent et auquel on se référera sous le terme de traitement « mono-prestation ». En réalité la problématique « multi-prestations » est l'étude d'une prestation globale (ou note globale) qui se décompose sur un ensemble de prestations intermédiaires (ou notes

partielles). Ces cas de multi-prestations sont fréquents dans la pratique. Dans la partie 4.1, nous détaillons un exemple industriel de multi-prestations. Jusqu'alors, chez RENAULT, il n'y a pas de méthodologie, et encore moins d'outil pour traiter de tels cas. On propose dans le chapitre 4 deux approches : une approche que l'on qualifie de directe et qu'on détaille dans la partie 4.3 et une approche que l'on qualifie de hiérarchique (voir partie 4.4). On obtient donc deux procédures d'estimation que l'on compare via leurs courbes ROC. Nous effectuons cette comparaison sur des données simulées, n'ayant pas eu accès à temps à un jeu de données réel.

Il apparaît que l'approche hiérarchique a de meilleures performances, dans tous les cas de figure. Pour l'appliquer telle qu'elle est décrite dans le chapitre 4, certaines hypothèses doivent être vérifiées, notamment qu'il n'y ait pas de variables explicatives qui ne seraient pas observées et qui pourtant seraient significativement explicatives d'une prestation. Le but du chapitre 5 est d'explorer ce problème. Les travaux que nous présentons dans ce chapitre 5 peuvent servir dans un contexte aussi bien de mono-prestation que de multi-prestations. En effet, la question de l'exhaustivité de l'ensemble des variables potentiellement explicatives est une question que les ingénieurs métiers posent souvent en pratique. L'introduction de variables cachées fait appel à des techniques relatives à l'algorithme EM. Dans la partie 5.2, on rappellera le principe de cet algorithme avant de voir, en partie 5.3, sa mise en œuvre dans le contexte des chemins de régularisation. On réalise des simulations où l'on met en concurrence, via les fonctions de contraste pénalisées calculées pour chacune d'elles, la modélisation avec variables cachées et celle sans variables cachées afin de déterminer à quel moment, le long du chemin de régularisation, le modèle avec variables cachées devient plus explicatif que le modèle sans variables cachées.

Chapitre 2

Sélection de modèle basée sur les chemins de régularisation

Cette partie est centrée autour de l'article reproduit en annexe 6.1 et intitulé *Pampering the Client : Calibration Vehicle Parts to Satisfy Customers* et publié dans le journal du Bentley College "Case Studies in Business, Industry and Government Statistics" (CSBIGS). On donne ci-après le résumé en français de l'article en anglais.

Dans la suite de ce chapitre, nous présentons le contexte général de l'objectivation des prestations dans lequel s'inscrit nos travaux (voir partie 2.1). Nous décrivons plus précisément la prestation sur laquelle porte notre étude, à savoir l'« Accroc-Croquement » (partie 2.2). On s'attache ensuite à décrire plus précisément les données sur lesquelles nous avons testé notre nouvelle approche de sélection de modèle (en 2.3). Après un bref rappel de la méthodologie appliquée (voir 2.4), nous décrivons les résultats que nous avons obtenus (voir partie 2.5). La partie 2.6 apporte des précisions sur la discussion débutée dans l'article. Après avoir décrit comment cette problématique était traitée auparavant au sein de RENAULT (voir paragraphe 2.7.1), et ce qu'a apporté notre nouvelle méthodologie, on détaille dans le paragraphe 2.7.2 comment cette méthodologie a été implémentée en tant que logiciel interne RENAULT.

Résumé de l'article Nous nous proposons dans ce papier de répondre au problème industriel suivant. Les constructeurs automobiles doivent mettre au point des véhicules dont le niveau de qualité doit satisfaire le client. Nous nous concentrons sur l'aspect passage de vitesse. Notre étude se base sur 507 évaluations réalisées sur 28 différentes configurations. Chaque configuration est décrite par 12 paramètres physiques. Nous proposons une procédure de sélection et de mise au point des

paramètres physiques qui ont un impact sur les évaluations des testeurs. Notre approche se déroule en deux temps. On construit d'abord un chemin de régularisation basé sur un modèle de régression logistique pénalisée L1. On extrait de ce chemin une suite croissante de modèles. Dans un deuxième temps, on sélectionne à l'aide du critère BIC un modèle parmi la suite de modèles obtenue lors de la première étape. Nous proposons dans cet article un algorithme de résolution numérique simple, spécifique à notre approche. Enfin, nous discutons son application aux données de l'étude.

Avant de rentrer à proprement parler dans la description de la problématique pour laquelle nous avons développé cette nouvelle méthodologie d'objectivation des prestations, il est nécessaire de préciser ne serait-ce que ces deux termes.

Objectivation On parle d'« objectivation » dès lors que l'on cherche à expliquer des données subjectives au moyen de données objectives. La donnée subjective à expliquer est souvent le ressenti du client et les données objectives sont souvent des critères physiques mesurés qui caractérisent de manière pertinente l'item sur lequel le client exprime son ressenti.

Prestations Pour un véhicule, l'ensemble des « prestations » est l'ensemble des items sur lesquels le client peut donner son ressenti, ou en d'autres termes, s'exprimer subjectivement. Une prestation est donc une donnée subjective que l'on va vouloir « objectiver ».

2.1 Le « V » de la prestation

On se propose ici de décrire comment sont analysées et déployées les prestations au sein de RENAULT. On parle du « V » de la prestation dans la mesure où le développement des prestations se déroule en deux phases : une phase descendante suivie d'une phase ascendante.
Chacune des phases se décompose en une suite de niveaux. Notons qu'à chaque niveau correspond un cahier des charges, à savoir un document qui liste un certain nombre d'engagements à tenir pour autoriser le nouveau véhicule à partir en production. On parlera de cahier des charges de niveau 1 pour le premier niveau, décrit ci-après, et ainsi de suite pour les niveaux suivants.

– La phase descendante commence par la définition de **ce que veut le client**. Pour déterminer ce que veut le client, RENAULT se base sur des études auprès de sa clientèle (et de la clientèle potentielle) afin de comprendre quelles sont les attentes des clients pour cette prestation en particulier. La presse spécialisée est aussi une source d'information sur ce que recherchent les clients. Prenons un exemple et supposons que les clients de véhicules

hauts de gamme veulent des véhicules avec peu de nuisances sonores dans l'habitacle.

- Vient ensuite **ce que RENAULT veut pour le client**, afin de définir un niveau de prestation en accord avec l'identité de marque de RENAULT. Il s'agit ici de se positionner par rapport aux attentes des clients. Pour rester sur l'exemple de l'acoustique des véhicules hauts de gamme, on va s'intéresser à la prestation « acoustique habitacle » et RENAULT va se positionner en ciblant un niveau pour cette prestation, par exemple être le meilleur du marché sur cet item « l'acoustique habitacle ».
- Le niveau précédent est ensuite décliné sur chacune des prestations. Les ingénieurs concernés par cette prestation doivent alors déterminer **comment traduire physiquement cette attente** dans le véhicule. Pour cela, les ingénieurs acousticiens - pour rester dans notre exemple - vont réaliser des études avec des clients qu'ils vont inviter à venir juger de la qualité acoustique de différents habitacles afin de recueillir leur ressenti. En parallèle, les différents habitacles sont mesurés sur plusieurs critères physiques qui caractérisent la prestation « acoustique habitacle ». Les ingénieurs construisent ainsi une base de données qui pourra être analysée à l'aide des outils d'objectivation, puisqu'il s'agit bien ici d'expliquer des données subjectives (le ressenti du client en terme de confort acoustique dans l'habitacle) au moyen de critères physiques (données objectives).
- Pour aller jusqu'à la conception des pièces du véhicule qui atteindra le niveau de prestation défini plus haut dans le « V » de la prestation, il s'agit maintenant de déterminer **comment tout cela se décline au niveau de l'organe**. Dans notre exemple, ce niveau de cahier des charges correspond au pré-dimensionnement des pièces spécifiques de l'habitacle qui vont jouer sur le rendu acoustique.
- Les derniers niveaux de définition correspondent à **la solution technique** adoptée. Ces niveaux sont aussi détaillés que nécessaire. Pour revenir à nos pièces acoustiques pour l'habitacle, il s'agit dans ces dernières étapes d'inscrire au cahier des charges tous les détails techniques, à savoir par exemple les endroits où l'on dispose telle quantité de tel isolant phonique.

La phase ascendante est une phase de validation et comporte autant de niveaux que la phase descendante. On commence par vérifier que les pièces de la solution technique respectent bien le cahier des charges spécifié. Puis on valide l'organe, puis enfin le véhicule. Les validations sont effectuées par des experts des prestations. Il serait plus rassurant d'avoir une validation venant des clients eux-mêmes plutôt qu'une validation venant des experts du domaine, mais ça n'est pas toujours possible. Le client ne peut pas toujours donner son ressenti sur des prestations trop ciblées et/ou sur des niveaux de détails trop précis, compte tenu du fait qu'un client

n'a pas conscience de l'effet de telle ou telle pièce de la solution technique sur son ressenti global. L'expert, lui, est formé pour en être capable. En revanche, il faut impérativement s'assurer que le ressenti du client dans le véhicule équipé de la solution technique spécifiée est bien conforme au cahier des charges. Il va sans dire qu'il est recommandé d'impliquer le plus tôt possible le client dans la phase de validation.

2.2 Description de la problématique : objectivation de la prestation « Accroc-Croquement »

Dans l'article, nous n'avons pas pu détailler la problématique qui est à l'origine de nos travaux, à savoir l'objectivation de la prestation « Accroc-Croquement ». Ce document est l'occasion de décrire plus concrètement ce dont il s'agit.

Il s'agit d'une prestation liée au passage de vitesse. Lors d'un passage de vitesse, un certain nombre de sensations sont transmises au conducteur par l'intermédiaire du levier de vitesse. En effet, quand le conducteur actionne le pommeau, il soumet la commande externe de boîte de vitesse à certains déplacements, soit en sélection (mouvements de gauche à droite), soit en passage (mouvements d'avant en arrière). Mais la commande externe de boîte soumet également le conducteur principalement à des efforts de réaction et à des vibrations.

Cette prestation s'appelle « l'Accroc-Croquement ». Elle correspond à la difficulté que le conducteur peut ressentir quand il change de vitesse (mouvement de passage). Souvent, on entendra dire que « la vitesse accroche » mais l'intensité du phénomène est variable. Les ingénieurs experts de cette prestation parleront dans ce cas de « leger accroc ». Mais le phénomène peut aller jusqu'à l'impossibilité d'enclencher la vitesse et le conducteur est obligé de s'y prendre à deux fois. Les experts parleront alors de « croquement » voire de « craquement » dans les cas extrêmes. Ce type de désagréments est évidemment à proscrire.

Tous les ans, RENAULT commande des enquêtes de satisfaction auprès des clients qui ont acquis récemment un nouveau véhicule, afin de recenser les problèmes les plus importants rencontrés au cours des six premiers mois. Ces enquêtes sont menées par un institut indépendant trans-constructeurs : il s'agit des études "JD Power". Sur l'item "Hard to operate and Gear grinding" (qu'on peut traduire approximativement en français par « Commande difficile et Grincement des vitesses ») de l'enquête "JD Power" de 2005, RENAULT se place en bas du classement à la fois pour les berlines et pour les monospaces du segment moyen (véhicules de type Mégane et Scénic), avec un nombre de plaintes pour 100 véhicules de 9 pour les

berlines et de 14 pour les monospaces (le véhicule recevant le moins de plaintes sur cet items est à 3). D'autre part, le « Top 20 France » de 2006 qui liste les 20 principales sources de plaintes des clients RENAULT, place l'item "Manual Trans - Difficult To Get In Gear(s)" (Passage de vitesse - Difficulté à enclencher les vitesses) en deuxième position. Les items "Manual Trans - Gears Grind When Shifting" (Passage de vitesse - Grincement au changement de vitesse) et "Manual Trans - Gearshift Hard To Operate" (Passage de vitesse - Difficultés à manipuler le levier de vitesses) apparaissent également dans les dix premières causes de plaintes.

D'où le fort intérêt pour cette problématique au sein de RENAULT. C'est donc dans ce contexte qu'a été décidée une action "métier" par le service Politique Technique Agrément au sein de la Direction de la Mécanique. Le but est de mieux comprendre les prestations liées à la commande externe de boîte de vitesse, ce qui pourrait conduire à une refonte du cahier des charges. En particulier, sur la prestation Accroc-Croquement, les métiers ont constaté que certains véhicules, validés en interne sur cette prestation (phase ascendante du « V » de la prestation), sont mal cotés par les clients et inversement, certains véhicules non-conformes au cahier des charges sont bien cotés par le client.

2.3 Description des données

Nous avons ainsi collaboré de manière très étroite avec le service Potilique Technique Agrément avec qui nous avons construit une base de données pour l'étude de la prestation Accroc-Croquement. Le constat du fort taux d'erreurs de reclassement [1] - aussi bien en termes de faux négatifs que de faux positifs - a conduit à remettre en cause le cahier des charges en vigueur. Le cahier des charges contenait un critère d'acception basé sur un unique critère physique, relatif à l'effort à fournir lors du passage.

Les métiers ont donc proposé une liste de nouveaux critères physiques, potentiellement explicatifs de la prestation Accroc-Croquement. Dans ce travail de recherche de nouveau critères physiques, nous avons invité les ingénieurs à être les plus exhaustifs possible dans le but de ne pas passer à côté d'une information susceptible d'expliquer la prestation.

[1] Il s'agit ici d'appliquer le cahier des charges et de reclasser les essais en deux catégories :
- les essais « bons », qui sont acceptés par le cahier des charges
- et les essais « mauvais », qui sont refusés par le cahier des charges.

On compte la proportion d'essais que l'on sait être bien jugés subjectivement et qui sont refusés au cahier des charges. Il s'agit du taux de faux négatifs. De même, on compte la proportion d'essais que l'on sait être mal jugés subjectivement et qui sont acceptés au cahier des charges. Il s'agit du taux de faux positifs.

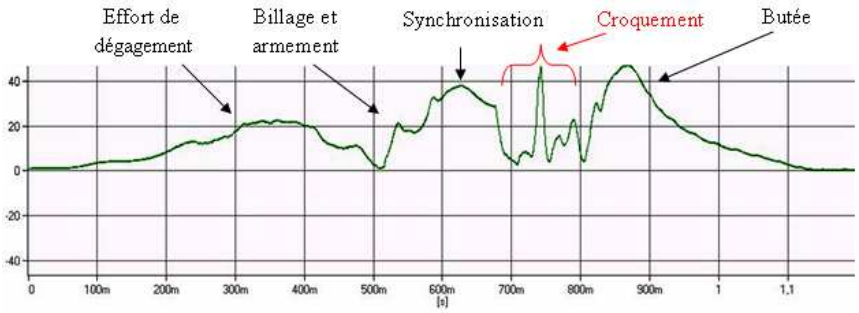

FIG. 2.1 – Courbe d'effort ressenti au pommeau en fonction du temps lors d'un passage de vitesse.

Cette recherche a abouti à la liste de critères physiques présentée ci-dessous. On trouvera en annexe 6.2 une description des différents critères physiques potentiellement explicatifs de cette prestation. Les critères physiques peuvent être regroupés en trois catégories : les critères d'effort (ou critères de force), les critères énergétiques et les autres critères.

Lors d'un passage, les efforts de dégagement, de synchronisation et de butée parviennent au client et sont interprêtés comme des informations importantes pour contrôler l'action, alors que l'effort de crabotage, s'il est trop important, est lui perçu comme une perturbation gênante (voir fig. 2.1). L'effort de dégagement permet de quitter la vitesse qui était enclenchée, l'effort de synchronisation indique que le nouveau rapport de boîte de vitesse commence à s'enclencher et l'effort de butée est simplement l'effort subi indiquant qu'on est au bout de la course du levier et que la vitesse est bien enclenchée. Le crabotage est la mise en contact du pignon fou de l'axe secondaire (après avoir été ralenti - ou accéléré par l'action au manchon baladeur) avec le pignon fixe correspondant sur l'axe primaire. Les pièces sont alors en phase et le crabotage consiste à les rapprocher jusqu'à avoir un mouvement solidaire. Ces pièces sont munies de dents hélicoïdales et bien que les pièces tournent à la même vitesse, le rapprochement peut provoquer un choc dû au non-alignement des dents. Le phénomène de crabotage a toujours lieu, tandis qu'on dira que le phénomène de croquement (quand l'effort de crabotage est très important) lui, est statistique, dans le sens où il n'arrive pas à chaque passage. La question n'est donc pas d'éviter l'apparition de l'accroc, mais plutôt de réduire le choc ressenti par le client au pommeau quand l'effort entre les dents des pignons est important.

La présence de chocs ont conduit les ingénieurs à considérer des critères énergé-

tiques, notamment en s'intéressant aux spectres fréquentiels des signaux observés. Enfin, d'autres critères, comme la température ont été jugés par les experts comme potentiellement explicatifs.

Voici la liste des 15 critères physiques jugés potentiellement explicatifs de la prestation Accroc-Croquement :

Critères de force

- Fc : effort de crabotage maximal
- Fc/Fs : rapport entre l'effort de crabotage maximal et l'effort de synchronisation maximal
- nombre de pics
- recul
- angle de recul

Critères énergétiques

- énergie cinétique
- impulsion de passage
- impulsion de crabotage
- impulsion de recul
- puissance de recul
- RMS1
- RMS2
- RMS3

Autres critères

- temps de passage
- T (température)

Ces critères physiques ont donc été mesurés sur chacun des 565 essais réalisés. Chaque essai est constitué d'un ensemble de cinq passages dans les mêmes conditions, pour prendre en compte le caractère non systématique du phénomène. Du fait du caractère très ciblé de la prestation que l'on cherche à étudier, les essayeurs ne sont pas ici des clients, mais des experts de l'agrément de conduite, compétents pour juger de la gravité du défaut rencontré. Le panel des essayeurs comprend six personnes différentes de manière à donner une certaine variété à la population test. Le panel s'est préalablement mis d'accord sur une grille de cotation subjective où le ressenti est classé dans une des quatre catégories de gravité suivante :

1. impact nul : bon, correct
2. impact faible : léger accroc, grattage
3. impact moyen : accroc, léger croquement, léger rebond
4. fort impact : croquement, craquement

La base de données a également été construite pour être assez proche de la variété des configurations existantes dans la rue. Cinq boîtes de vitesse ont été testées :

- boîte PK4 : 301 essais/565 essais
- boîte JR5 : 98/565
- boîte TL4 : 58/565
- boîte ND0 : 49/565
- boîte HONDA : 59/565

Il y a globalement autant d'essais qui ont été réalisés à chaud qu'à froid.

- essais réalisés à chaud : 278/565
- essais réalisés à froid : 287/565

De même, il y a un bon partage entre les essais réalisés sur banc moteur (essai moins coûteux) que d'essais réalisés sur véhicule, en condition réelle de conduite.

- essais réalisés sur banc : 306/565
- essais réalisés sur véhicule : 259/565

2.4 Méthodologie appliquée : principe général

La méthodologie que nous proposons d'appliquer pour analyser et déployer les prestations est la suivante. Le but final de la démarche est de trouver un modèle de discrimination optimal. Nous définissons ci-après en quoi il est optimal. Ce modèle fait intervenir un certain nombre de variables explicatives. Il est possible que toutes les variables explicatives ne soient pas nécessaires à l'explication de la prestation. La taille du modèle, telle qu'elle a été définie dans l'équation (1.6) est liée au nombre de variables explicatives qui interviennent dans le modèle optimal. Ce nombre est évidemment inconnu. Nous proposons de procéder en deux étapes.

La première étape de la méthodologie consiste à faire varier la taille du modèle, pour connaître le modèle associé à chacune des tailles de modèle parcourue. On trouve ainsi le meilleur modèle à une variable, puis le meilleur modèle à deux variables, etc. Cette étape peut être vue comme une hiérarchisation des variables explicatives ou encore comme la construction d'une suite croissante - en taille - de modèles optimaux.

La seconde étape est une étape classique de sélection de modèle. Parmi la suite de meilleurs modèles que l'on vient de construire, on sélectionne le *meilleur* modèle.

2.4.1 Première étape

La première partie consiste en la construction du chemin de régularisation. On rappelle que le chemin de régularisation est défini comme étant l'ensemble des vecteurs $\hat{\boldsymbol{\beta}}(\lambda)$ quand λ parcourt $[0, +\infty[$. $\hat{\boldsymbol{\beta}}(\lambda)$ est défini par :

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \{-\log L_n(\boldsymbol{\beta}) + \lambda J(\boldsymbol{\beta})\}$$

où L_n est donc ici la vraisemblance logistique et $J(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_1$.

Comme on l'a vu également plus haut, il n'existe pas de résolution analytique de ce problème dans le cas de la régression logistique, contrairement au cas de la régression linéaire (résolution par l'algorithme LARS).

L'objectif de cette première étape n'est d'ailleurs pas de trouver tous les vecteurs $\hat{\boldsymbol{\beta}}(\lambda)$ quand λ parcourt $[0, +\infty[$. Seuls nous importent les instants où l'ensemble des variables actives est modifié. On décide donc de développer un algorihme simplifié recherchant uniquement les instants où les variables entrent dans le modèle.

On remarque tout d'abord qu'il existe une certaine valeur pour λ au-delà de laquelle toutes les composantes du vecteur $||J(\beta)||_1$ sont nulles. On appelle cette valeur λ_{max}. On calcule le modèle associé à $\lambda = \lambda_{max}$, qui se trouve être le modèle constant. On calcule également le modèle associé à $\lambda = 0$, qui contient les p variables explicatives. On procède ensuite à une *recherche dichotomique* des λ intéressants. En l'occurence, on commence par calculer le modèle associé à $\lambda = \lambda_{max}/2$:

$$\hat{\boldsymbol{\beta}}(\lambda_{max}/2) = \arg\min_{\boldsymbol{\beta}} \left\{ -\log L_n(\boldsymbol{\beta}) + \frac{\lambda_{max}}{2} J(\boldsymbol{\beta}) \right\}$$

On compare l'ensemble actif associé à $\hat{\boldsymbol{\beta}}(\lambda_{max}/2)$ aux ensembles actifs déjà calculés. On continue la recherche dichotomique jusqu'à obtenir une suite d'ensembles actifs qui diffèrent au plus d'une variable. Il est possible que si l'on classe les ensembles actifs en fonction de λ par exemple, deux ensembles actifs consécutifs soient identiques. On verra dans la suite comment éliminer ces éventuels doublons. Pour des raisons d'efficacité algorithmique, on se fixe un seuil en-dessous duquel on ne cherchera pas à poursuivre la recherche dichotomique. Ce seuil est arbitrairement fixé à $\Delta\lambda = \lambda_{max}/1000$. On s'autorise ainsi à ce que deux ensembles actifs consécutifs diffèrent de plus d'une variable. Cela n'a pas d'incidence sur les résultats : il suffit d'en tenir compte lors de l'interprétation.

2.4.2 Seconde étape

Une fois la première étape réalisée, on a à disposition une suite d'ensembles actifs et les modèles associés. Il est très probable que la recherche dichotomique ait conduit cette suite d'ensembles à contenir des doublons, c'est-à-dire des ensembles consécutifs identiques. D'où la nécessité d'enlever ces doublons pour ne garder, par exemple, qu'un seul ensemble actif à k variables. On garde l'ensemble actif correspondant au λ le plus petit (l'ensemble retenu est celui qui correspond au plus grand $||\hat{\boldsymbol{\beta}}||_1$). On obtient ainsi une suite croissante de modèles explicatifs où deux modèles successifs différent d'exactement une variable (sauf cas limite où plus d'une variable entrent simultanément).

La suite de l'algorithme consiste à sélectionner, parmi cette suite croissante de modèles, le meilleur modèle pour expliquer la prestation. Cette sélection est effectuée grâce au *Bayesian Information Criterion* (BIC).

2.5 Description des résultats

La méthodologie appliquée à ces données fournit les résultats suivants. Le chemin de régularisation a permis de hiérarchiser les variables, en donnant la liste des variables par ordre d'importance décroissante quant à leur pouvoir explicatif de la prestation Accroc-Croquement. On choisit ensuite le meilleur de ces modèles, meilleur au sens qu'il minimise le BIC. Ce meilleur modèle contient deux critères physiques, et deux seulement. Ces deux critères physiques sont les suivants : Var 5 et Var 1.
Le fait que Var 5, critère physique en application dans le cahier des charges avant le lancement de cette nouvelle étude, ressorte parmi les critères physiques significativement explicatifs est rassurant du point de vue de l'expertise RENAULT. Cela signifie que le travail mené jusqu'à présent n'est pas dénué de sens. À ce critère physique s'ajoute un autre critère qui est un niveau d'énergie : Var 1. Il apparaît donc que l'approche énergétique du phénomène est intéressante.

La phase suivante, pour aller jusqu'à la refonte du cahier des charges est une phase de validation. La combinaison des deux critères physiques fournit un nouveau critère qui doit être testé pour que le cahier des charges soit modifié. Une nouvelle base de presque 4000 essais va servir de validation à ce nouveau critère.

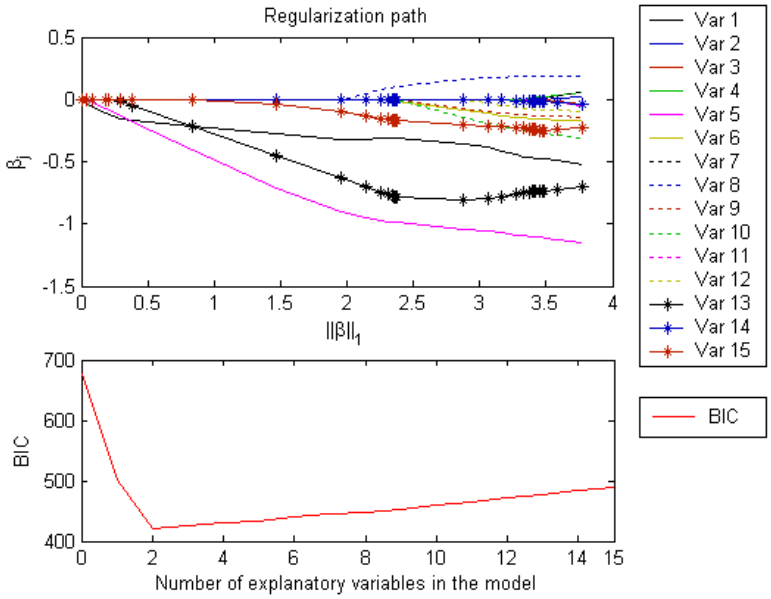

FIG. 2.2 – Chemin de régularisation et BIC sur les 15 variables initialement présentes dans l'étude

2.6 Précisions sur la discussion

Dans cette section, nous revenons sur la discussion qui conclut l'article, en y apportant quelques précisions.

On rappelle en fig. 2.2 le chemin de régularisation obtenu avec les 15 critères physiques listés dans le paragraphe 2.3. Comme on vient de le rappeler, le BIC sélectionne le modèle à deux variables. Les deux variables qui constituent alors le modèle sont les suivantes : Var 5 et (Var 1).
D'autre part, on voit qu'à mesure que croît $||\beta||_1$, Var 1 et Var 5 ne restent pas les deux variables ayant les deux plus fortes amplitudes en valeur absolue.

Les connaissances actuelles sur les chemins de régularisation appliqués sur des cas pratiques sont assez minces. On ne sait pas bien aujourd'hui comment doit s'interpréter ce type de chemin, notamment quand celui-ci est emmêlé. Se pose alors la question de savoir quel modèle considérer : est-ce qu'il faut considérer :

- les deux premières variables qui entrent dans la modèle, à savoir Var 5 et Var 1
- ou les deux variables qui possèdent les coefficients les plus grands en valeur

absolue au long du chemin, à savoir Var 5 et Var 13 ?

Le principal argument en faveur de la première approche est le suivant : si l'on cherche le meilleur modèle à une variable, par exemple en contraignant le problème d'optimisation à ne retenir qu'une et une seule variable pour expliquer la prestation, alors c'est la variable globalement la plus corrélée avec la réponse qui est retenue. Dans l'algorithme, c'est donc la première variable à entrer dans le modèle.

La façon appropriée de comprendre comment ces chemins de régularisation doivent être interprétés est d'étudier leur comportement asymptotique afin d'en déduire des préconisations dans le cas non asymptotique. Les travaux récents de Greenshtein et Ritov, dont les principaux résultats en termes de persistence ont été rappelés dans l'introduction, viennent plutôt accréditer la seconde approche. En effet, leurs résultats semblent sélectionner de plus grands modèles que ceux sélectionnés par le BIC. Dans les cas pratiques, et encore plus particulièrement dans le nôtre, nous ne pouvons raisonnablement pas supposer être dans le cas asymptotique. L'hypothèse de Greenshtein et Ritov par exemple est que le nombre de variables explicatives p croît exponentiellement avec le nombre d'essais n : $p = n^{\alpha}$, avec $\alpha > 1$. Or dans notre cas pratique, $p = 15$ et $n = 565$.

D'où la difficulté d'interprétation posée par les chemins de régularisation où les premières variables qui entrent dans le modèle ne sont pas celles qui ont les plus fortes valeurs de coefficients au cours du chemin. Pour résoudre ce point bloquant, nous avons conclu, après duscussion avec les ingénieurs, que l'interprétation serait facilitée si les deux approches coïncidaient. On a donc cherché à faire en sorte que le chemin soit « peigné », c'est-à-dire que l'on se trouve dans la situation où :

- les premières variables qui entrent dans le modèle
- sont celles qui conservent les coefficients les plus grands en valeurs absolues au long du chemin.

Ainsi le chemin de régularisation est « peigné », *i.e.* qu'il n'y a pas de croisements des différentes trajectoires des coefficients β_j comme fonctions de $||\beta||_1$.

Le travail réalisé avec les experts RENAULT a abouti à remarquer que les deux ensembles $\{$Var 5,Var 1$\}$ et $\{$Var 5,Var 13$\}$ sont très comparables. A partir du constat que les deux critères physiques Var 1 et Var 13 décrivent des réalités physiques très proches, les ingénieurs ont jugé recevable de ne conserver qu'un des deux critères physiques. Et plus généralement, des trois critères physiques Var 1, Var 13 et Var 14, on n'a conservé que Var 1 comme « représentant ». De même, il est apparu que Var 15 et Var 5 avaient des influences similaires sur le chemin. En accord avec les ingénieurs, Var 15 a également été sorti de l'étude. On obtient ainsi le chemin tel que représenté dans la figure 1 de l'article, figure qui est rappelée en fig. 2.3. On note que le chemin n'est pas parfaitement « peigné » puisque les trajectoires des variables Var 1 et Var 5 se croisent encore. En accord avec les

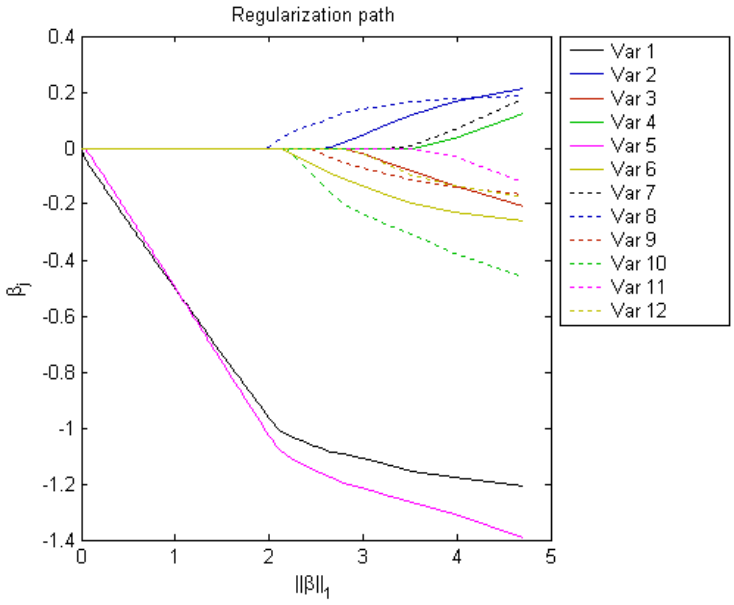

FIG. 2.3 – Chemin de régularisation sur les 15 variables sauf Var 13, Var 14 et Var 15

ingénieurs, ce chemin de régularisation a été jugé satisfaisant, du moment que les entrées dans le modèle de ces deux variables sont quasiment simultanées : cela signifie qu'un faible relâchement de la contrainte sur $||\boldsymbol{\beta}||_1$ suffit pour les autoriser à entrer toutes les deux. Elles constituent donc un groupe que l'on opposera aux variables restantes, prises elles séparément.

Même avec ce travail permettant d'accorder les deux approches différentes d'interprétation du chemin de régularisation, on peut émettre une critique quant à l'utilisation du BIC dans la seconde phase de la méthodologie. En effet, on rappelle que pour utiliser le BIC en sélection de modèle, il est nécessaire d'avoir une suite de modèles *emboîtés*. Or il arrive qu'au cours de l'algorithme, certaines variables sortent du modèle courant - on dit qu'une variable « sort » si le coefficient correspondant s'annule à nouveau. Nous ne sommes donc pas assurés d'avoir à la fin une suite de modèles qui soit réellement emboîtée. Néanmoins,à défaut d'un autre critère de sélection de modèle, l'utilisation du BIC est un choix qui permet d'éviter un choix encore plus arbitraire. D'autre part, l'expérience nous montre que le phénomène de variables sortantes reste finalement assez rare et les chemins sont souvents "peignés". Dans l'optique d'implémenter la méthodologie en une solu-

tion logicielle, comme cela a été le cas ici, on doit de toutes façons savoir traiter le cas des variables sortantes, aussi rare soit-il.

2.7 Implémentation de la solution logicielle

Nous avons en effet eu la chance de pouvoir participer à l'industrialisation de notre solution algorithmique pour l'objectivation des prestations. Par industrialisation, on entend développement d'un logiciel maison, disponible en interne par tous les ingénieurs désireux d'utiliser cet outil.

2.7.1 Outil existant : PRESTool

Un outil d'objectivation des prestations existait déjà, du nom de PRESTool, pour « prestations » et « *tool* ». Notre travail a permis de mettre à jour le module qui calcule le meilleur modèle de discrimination. Dans la suite du document, nous ferons référence à l'ancienne solution logicielle sous le nom de PRESTool et nous ferons référence à notre nouvelle version sous le nom de PRESTool-OCR (pour Objectivation par Chemins de Régularisation).

La présence de nombreuses contraintes dans la résolution numérique du problème d'optimisation rend les calculs assez lourds, notamment lorsque les dimensions du problème augmentent (en nombre d'observations n, comme en nombre de variables explicatives p). La principale critique formulée vis-à-vis de l'outil est donc l'importance des temps de calcul. Pour prendre un exemple, avec les 565 essais et les 15 variables explicatives de notre problème, l'algorithme nécessite plusieurs dizaines d'heures pour donner un résultat. L'objectif était de fournir aux ingénieurs RENAULT un outil plus rapide et qui réponde au moins de manière aussi pertinente à la problématique d'objectivation des prestations, pour qu'ils utilisent ce logiciel plutôt que de faire appel, par exemple, à des sociétés externes de prestations en statistiques. La nouvelle méthodologie implémentée dans l'outil PRESTool-OCR est rapide et permet une grande interactivité : on peut modifier les variables explicatives prises en compte et/ou les essais à considérer et voir rapidement l'effet sur le chemin de régularisation et/ou sur le BIC. En terme de temps de calcul, on note que l'exemple de l'Accroc-Croquement donne un résultat en un peu moins de 4 minutes.

2.7.2 PRESTool-OCR

Tout d'abord on se place dans le cas d'une réponse binaire. Si le nombre de modalités de la variable réponse est supérieur à deux, alors l'outil ne détermine plus

automatiquement le meilleur regroupement de modalités possibles. Le choix des deux modalités à opposer est laissé aux ingénieurs. En effet, les modalités qu'ont à gérer les ingénieurs sont souvent du type : {très bon, bon, mauvais, très mauvais}. Selon qu'ils veulent savoir ce qui rend la prestation très bonne ou s'ils veulent simplement savoir ce qui différencie globalement les occurences bonnes des occurences mauvaises, les regroupements sont les suivants : très bon contre {bon, mauvais, très mauvais} dans le premier cas et {très bon, bon} contre {mauvais, très mauvais} dans le second cas. La rapidité du nouvel outil autorise de faire plusieurs tests différents consécutifs.
Une autre raison du choix de travailler uniquement avec des données binaire est la suivante : les jeux de données sont rarement grands, donc on privilégie l'estimation d'un nombre plus réduit de paramètres (cas binaire) plutôt qu'un grand nombre de paramètres qui seraient mal estimés (cas multi-modal).

Parmi les contraintes que nous n'avons pas citées plus haut, il y a dans PRESTool des fonctions de pénalisation supplémentaires qui forcent le modèle de discrimination à ne pas conserver toutes les variables explicatives. On peut également forcer PRESTool à trouver le meilleur modèle à exactement k variables. Un des avantages à travailler grâce aux chemins de régularisation est que l'on a accès simultanément, et par simple lecture graphique du graphique des évolutions des coefficients en fonction de $||\beta||_1$, à tous les modèles de discrimination possibles. Le BIC est ensuite une aide au choix de la taille optimale du modèle de discrimination.

L'implémentation de la solution logicielle PRESTool-OCR suit l'algorithme décrit dans l'article cité en annexe 6.1, c'est-à-dire qu'on applique bien l'approche en deux étapes, à savoir : tout d'abord, une sélection de variables issue de la hiérarchisation des critères physiques, puis sélection de modèle, via le BIC, parmi une suite de modèles emboités.

La hiérarchisation des variables explicatives répond aux attentes des ingénieurs RENAULT. En effet, ces derniers ont l'habitude de se poser les questions suivantes :

- « S'il n'y avait qu'une variable explicative à conserver, quelle serait-elle ? », ou encore
- « Et si on ne devait ne garder que deux variables explicatives ? »

Nous avons donc proposé d'utiliser les chemins de régularisation dans cette optique, en privilégiant les premières variables comme étant les plus explicatives.

2.7.3 Cas des variables sortantes

Lorsqu'on se place dans l'optique de développer un outil logiciel à l'intention d'utilisateurs qui ne sont pas familiers de la méthodologie, on se doit de proposer un outil le plus intuitive possible. C'est pourquoi l'on propose d'utiliser directement les sorties graphiques des chemins de régularisation, telles que celle représentée fig. 2.3. En effet, un des atouts des chemins de régularisation réside dans le fait que les représentations graphiques sont faciles à appréhender.

Malheureusement, la présence de variables sortantes complique grandement cette compréhension. Là encore, c'est l'expertise acquise avec les ingénieurs RENAULT qui a permis de proposer une solution pratique. Dans les cas où des variables sortent du modèle, on s'est aperçu que cette sortie de variable s'accompagne de l'entrée d'une - ou plusieurs - variables, ce qui est logique d'un point de vue optimisation, en regardant la forme de la fonction optimisée (équation (2) de l'article). La (ou les) variable qui remplace la variable sortante est proche de celle-ci quant à sa signification physique. Il n'y a donc pas lieu de s'attacher plus particulièrement à l'une ou à l'autre des variables qui représentent la même réalité physique. C'est ce que nous avons observé avec la classe de variables explicatives « équivalentes » $\{$Var 1, Var 13, Var 14$\}$: considérer l'une plutôt qu'une autre n'a pas d'apport significatif quant à l'explication de la prestation. Il en va de même pour l'ensemble des variables explicatives $\{$Var 15, Var 5$\}$.

Comme on l'a expliqué plus haut, par souci de clarté d'interprétation pour les utilisateurs de l'outil PRESTool-OCR, nous avons été amenés à *simplifier* le vrai chemin de régularisation. En effet, nous avons fait le choix, en accord avec les ingénieurs, de pleinement tirer profit de cette substitution entre variables explicatives de même "classe de variables équivalentes", pour simplifier le chemin sans altérer la pertinence du résultat. Pratiquement, dans l'algorithme dichotomique, on ne cherche pas les valeurs de λ où l'ensemble de variables actives change, mais on cherche uniquement les λ où le cardinal de cet ensemble change. Ainsi notre recherche dichotomique s'arrête quand on passe de k variables actives à $k+1$ variables actives, sauf si la recherche dichotomique atteint le seuil $\delta\lambda$ en-dessous duquel on arrête les subdivisions. En effet, connaître le λ exact qui sépare l'entrée d'une variable par rapport à l'entrée de l'autre n'apporte pas de gain significatif quant à l'interprétation.
Si une variable sort pour être remplacée par une autre, on peut être amené à calculer les ensembles actifs pour deux λ et tomber sur deux ensembles de variables différents, mais de même cardinal. Ce genre de cas est traité algorithmiquement comme suit : pour un cardinal d'ensemble de variables actives donné, on choisit l'ensemble actif correspondant au λ le plus petit, donc au $||\beta||_1$ le plus grand.

Le problème est légérement plus compliqué quand une variable est remplacée

par plusieurs autres variables. Dans ce cas, on préfère garder la trace de la variable qui sort, même si cela rend le chemin de régularisation peu lisible. Supposons que la première variable à entrer dans le modèle sort et soit remplacée par deux variables, alors le BIC peut choisir pour meilleur modèle, le modèle à une variable, plutôt que celui à deux variables (même si cela semble peu probable, au vu des considérations de proximité en sens physique des deux groupes de variables).

Chapitre 3

Consistence uniforme et théorème de la limite centrale pour des M-estimateurs pénalisés

Dans cette partie, nous développons quelques applications des résultats présentés dans l'article cité en annexe 6.3. Les travaux présentés dans ce chapitre 3 étendent les résultats de Knight et Fu [29] en étudiant le comportement asymptotique de M-estimateurs pénalisés. Ces estimateurs incluent les estimateurs de type Lasso. D'autres estimateurs du maximum de vraisemblance peuvent également être pénalisés, comme la régression logistique, qui est un cas particulier des modèles exponentiels (voir paragraphe 3.3).

L'article cité en annexe 6.3 s'articule comme suit. La Section 2 montre la consistence uniforme des M-estimateurs dépendant d'un paramètre ainsi qu'une application de ce résultat pour les M-estimateurs auxquels on ajoute une fonction de pénalité, par l'intermédiaire d'une constante de régularisation λ. Ce résultat de consistence des M-estimateurs pénalisés fait l'objet du Théorème 3. Dans la Section 3, un résultat similaire est obtenu sous des hypothèses supplémentaires de convexité. La Section 4 donne un résultat de convergence au second ordre pour les fonctions de contraste dépendant d'un paramètre (Théorème 5). Ce résultat permet de montrer un Théorème de la Limite Centrale (TLC) pour les M-estimateurs dépendant d'un paramètre (Théorème 6 de la Section 5) puis de la même manière un TLC pour les M-estimateurs pénalisés (Théorème 7 de la Section 6). On applique ensuite ces résultats de consistence et de TLC au cas du Lasso, dans la Section 7, où l'on trouve les preuves des Théorèmes 1 et 2, ainsi qu'une application au test d'hypothèse statistique, basé sur le chemin de régularisation. D'autres exemples sont donnés dans la Section 8, comme par exemple la consistence et le

TLC pour les modèles linéaires généralisés dans leur version avec pénalisation L_1, ainsi que pour le LAD pénalisé L_1.

Dans ce chapitre, on se propose de rappeler les résultats obtenus pour le Lasso. On montre en effet que sont vérifiées les conditions décrites dans l'article cité en annexe 6.3 pour la consistence uniforme et le théorème de la limite centrale fonctionnel. On montre, dans la partie 3.2, que les fonctions de pénalisation classiques, comme les pénalisations L_1 et L_2, vérifient les hypothèses du Théorème 4 de l'article de l'annexe 6.3. Dans la partie 3.3, on s'intéresse à d'autres formes de fonctions d'attache aux données (on dit aussi fonctions de contraste), comme par exemple les modèles exponentiels (voir partie 3.3.1). Enfin on décrit une application simple des résultats asymptotiques cités dans l'article en annexe 6.3 concernant le test d'hypothèse. On propose une statistique pour tester l'hypothèse nulle $H_0 : \{\boldsymbol{\beta} = 0\}$. On verra, dans la partie 3.4, que les résultats asymptotiques permettent de déterminer la loi asymptotique de cette statistique afin d'en tirer des p-valeurs. On comparera également cette statistique à la statistique de Fischer.

3.1 Application au Lasso

Dans cette partie, nous proposons d'appliquer au cas du Lasso les résultats en termes de consistence uniforme (paragraphe 3.1.1) et de théorème de la limite centrale (paragraphe 3.1.2).

Conformément aux notations utilisées dans l'article, on notera dans cette partie la constante de régularisation $\mathbf{t}$, au lieu de λ jusqu'à présent. Avec cette nouvelle notation, on rappelle la formulation du Lasso (équation (1.7)) :

$$\hat{\boldsymbol{\beta}}_n(\mathbf{t}) = \arg\min_{\boldsymbol{\beta}} \|Y_n - \boldsymbol{X}_n\boldsymbol{\beta}\|^2 + \mathbf{t}\|\boldsymbol{\beta}\|_1 \tag{3.1}$$

Le modèle considéré est un modèle linéaire : $Y_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n$, où $Y_n = (y_k)_{k=1,\dots,n} \in \mathbb{R}^n$,$\boldsymbol{X}_n \in \mathcal{M}_{n\times p}(\mathbb{R})$ ou encore : $y_k = \boldsymbol{x}_k^T\boldsymbol{\beta} + \varepsilon_k$, $((\boldsymbol{x}_k)_{k=1,\dots,n} \in \mathbb{R}^p)$. (ε_n) est supposé être un bruit fort centré de variance σ^2. En plus des notations de l'article de l'annexe 6.3 concernant la fonction de contraste, on adopte les notations prises dans Knight et Fu [29], concernant la fonction de pénalisation.

$$\begin{aligned} \Lambda_n(\boldsymbol{\phi}, t) &= M_n(\boldsymbol{\phi}) + \mathbf{t}J_n(\boldsymbol{\phi}) \\ &= \frac{1}{n}\|Y_n - \boldsymbol{X}_n\boldsymbol{\beta}\|^2 + \mathbf{t}\lambda_n\|\boldsymbol{\beta}\|_1 \end{aligned} \tag{3.2}$$

Nous retrouvons ci-dessous, sous des hypothèses identiques, les résultats de Knight et Fu [29] quant à la consistence et au théorème de la limite centrale en les améliorant, puisque nous en donnons des versions fonctionnelles.

Les hypothèses sont :

(KF–1) [Conditions pour la consistence]
$\lambda_n \to 0$ et $C_n = n^{-1}\boldsymbol{X}_n^T\boldsymbol{X}_n \to C$, où C est une matrice symétrique définie positive.

(KF–2) [Conditions pour le TLC]
(KF–1) est vraie, $\lambda_n = n^{-1/2}$ et $\max_{1\leq k\leq n}\|\boldsymbol{x}_k\|^2 = o(n)$.

3.1.1 Consistence uniforme du Lasso

Comme la fonction $\boldsymbol{\phi} \in \boldsymbol{\Phi} \mapsto M_n(\boldsymbol{\phi})$ est convexe, nous appliquons le résultat du Théorème 4 de l'article cité en annexe 6.3. Pour reprendre les mêmes notations, on écrit $M_n(\boldsymbol{\phi})$ sous la forme :

$$M_n(\boldsymbol{\phi}) = \frac{1}{n}\sum_{k=1}^{n}\delta((\boldsymbol{x}_k, y_k), \boldsymbol{\phi})$$

avec $\delta((\boldsymbol{x}_k, y_k), \boldsymbol{\phi}) = (y_k - \boldsymbol{x}_k^T\boldsymbol{\phi})^2$.
Pour pouvoir appliquer le Théorème 4, il faut vérifier les trois conditions suivantes :

(i) pour tout $y \in \mathcal{Y}$, $\delta(y, \cdot)$ est une fonction convexe de $\boldsymbol{\phi}$;

(ii) pour tout $\boldsymbol{\phi} \in \boldsymbol{\Phi}$, $M_n(\boldsymbol{\phi}) \to_P \Delta(\boldsymbol{\phi})$, strictement convexe ;

(iii) pour tout $\boldsymbol{\phi} \in \boldsymbol{\Phi}$, $\Delta(\boldsymbol{\phi}) \geq \Delta(\boldsymbol{\beta})$.

La première condition (i) est vérifiée de manière évidente.
Pour montrer la condition (ii), on récrit $M_n(\boldsymbol{\phi})$ de la manière suivante :

$$\begin{aligned} M_n(\boldsymbol{\phi}) &= \frac{1}{n}\|Y_n - \boldsymbol{X}_n\boldsymbol{\phi}\|^2 \\ &= \frac{1}{n}\|\varepsilon_n\boldsymbol{\beta} - \boldsymbol{X}_n(\boldsymbol{\phi}-\boldsymbol{\beta})\|^2 \\ &= \frac{1}{n}\|\varepsilon_n\boldsymbol{\beta}\|^2 + (\boldsymbol{\phi}-\boldsymbol{\beta})^T C_n(\boldsymbol{\phi}-\boldsymbol{\beta}) - \frac{2}{n}\varepsilon_n^T\boldsymbol{X}_n(\boldsymbol{\phi}-\boldsymbol{\beta}) \end{aligned}$$

avec $\varepsilon_n = Y_n - \boldsymbol{X}_n\boldsymbol{\beta}$. On a donc :

$$M_n(\boldsymbol{\phi}) - M_n(\boldsymbol{\beta}) = (\boldsymbol{\phi}-\boldsymbol{\beta})^T C_n(\boldsymbol{\phi}-\boldsymbol{\beta}) - \frac{2}{n}\varepsilon_n^T\boldsymbol{X}_n(\boldsymbol{\phi}-\boldsymbol{\beta}) \qquad (3.3)$$

$\|\boldsymbol{X}_n^T\varepsilon_n\| = O_P(\sqrt{n})$, car :

$$\mathbb{E}\left[\|\boldsymbol{X}_n^T\varepsilon_n\|^2\right] = \mathbb{E}[\mathrm{Tr}(\varepsilon_n^T\boldsymbol{X}_n\boldsymbol{X}_n^T\varepsilon_n)] = \mathrm{Tr}(\boldsymbol{X}_n\boldsymbol{X}_n^T) = O(n)\ ,$$

par (KF–1). Par suite $-\frac{2}{n}\varepsilon_n^T\boldsymbol{X}_n(\boldsymbol{\phi}-\boldsymbol{\beta}) = O_P(\frac{1}{\sqrt{n}})$. Et en particulier, ce terme tend vers 0 en probabilité. Ainsi, par (KF–1) :

$$M_n(\boldsymbol{\phi}) - M_n(\boldsymbol{\beta}) \to_P (\boldsymbol{\phi}-\boldsymbol{\beta})^T C(\boldsymbol{\phi}-\boldsymbol{\beta}) = \Delta(\boldsymbol{\phi})\ .$$

La condition (iii) est immédiate, du moment que l'on a supposé que C est définie positive.
Les trois conditions sont donc vérifiées. On applique alors le Théorème 4 : si $\widehat{\boldsymbol{\beta}}_n$ vérifie

$$\sup_{\mathbf{t}\in[0,L]} \left\{ \Lambda_n(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}), \mathbf{t}) - \Lambda_n(\boldsymbol{\beta}, \mathbf{t}) \right\}_+ \xrightarrow{P} 0 \tag{3.4}$$

alors $\widehat{\boldsymbol{\beta}}_n$ converge en probabilité vers $\boldsymbol{\beta}$, uniformément en $\mathbf{t}$ sur $T = [0, L]$, pour la probabilité P. On obtient la propriété :

$$\sup_{\mathbf{t}\in[0,L]} \|\widehat{\boldsymbol{\beta}}_n(\mathbf{t}) - \boldsymbol{\beta}\| \xrightarrow{P} 0 \tag{3.5}$$

Remarques :

- L'ensemble $T = [0, L]$ des valeurs prises par la constante de régularisation est à mettre en correspondance avec l'ensemble $[0, \lambda_{max}]$ des valeurs prises par λ qu'on trouve dans le chapitre 2.
- Définir $\widehat{\boldsymbol{\beta}}_n$ comme l'argument minimum de Λ_n est une condition plus forte que (3.4) et donc, en particulier, implique (3.4).

3.1.2 Théorème de la limite centrale pour le Lasso

On va maintenant appliquer le Théorème 6 de l'article en annexe 6.3 dans le cas du Lasso.

Consistence uniforme en probabilité : On vient de montrer (paragraphe 3.1.1) que le Lasso est uniformément consistent en $\mathbf{t}$.

Décomposition de la fonction de contraste pénalisée : On décompose $\Lambda_n(\boldsymbol{\phi}, \mathbf{t}) - \Lambda_n(\boldsymbol{\beta}, \mathbf{t})$. Grâce à la décomposition de $M_n(\boldsymbol{\phi}) - M_n(\boldsymbol{\beta})$ donnée par l'équation (3.3), on a :

$$\begin{aligned}
\Lambda_n(\boldsymbol{\phi}, \mathbf{t}) - \Lambda_n(\boldsymbol{\beta}, \mathbf{t}) &= (\boldsymbol{\phi} - \boldsymbol{\beta})^T C_n (\boldsymbol{\phi} - \boldsymbol{\beta}) - \frac{2}{n} \varepsilon_n^T \boldsymbol{X}_n (\boldsymbol{\phi} - \boldsymbol{\beta}) \\
&\quad + \mathbf{t} n^{-1/2} \left(\|\boldsymbol{\phi}\|_1 - \|\boldsymbol{\beta}\|_1 \right) \\
&= (\boldsymbol{\phi} - \boldsymbol{\beta})^T C (\boldsymbol{\phi} - \boldsymbol{\beta}) \\
&\quad + (\boldsymbol{\phi} - \boldsymbol{\beta})^T (C_n - C)(\boldsymbol{\phi} - \boldsymbol{\beta}) \\
&\quad - \frac{2}{n} \varepsilon_n^T \boldsymbol{X}_n (\boldsymbol{\phi} - \boldsymbol{\beta}) + \mathbf{t} n^{-1/2} \left(\|\boldsymbol{\phi}\|_1 - \|\boldsymbol{\beta}\|_1 \right)
\end{aligned}$$

On pose alors :
- $G_n(\boldsymbol{\phi}, \mathbf{t}) = -\frac{2}{n} \varepsilon_n^T \boldsymbol{X}_n (\boldsymbol{\phi} - \boldsymbol{\beta}) + \mathbf{t} n^{-1/2} \left(\|\boldsymbol{\phi}\|_1 - \|\boldsymbol{\beta}\|_1 \right)$
- $H(\boldsymbol{\phi}, \mathbf{t}) = (\boldsymbol{\phi} - \boldsymbol{\beta})^T C (\boldsymbol{\phi} - \boldsymbol{\beta})$

– $R_n(\boldsymbol{\phi}, \mathbf{t}) = \|\boldsymbol{\phi} - \boldsymbol{\beta}\|^{-1}(\boldsymbol{\phi} - \boldsymbol{\beta})^T(C_n - C)(\boldsymbol{\phi} - \boldsymbol{\beta})$

Condition sur la fonction G_n : $G_n(\boldsymbol{\phi}, \mathbf{t})$ comporte deux termes.

On note :

$$G_{n,1}(\boldsymbol{\phi}, \mathbf{t}) = -\frac{2}{n}\varepsilon_n^T \boldsymbol{X}_n(\boldsymbol{\phi} - \boldsymbol{\beta}) \text{ et } G_{n,2}(\boldsymbol{\phi}, \mathbf{t}) = \mathbf{t}\frac{1}{\sqrt{n}}\left(\|\boldsymbol{\phi}\|_1 - \|\boldsymbol{\beta}\|_1\right) .$$

Par Cauchy-Schwarz, on a :

$$n|G_{n,1}(\boldsymbol{\phi}, \mathbf{t})| \leq 2\|\varepsilon_n^T \boldsymbol{X}_n\|.\|(\boldsymbol{\phi} - \boldsymbol{\beta})\| .$$

On a déjà vu que $\|\varepsilon_n^T \boldsymbol{X}_n\| = O_{P^*}(\sqrt{n})$. Donc $\boldsymbol{X}_n^T\varepsilon_n = \sqrt{n}U_n$, avec $U_n = O_{P^*}(1)$ et on a : $n|G_{n,1}(\boldsymbol{\phi}, \mathbf{t})| \leq 2\sqrt{n}\|U_n\|\|\boldsymbol{\phi} - \boldsymbol{\beta}\|$.

D'autre part :

$$n|G_{n,2}(\boldsymbol{\phi}, \mathbf{t})| = \mathbf{t}\sqrt{n}\Big|\|\boldsymbol{\beta}\|_1 - \|\boldsymbol{\phi}\|_1\Big| \leq \mathbf{t}\sqrt{n}\|\boldsymbol{\phi} - \boldsymbol{\beta}\|_1 .$$

Et d'après l'équivalence des normes sur $\boldsymbol{\Phi}$:

$$\exists c > 0 \text{ tel que } \|\boldsymbol{\phi} - \boldsymbol{\beta}\|_1 \leq c\|\boldsymbol{\phi} - \boldsymbol{\beta}\| .$$

Donc $n|G_{n,2}(\boldsymbol{\phi}, \mathbf{t})| \leq c\mathbf{t}\sqrt{n}\|\boldsymbol{\phi} - \boldsymbol{\beta}\|$.

On a donc, comme $\mathrm{T} = [0, L]$:

$$\begin{aligned}
& \sup_{(\boldsymbol{\phi},\mathbf{t})\in(\boldsymbol{\Phi},T)} \frac{n|G_n(\boldsymbol{\phi}, \mathbf{t})|}{1 + \sqrt{n}\|\boldsymbol{\phi} - \boldsymbol{\beta}\|} \\
= & \sup_{(\boldsymbol{\phi},\mathbf{t})\in(\boldsymbol{\Phi},T)} \frac{n|G_{n,1}(\boldsymbol{\phi}, \mathbf{t}) + G_{n,2}(\boldsymbol{\phi}, \mathbf{t})|}{1 + \sqrt{n}\|\boldsymbol{\phi} - \boldsymbol{\beta}\|} \\
\leq & \sup_{(\boldsymbol{\phi},\mathbf{t})\in(\boldsymbol{\Phi},T)} \frac{n|G_{n,1}(\boldsymbol{\phi}, \mathbf{t})|}{1 + \sqrt{n}\|\boldsymbol{\phi} - \boldsymbol{\beta}\|} + \sup_{(\boldsymbol{\phi},\mathbf{t})\in(\boldsymbol{\Phi},T)} \frac{n|G_{n,2}(\boldsymbol{\phi}, \mathbf{t})|}{1 + \sqrt{n}\|\boldsymbol{\phi} - \boldsymbol{\beta}\|} \\
\leq & \sup_{(\boldsymbol{\phi},\mathbf{t})\in(\boldsymbol{\Phi},T)} 2\|U_n\|\frac{\sqrt{n}\|\boldsymbol{\phi} - \boldsymbol{\beta}\|}{1 + \sqrt{n}\|\boldsymbol{\phi} - \boldsymbol{\beta}\|} + \sup_{(\boldsymbol{\phi},\mathbf{t})\in(\boldsymbol{\Phi},T)} cL\frac{\sqrt{n}\|\boldsymbol{\phi} - \boldsymbol{\beta}\|}{1 + \sqrt{n}\|\boldsymbol{\phi} - \boldsymbol{\beta}\|} \\
= & O_P(1)
\end{aligned}$$

Comme $\frac{\sqrt{n}\|\boldsymbol{\phi}-\boldsymbol{\beta}\|}{1+\sqrt{n}\|\boldsymbol{\phi}-\boldsymbol{\beta}\|}$ est toujours borné par 1, donc on a bien que :

$$\sup_{(\boldsymbol{\phi},\mathbf{t})\in(\boldsymbol{\Phi},T)} \frac{n|G_n(\boldsymbol{\phi}, \mathbf{t})|}{1 + \sqrt{n}\|\boldsymbol{\phi} - \boldsymbol{\beta}\|} = O_P(1) .$$

Conditions sur la fonction H : Il suffit de poser $\Gamma(\mathbf{t}) = C$, pour tout $\mathbf{t} \in T$.

– Comme C est définie positive, toutes ses valeurs propres sont strictement positives.
– D'autre part, l'autre condition sur H est trivialement vérifiée car $H(\boldsymbol{\phi}, \mathbf{t}) = (\boldsymbol{\phi} - \boldsymbol{\beta})^T C(\boldsymbol{\phi} - \boldsymbol{\beta})$.

Condition sur la fonction R_n : $R_n(\boldsymbol{\phi}, \mathbf{t}) = \|\boldsymbol{\phi}-\boldsymbol{\beta}\|^{-1}(\boldsymbol{\phi}-\boldsymbol{\beta})^T(C_n-C)(\boldsymbol{\phi}-\boldsymbol{\beta})$.
Donc :

$$|R_n(\boldsymbol{\phi}, \mathbf{t})| \leq \rho(C_n - C) \, \|\boldsymbol{\phi} - \boldsymbol{\beta}\| \, ,$$

où $\rho(C_n - C)$ est le maximum des valeurs absolues des valeurs propres de $(C_n - C)$.
Comme $C_n \xrightarrow{P} C$, on a que $\rho(C_n - C) = o_P(1)$. Et donc il vient :

$$\sup \{R_n(\boldsymbol{\phi}, \mathbf{t}), \boldsymbol{\phi} \in \boldsymbol{\Phi}, \|\boldsymbol{\phi} - \boldsymbol{\beta}\| \leq r_n\} = o_P(r_n)$$

Convergence de la fonction $\widehat{G}_n$: De même que l'on a décomposé $G_n(\boldsymbol{\phi}, \mathbf{t})$ en $G_{n,1}(\boldsymbol{\phi}, \mathbf{t}) + G_{n,2}(\boldsymbol{\phi}, \mathbf{t})$, on définit :

$$\widehat{G}_{n,1}(\boldsymbol{\phi}, \mathbf{t}) = nG_{n,1}\left(\boldsymbol{\beta} + \frac{1}{\sqrt{n}}\boldsymbol{\phi}, \mathbf{t}\right) \text{ et } \widehat{G}_{n,2}(\boldsymbol{\phi}, \mathbf{t}) = nG_{n,2}\left(\boldsymbol{\beta} + \frac{1}{\sqrt{n}}\boldsymbol{\phi}, \mathbf{t}\right)$$

On a d'une part :

$$\widehat{G}_{n,1}(\boldsymbol{\phi}, \mathbf{t}) = -\frac{2\varepsilon_n^T \boldsymbol{X}_n}{\sqrt{n}}\boldsymbol{\phi} = -2U_n^T\boldsymbol{\phi} \, .$$

On se trouve dans le cas classique de la régression des moindres carrés où l'on sait, par le théorème de Lindeberg-Feller et sous la condition (KF–2), que U_n converge en loi vers U, de loi $\mathcal{N}(0, \sigma^2 C)$.
Soit K un compact de $\boldsymbol{\Phi}$. On considère la fonction f définie sur $\mathbb{R}^p$ à valeurs dans $\ell^\infty(K \times T)$ qui à tout élément u associe la fonction $(\boldsymbol{\phi}, \mathbf{t}) \mapsto u^T\boldsymbol{\phi}$. Comme f est continue, on en conlut que $\widehat{G}_{n,1}$ converge dans $\ell^\infty(K \times T)$ vers $G_1 = -2f(U)$.
D'autre part,

$$\begin{aligned}
\widehat{G}_{n,2}(\boldsymbol{\phi}, \mathbf{t}) &= \mathbf{t}\sqrt{n}\sum_{j=1}^{p}\left\{\left|\beta_j + \frac{\phi_j}{\sqrt{n}}\right| - |\beta_j|\right\} \\
&= \mathbf{t}\sqrt{n}\sum_{j=1}^{p}\left\{|\beta_j|\left(\left|1 + \frac{\phi_j}{\sqrt{n}\beta_j}\right| - 1\right)\mathbb{1}_{\beta_j \neq 0} + \left|\frac{\phi_j}{\sqrt{n}}\right|\mathbb{1}_{\beta_j = 0}\right\} \\
&= \mathbf{t}\sqrt{n}\sum_{j=1}^{p}\left\{|\beta_j|\frac{\phi_j}{\sqrt{n}\beta_j}\mathbb{1}_{\beta_j \neq 0} + \left|\frac{\phi_j}{\sqrt{n}}\right|\mathbb{1}_{\beta_j = 0}\right\}
\end{aligned}$$

La dernière égalité est correcte pour $\boldsymbol{\phi} \in K$ et pour n suffisamment grand. Donc, pour n suffisamment grand, on obtient :

$$\begin{aligned} \widehat{G}_{n,2}(\boldsymbol{\phi}, \mathbf{t}) &= \mathbf{t} \sum_{j=1}^{p} \left\{ \frac{|\beta_j|}{\beta_j} \phi_j \mathbb{1}_{\beta_j \neq 0} + \mathbf{t} |\phi_j| \mathbb{1}_{\beta_j = 0} \right\} \\ &= \mathbf{t} J_\infty(\boldsymbol{\phi}) \end{aligned} \tag{3.6}$$

qui est maintenant indépendante de n. Cette expression est donc la limite quand n tend vers ∞ de $\widehat{G}_{n,2}(\boldsymbol{\phi}, \mathbf{t})$.

Pour conclure, $\widehat{G}_n(\boldsymbol{\phi}, \mathbf{t}) = \widehat{G}_{n,1}(\boldsymbol{\phi}, \mathbf{t}) + \widehat{G}_{n,2}(\boldsymbol{\phi}, \mathbf{t})$ a bien une limite dans $\ell^\infty(K \times T)$.

On peut maintenant appliquer le Théorème 6. Supposons que la condition (3.4) est remplacée par la condition renforcée :

$$\sup_{\mathbf{t} \in [0,L]} \left\{ \Lambda_n(\widehat{\boldsymbol{\beta}}_n, \mathbf{t}) - \Lambda_n(\boldsymbol{\beta}, \mathbf{t}) \right\}_+ = o_P(1/n)$$

On montre facilement qu'il existe un processus $\{\widehat{\boldsymbol{u}}(\mathbf{t}), \mathbf{t} \in T\}$ qui minimise :

$$\mathbb{L}(\boldsymbol{\phi}, \mathbf{t}) = -2U^T \boldsymbol{\phi} + \boldsymbol{\phi}^T C \boldsymbol{\phi} + \mathbf{t} J_\infty(\boldsymbol{\phi})$$

en $\boldsymbol{\phi} \in \mathbb{R}^p$ et tel que les conditions (ii) et (iii) du Théorème 5 de l'annexe 6.3 soient vérifiées. On a donc, par application du Théorème 6 :

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \rightsquigarrow \widehat{\boldsymbol{u}}$$

avec $\widehat{\boldsymbol{u}}$ dans l'espace des fonctions bornées de $\mathrm{T} = [0, L]$ dans $\mathbb{R}^p$, noté $\ell^\infty(T, p)$.

Le résultat de Knight et Fu dans [29] est obtenu en appliquant cette convergence fonctionnelle en un point $\mathbf{t} \in [0, L]$ donné.

3.2 Fonctions de pénalisation

Dans cette partie, on montre que les fonctions de pénalisation de la forme

$$J_n(\boldsymbol{\phi}) = n^{-1/2} \sum_{j=1}^{p} |\phi_j|^\gamma = n^{-1/2} \|\boldsymbol{\phi}\|_\gamma^\gamma \tag{3.7}$$

vérifient l'hypothèse requise sur la pénalisation pour pouvoir appliquer le Théorème 7 de l'article cité en annexe 6.3.

On suppose dans la suite que $\gamma \geq 1$. Pour ces valeurs de γ, $\|.\|_\gamma$ est une norme. Il vient la propriété suivante :

Proposition 3.1 *Soient $\boldsymbol{\beta} \in \mathbb{R}^p$ et $\gamma \geq 1$. Il existe $c_{\gamma,\boldsymbol{\beta}}$, une constante ne dépendant que de γ et de $\boldsymbol{\beta}$, telle que :*

$$\forall \boldsymbol{\phi} : \left| \|\boldsymbol{\phi}\|_\gamma^\gamma - \|\boldsymbol{\beta}\|_\gamma^\gamma \right| \leq c_{\gamma,\boldsymbol{\beta}} \left(\|\boldsymbol{\phi} - \boldsymbol{\beta}\|_\gamma^\gamma + \|\boldsymbol{\phi} - \boldsymbol{\beta}\| \right) \tag{3.8}$$

Preuve : La propriété (3.8) est évidente pour $\gamma = 1$. Supposons que $\gamma > 1$. On a, pour $x \in \mathbb{R}_+^*$:

$$\frac{d}{dx} x^\gamma = \gamma x^{\gamma-1} \tag{3.9}$$

Donc il vient, pour tout $j \in \{1, ..., p\}$:

$$\begin{aligned} |\phi_j|^\gamma - |\beta_j|^\gamma| &\leq \gamma \big| |\phi_j| - |\beta_j| \big| \max \left(|\phi_j|^{\gamma-1}, |\beta_j|^{\gamma-1} \right) \\ &\leq \gamma |\phi_j - \boldsymbol{\beta}_j| \left(|\phi_j|^{\gamma-1} + |\beta_j|^{\gamma-1} \right) \end{aligned}$$

par majoration du maximum par la somme. D'autre part, on a la propriété suivante :

$$\forall a, b, p > 0 : (a+b)^p \leq c_p \left(a^p + b^p \right)$$

où c_p est une constante. Cette propriété s'obtient par convexité (respectivement par concavité) de la fonction $x \mapsto x^p$ quand $p \geq 1$ (resp. $p < 1$). On l'applique à $|(\phi_j - \beta_j) + \beta_j|^{\gamma-1}$. On a alors :

$$|\phi_j|^\gamma - |\beta_j|^\gamma| \leq c_\gamma^T |\phi_j - \boldsymbol{\beta}_j| \left(|\phi_j - \beta_j|^{\gamma-1} + |\beta_j|^{\gamma-1} \right)$$

Puis, par sommation :

$$\begin{aligned} \left| \|\boldsymbol{\phi}\|_\gamma^\gamma - \|\boldsymbol{\beta}\|_\gamma^\gamma \right| &\leq c_\gamma^T \left(\sum_{j=1}^p |\phi_j - \beta_j|^\gamma + \sum_{j=1}^p |\phi_j - \beta_j| |\beta_j|^{\gamma-1} \right) \\ &\leq c_{\gamma,\boldsymbol{\beta}} \left(\|\boldsymbol{\phi} - \boldsymbol{\beta}\|_\gamma^\gamma + \|\boldsymbol{\phi} - \boldsymbol{\beta}\|_1 \right) \end{aligned}$$

Et comme les normes sont équivalentes sur $\mathbb{R}^p$, on obtient finalement (3.8). □

On vérifie maintenant aisément que la condition :

$$n \ |J_n(\boldsymbol{\phi}) - J_n(\boldsymbol{\beta})| \leq C \left(1 + \sqrt{n} \, \|\boldsymbol{\phi} - \boldsymbol{\beta}\| \right) \quad \text{for} \quad \|\boldsymbol{\phi} - \boldsymbol{\beta}\| \leq 1 \ , \tag{3.10}$$

est vérifiée pour $J_n(\boldsymbol{\phi}) = n^{-1/2} \|\boldsymbol{\phi}\|_\gamma^\gamma$. En effet :

$$\begin{aligned} n \ |J_n(\boldsymbol{\phi}) - J_n(\boldsymbol{\beta})| &= \sqrt{n} \ \left| \|\boldsymbol{\phi}\|_\gamma^\gamma - \|\boldsymbol{\beta}\|_\gamma^\gamma \right| \\ &\leq \sqrt{n} c_{\gamma,\boldsymbol{\beta}} \left(\|\boldsymbol{\phi} - \boldsymbol{\beta}\|_\gamma^\gamma + \|\boldsymbol{\phi} - \boldsymbol{\beta}\| \right) \ . \end{aligned}$$

Si $\|\boldsymbol{\phi} - \boldsymbol{\beta}\| \leq 1$, on a :

$$n \ |J_n(\boldsymbol{\phi}) - J_n(\boldsymbol{\beta})| \leq \sqrt{n} c_{\gamma,\boldsymbol{\beta}} \left(\|\boldsymbol{\phi} - \boldsymbol{\beta}\| \right) \ .$$

Donc il vient que la quantité :

$$\frac{n\ |J_n(\boldsymbol{\phi}) - J_n(\boldsymbol{\beta})|}{1 + \sqrt{n}\,\|\boldsymbol{\phi} - \boldsymbol{\beta}\|}$$

est bien majorée par une constante.

On peut également donner l'expression de $J_\infty(\boldsymbol{\phi})$ qui est définie dans le Théorème 4 de l'article de l'annexe 6.3.

$$\forall K \text{ compact de } \mathbb{R}^p : \quad \sup_{\boldsymbol{\phi}\in K} \left| n\ J_n(\boldsymbol{\beta} + n^{-1/2}\boldsymbol{\phi}) - n\ J_n(\boldsymbol{\beta}) - J_\infty(\boldsymbol{\phi}) \right| \to 0\ . \tag{3.11}$$

Pour $\gamma = 1$, $J_\infty(\boldsymbol{\phi})$ a déjà été calculé en (3.6). On suppose maintenant que $\gamma > 1$. Soit K un compact de $\mathbb{R}^p$:

$$\begin{aligned}
& n\ J_n(\boldsymbol{\beta} + n^{-1/2}\boldsymbol{\phi}) - n\ J_n(\boldsymbol{\beta}) \\
= \ & \sqrt{n}\ \|\boldsymbol{\beta} + n^{-1/2}\boldsymbol{\phi}\|_\gamma^\gamma - \|\boldsymbol{\beta}\|_\gamma^\gamma \\
= \ & \sqrt{n}\ \sum_{j=1}^p \left|\beta_j + n^{-1/2}\phi_j\right|^\gamma - |\beta_j|^\gamma \\
= \ & n^{\frac{1-\gamma}{2}}\ \sum_{j=1}^p |\phi_j|^\gamma\ \mathbb{1}_{\beta_j=0} + \sqrt{n}\sum_{j=1}^p \left(\left|\beta_j + n^{-1/2}\phi_j\right|^\gamma - |\beta_j|^\gamma\right) \mathbb{1}_{\beta_j\neq 0}
\end{aligned}$$

D'une part le premier terme de la somme tend vers 0 sur tout compact de $\mathbb{R}$, car $\gamma > 1$. D'autre part, d'après (3.9), on a l'équivalent suivant, quand $n \to \infty$, pour $\beta_j \neq 0$:

$$\left|\beta_j + n^{-1/2}\phi_j\right|^\gamma - |\beta_j|^\gamma \sim \operatorname{sgn}\beta_j \gamma |\beta_j|^{\gamma-1} n^{-1/2}\phi_j$$

Et donc :

$$J_\infty(\boldsymbol{\phi}) = \gamma \sum_{j=1}^p \operatorname{sgn}\beta_j |\beta_j|^{\gamma-1}\phi_j \mathbb{1}_{\beta_j\neq 0}.$$

De plus, la convergence est bien uniforme pour $\boldsymbol{\phi}$ dans un compact.

3.3 Généralisations

3.3.1 Application aux modèles exponentiels

Comme pour le Lasso (voir section 3.1), nous proposons d'appliquer les résultats en termes de consistence uniforme et de théorème de la limite centrale au cas des modèles exponentiels.

On suppose dans cette section que $\boldsymbol{Y}$ suit une loi de la famille exponentielle canonique, i.e. $\boldsymbol{Y}$ admet pour densité :

$$p(y|\theta) = h(y)\exp\{y\theta - b(\theta)\}$$

par rapport à une mesure de domination μ.
$b(\theta)$ est la fonction de log-répartition. Elle vérifie

$$b(\theta) = \log \int h(y)\exp\{y\theta\}\mu(y)dy \ .$$

$b(\theta)$ est donc strictement convexe et C^∞. La fonction de contraste non–pénalisée s'écrit donc, à une constante additive près, indépendante de $\boldsymbol{\phi}$:

$$M_n(\boldsymbol{\phi}) = n^{-1}\sum_{k=1}^{n} g((\boldsymbol{x}_k, y_k), \boldsymbol{\phi})$$

où $g((\boldsymbol{x}_k, y_k), \boldsymbol{\phi}) = -y_k\boldsymbol{x}_k^T\boldsymbol{\phi} + b(\boldsymbol{x}_k^T\boldsymbol{\phi})$.

Le modèle de régression généralisé étudié dans [36] consiste à observer un $(y_k, \boldsymbol{x}_k)$, i.i.d. à valeurs dans $\mathbb{R}\times\mathbb{R}^p$, avec $f_{y_k|\boldsymbol{x}_k} = p(y|\boldsymbol{x}^T\boldsymbol{\beta})$.

Nous allons obtenir un comportement asymptotique similaire à celui obtenu au paragraphe 3.1 pour le Lasso sous des hypothèses similaires à (KF–1) et (KF–2) pour des $(\boldsymbol{x}_k)$ i.i.d., c'est-à-dire pour $\mathbb{E}[\boldsymbol{x}_k] = 0$ et $\mathrm{Cov}(\boldsymbol{x}_k) = C$, avec C définie positive.

On applique là encore le Théorème 4 de l'article cité en annexe 6.3 pour montrer la consistence uniforme de l'estimateur pénalisé.

– g est une fonction convexe car b est strictement convexe ;
– $\forall\boldsymbol{\phi}\in\boldsymbol{\Phi}, M_n(\boldsymbol{\phi}) \to_P M(\boldsymbol{\phi})$ avec

$$M(\boldsymbol{\phi}) = \lim_{n\to\infty} n^{-1}\sum_{k=1}^{n} g((\boldsymbol{x}_k, y_k), \boldsymbol{\phi}) = \mathbb{E}\left[g((\boldsymbol{x}, y), \boldsymbol{\phi})\right]$$

où $(\boldsymbol{x}, y)$ a la même distribution que les $(\boldsymbol{x}_k, y_k)$. $M_n(\boldsymbol{\phi})$ est strictement convexe dès que $\boldsymbol{X}_n^T\boldsymbol{X}_n$ est inversible. Comme $n^{-1}\boldsymbol{X}_n^T\boldsymbol{X}_n \to C$ alors $n^{-1}\det(\boldsymbol{X}_n^T\boldsymbol{X}_n) \to \det(C)$ Ceci arrive presque sûrement quand n est suffisamment grand ;

– $\forall\boldsymbol{\phi}\in\boldsymbol{\Phi}$, $M(\boldsymbol{\phi}) \geq M(\boldsymbol{\beta})$. En effet :

$$M(\boldsymbol{\phi}) - M(\boldsymbol{\beta}) = -\mathbb{E}\left[\mathbb{E}\left[\log\frac{p(y|\boldsymbol{x}^T\boldsymbol{\phi})}{p(y|\boldsymbol{x}^T\boldsymbol{\beta})}|\boldsymbol{x}\right]\right]$$

C'est donc l'opposé de l'espérance d'une distance de Kullback-Leibler.

Pour montrer le théorème de la limite centrale, on vérifie les conditions (P-1)-(P-4) de Pollard afin d'appliquer le Théorème 7 de l'article en annexe 6.3.
g admet autour de $\boldsymbol{\beta}$ un développement de Taylor de la forme :

$$g((\boldsymbol{x}_k, y_k), \boldsymbol{\phi}) = g((\boldsymbol{x}_k, y_k), \boldsymbol{\beta}) + (\boldsymbol{\phi} - \boldsymbol{\beta})^T \, \Delta((\boldsymbol{x}_k, y_k)) + |\boldsymbol{\phi} - \boldsymbol{\beta}| \, r((\boldsymbol{x}_k, y_k), \boldsymbol{\phi})$$

où $\Delta((\boldsymbol{x}_k, y_k)) = \boldsymbol{x}_k(-y_k + g'(\boldsymbol{x}_k^T \boldsymbol{\beta}))$ et $r((\boldsymbol{x}_k, y_k), \boldsymbol{\phi}) = |\boldsymbol{\phi} - \boldsymbol{\beta}| \, \frac{b''(\boldsymbol{x}_k^T \boldsymbol{\phi}_0)}{2}$, avec $\boldsymbol{\phi}_0$ s'écrivant sous la forme $\boldsymbol{\beta} + t_0 \, (\boldsymbol{\phi} - \boldsymbol{\beta})$ avec $0 < t_0 < 1$. Le gradient et la hessienne de $M_n(\boldsymbol{\phi})$ prennent les expressions suivantes :

$$\nabla M_n(\boldsymbol{\beta})^T = \sum_{k=1}^{n} \Delta((\boldsymbol{x}_k, y_k)) \tag{3.12}$$

$$\nabla^2 M_n(\boldsymbol{\beta}) = n^{-1} \, \sum_{k=1}^{n} g''(\boldsymbol{x}_k^T \boldsymbol{\beta}) \rightarrow_{n \rightarrow \infty} \Gamma(\boldsymbol{\beta}) \tag{3.13}$$

- $M(\boldsymbol{\phi})$ possède une hessienne définie positive en $\boldsymbol{\beta}$, valeur pour laquelle M atteint son minimum.
- Δ est d'espérance nulle et de covariance finie (propriétés du score pour la famille exponentielle).
- On a la propriété de différentiabilité stochastique qu'on rappelle ici :

$$\sup_{V_n} \frac{|\nu_n r(., \boldsymbol{\phi})|}{1 + \sqrt{n}|\boldsymbol{\phi} - \boldsymbol{\beta}|} \xrightarrow{P} 0 \tag{3.14}$$

  où V_n est l'ensemble des suites qui convergent vers $\boldsymbol{\beta}$. En effet, on rappelle qu'ici :
  $r((\boldsymbol{x}_k, y_k), \boldsymbol{\phi}) = |\boldsymbol{\phi} - \boldsymbol{\beta}| \, \frac{b''(\boldsymbol{x}_k^T \boldsymbol{\phi}_0)}{2}$. Comme on regarde des voisinages de $\boldsymbol{\beta}$, on peut considérer qu'à partir d'un certain rang, $\boldsymbol{\phi}_0$ appartient à un compact de $\mathbb{R}$. On peut donc borner b'' sur ce compact. Il vient donc (3.14) (voir aussi [38, Section 4]).

3.3.2 Autres applications

Les applications du Théorème 7 de notre article sont nombreuses. En effet, si l'on prend les exemples de M-estimateurs donnés par Pollard dans [38], les résultats de consistence de l'estimateur comme fonction du paramètre de régularisation et l'existence d'un théorème de la limite centrale sont encore vrais si l'on ajoute une fonction de pénalisation, sous des hypothèses peu contraignantes sur la pénalisation. Ainsi on a ces résultats asymptotiques pour la régression Ridge, mais

également pour des fonctions de constraste non-dérivables, comme le LAD (Least Absolute Deviation) qui est un critère basé sur la valeur absolue des écarts. Ceci est une conséquence de la propriété de différentiabilité stochastique montrée dans [38, Exemple 8]. Nos travaux donnent par exemple des résultats asymptotiques similaires pour le LAD dans sa version pénalisé L_2.

3.4 Application au test d'hypothèse

Dans cette partie, on montre une application simple que peuvent apporter les résultats asymptotiques présentés dans l'article en annexe 6.3. Pour tester l'hypothèse nulle $H_0 : \{\boldsymbol{\beta} = 0\}$, on propose la statistique S_n suivante.

$$S_n = \inf_{\mathbf{t}\in[0,L]} \|\boldsymbol{X}_n \widehat{\boldsymbol{\beta}}(\mathbf{t})\|^2$$

Les résultats du Théorème 2 de l'article nous donne la convergence en loi de cette statistique. En effet, on a :

$$S_n = \inf_{\mathbf{t}\in[0,L]} \sqrt{n}\widehat{\boldsymbol{\beta}}(\mathbf{t}) \frac{\boldsymbol{X}_n^T \boldsymbol{X}_n}{n} \sqrt{n}\widehat{\boldsymbol{\beta}}(\mathbf{t}) \rightsquigarrow \inf_{\mathbf{t}\in[0,L]} \widehat{\boldsymbol{u}}(\mathbf{t})^T C \widehat{\boldsymbol{u}}(\mathbf{t}) = S_\infty$$

où $\widehat{\boldsymbol{u}}(\mathbf{t})$ minimise la fonctionnelle

$$\mathbb{L}(\phi, \mathbf{t}) = -2U^T C + \phi^T C \phi + \mathbf{t} \sum_{j=1}^{p} |\phi_j| \tag{3.15}$$

qui est l'expression du contraste pénalisé limite sous H_0.

On note qu'en pratique, l'algorithme LARS permet de calculer facilement S_n, car le chemin de régularisation est affine par morceaux. Il est de même aisé d'obtenir des simulations de S_∞ sous H_0 : on simule $U \sim \mathcal{N}(0, \sigma^2 C)$, on calcule, là encore à l'aide de la méthode LARS, le chemin de régularisation $\widehat{\boldsymbol{u}}(\mathbf{t})$ qui minimise (3.15) et on obtient ainsi la valeur de S_∞ correspondante. On peut donc facilement calculer des approximations de p-valeurs asymptotiques pour la statistique S_n.

Pour tester les performances de la statistique de test S_n, on a construit des courbes ROC à partir de données simulées. Nous avons pris $n = 30$ essais, $p = 20$ variables explicatives et nous avons simulé un modèle linéaire sous H_0 et sous H_1. Nous avons considéré que sous H_1, les coefficients de $\boldsymbol{\beta}$ sont uniformément distribués dans $[-1, 1]$. Les vecteurs de régression sont pris gaussiens de variance unité. Nous avons considéré les deux types de bruit additif au modèle linéaire suivants :

– un bruit gaussien de loi $\mathcal{N}(0, 4)$

– un mélange de bruits gaussiens, de loi résultante $0.5\,\mathcal{N}\,(0, 0.8)+0.5\,\mathcal{N}\,(0, 7.2)$

On compare les courbes ROC de S_n à celles obtenues avec la F-statistique suivante :

$$F_n = \frac{(n-p)\left\|\boldsymbol{X}_n\widehat{\boldsymbol{\beta}}(0)\right\|^2}{p\left\|\boldsymbol{Y}_n - \boldsymbol{X}_n\widehat{\boldsymbol{\beta}}(0)\right\|^2}$$

où $\boldsymbol{Y}_n = [y_1 \dots y_n]^T$ est construit suivant le même modèle. Cette statistique est due à Fisher et est classiquement utilisée par tester cette hypothèse, par exemple dans la méthode d'analyse de la variance (ANOVA). Les résultats de la figure 3.1 montrent que les performances de S_n sont meilleures que celles de F_n.

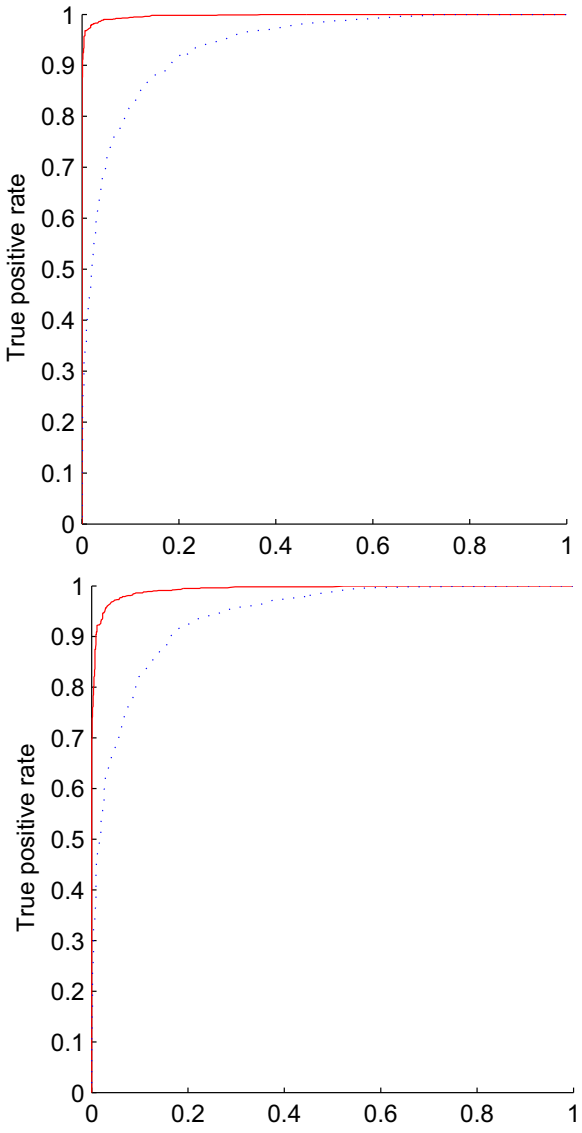

FIG. 3.1 – Courbes ROC de S_n en trait plein rouge et F_n en trait pointillé bleu. 1000 simultations de Monte-Carlo ont été tirées sous H_0 et sous H_1. En haut : cas d'un bruit gaussien. En bas : cas d'un mélange de bruits gaussiens

Chapitre 4

Objectivation multi-prestations

4.1 Le contexte de la multi-prestations

4.1.1 Définition de la multi-prestations

Comme vu précédemment dans la partie 2.1, le déploiement des prestations passe par la déclinaison de divers cahiers des charges. Le but est de cibler quelles sont les pièces et plus précisément encore quelles sont les caractéristiques techniques de celles-ci qui influent sur le ressenti final du client. C'est le rôle de l'objectivation : parvenir à mettre au point les pièces et les organes pour satisfaire au mieux le client.

Cependant, la mise au point d'une pièce ou d'un organe peut influer sur plus d'une prestation. Et pour compliquer les choses, deux prestations peuvent être du ressort de différents domaines d'ingénierie. Actuellement, l'objectivation est réalisée prestation par prestation. Il peut ainsi arriver que deux objectivations séparées conduisent à des préconisations contradictoires pour un même organe. Il faut dans ce cas pouvoir trouver un compromis qui satisfait au mieux le client. On parle alors d'objectivation *multi-prestations*.

4.1.2 Exemple de multi-prestations

Prenons un exemple concret pour mettre en évidence la nécessité, dans certains cas, d'objectiver simutanément plusieurs prestations.

Pour l'exemple, nous restons dans le périmètre de l'agrément des commandes. On considère la pédale d'embrayage et plus précisément, la mise au point de sa longueur de course [1] . La mise au point d'un tel organe est un cas typique de multi-

[1] La longueur de course est le trajet entre les positions « pédale levée » et « pédale totalement enfoncée ».

prestations car elle affecte différents métiers. En effet, les deux cahiers des charges suivants doivent être respectés.

Agrément des commandes Le métier agrément des commandes s'assure que les commandes sont faciles à manipuler. En ce qui concerne la pédale d'embrayage, l'action à réaliser lors d'un débrayage (*i.e.* enfoncer la pédale) demande un effort important. En effet, le conducteur doit appuyer suffisamment sur la pédale pour permettre au disque d'embrayage de se désolidariser du volant moteur. Un moyen efficace de diminuer cet effort est de le démultiplier en allongeant la course de la pédale. C'est pourquoi les experts agrément préconisent une course de pédale d'embrayage plutôt longue.

Ergonomie Le métier ergonomie s'assure que l'atteinte et la manipulation des commandes n'engendre pas de gêne chez le conducteur. Les courses des deux autres pédales (accélérateur et frein) sont en général beaucoup plus courtes, car les mécanismes mis en jeu nécessitent de moindres efforts. Plus la course de la pédale d'embrayage est longue, plus le conducteur doit lever la jambe, ce qui est une source d'inconfort. Il faut aussi éviter que le genou du conducteur vienne heurter le bas du volant. C'est pourquoi les experts agrément préconisent, eux, une course de pédale d'embrayage plutôt courte.

Si les objectivations sont réalisées séparément, alors l'objectivation de la prestation agrément des commandes fournit une grande valeur pour la course de la pédale d'embrayage tandis que l'objectivation de la prestation ergonomie fournit une petite valeur pour la même course. Il y a donc un compromis à trouver entre les deux métiers. Actuellement, dans la pratique, c'est au chef de projet véhicule qu'il incombe de trancher, après que les experts de chaque prestation ont expliqué leur point de vue et leur préconisation. Nous proposons ici de fournir les informations suffisantes pour que ce compromis ne soit plus l'issue de négociations mais qu'il soit optimisé du point de vue du client.

Pour cela, on suppose qu'il est possible de recueillir le ressenti global du client (sur la manipulation de la pédale d'embrayage par exemple) ainsi que son ressenti partiel sur chacune des prestations (ergonomie, agrément des commandes). On rappelle que ce n'est pas toujours le cas notamment dans le cas de prestations complexes. On pourra alors se contenter d'évaluations subjectives émanant d'experts des différentes prestations, plutôt que des évaluations client.

4.2 Formalisation mathématique du problème

Soit $\boldsymbol{Y}^G$ la variable à expliquer - qu'on appelle aussi variable réponse. $\boldsymbol{Y}^G = \left(y_i^G\right)_{i=1\ldots n}$ est le vecteur des notes subjectives globales : $y_i^G \in \{0,1\}, \forall i \in [1,n]$.

On considère maintenant un niveau intermédiaire. On appelle $\boldsymbol{Y}^1, ..., \boldsymbol{Y}^Q$ les Q prestations intermédiaires pour lesquelles on dispose également de notes subjectives.
On appellera dans la suite notes *partielles* les quantités : $\boldsymbol{Y}^q = (y_i^q)_{i=1...n}$ avec $y_i^q \in \{0,1\}, \forall i \in [1,n]$. On notera par $\boldsymbol{Y} = \left(\boldsymbol{Y}^1, ..., \boldsymbol{Y}^Q\right) \in \mathcal{M}_{n\times Q}$ le vecteur des notes partielles.
Le but final est d'expliquer la note globale $\boldsymbol{Y}^G$ à l'aide d'un ensemble de p variables explicatives mesurées : $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_p$.
On note $\boldsymbol{X}$ la matrice d'expériences. $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_p] \in \mathcal{M}_{n\times p}$.

Deux approches sont envisageables pour faire le lien entre note globale et critères physiques :

Approche directe On explique directement la note globale par les critères physiques. C'est l'approche intuitive pour modéliser la note globale.

Approche hiérarchique Le lien entre la note globale et les critères physiques tient compte des prestations intermédiaires. On propose d'objectiver séparément les différentes prestations intermédiaires puis de modéliser la note globale à l'aide des notes partielles.

Ces deux approches sont décrites plus précisément ci-dessous.

4.3 Approche directe

On se propose dans cette partie de décrire comment on explique directement la prestation globale à partir des variables explicatives. La méthode utilisée est la même que présentée précédemment dans le cas mono-prestation.

On conserve le même cadre que dans la partie 2, à savoir la régression logistique pénalisée L_1. Pour ce qui est des notations, on notera par $\boldsymbol{Y}_i^G$ le $i^{\text{ème}}$ tirage de la variable aléatoire $\boldsymbol{Y}^G$. De même, on confondra la variable aléatoire $\boldsymbol{X}_i$ et la ligne de la matrice $\boldsymbol{x}_i^T$. La vraisemblance s'écrit alors :

$$L_n(\boldsymbol{\beta}) = \exp\left\{\sum_{i=1}^{n} y_i^G \boldsymbol{x}_i^T\boldsymbol{\beta} - \log\left(1 + e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)\right\}$$

Le but est donc d'estimer $\widehat{\boldsymbol{\beta}}_{AD}$ (AD pour Approche Directe), le vecteur de paramètres qui maximise cette vraisemblance.

$$\widehat{\boldsymbol{\beta}}_{AD} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p}\left(L_n(\boldsymbol{\beta})\right)$$

Le modèle direct s'écrit alors de la façon suivante :

$$\mathbb{P}\left(\boldsymbol{Y}_i^G = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i\right) = \frac{e^{\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_{AD}}}{1 + e^{\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_{AD}}}$$

On voit que l'information contenue dans les prestations intermédiaires n'intervient pas dans l'approche directe. Voyons maintenant comment tirer profit de cette information supplémentaire.

4.4 Approche hiérarchique

Dans l'approche hiérarchique, nous prenons en compte l'information contenue dans les notes partielles.

4.4.1 Description de l'approche hiérarchique

Nous allons procéder en deux temps, d'où le terme d'approche *hiérarchique*. Tout d'abord, on estime les modèles qui expliquent chacune des notes partielles $\boldsymbol{Y}^q$,$q \in \{1, ..., Q\}$ en fonction des variables explicatives $\boldsymbol{X}_1, ..., \boldsymbol{X}_p$. Ainsi dans cette étape, on réalise Q objectivations mono-prestation. On s'intéresse dans un deuxième temps au lien entre la note globale $\boldsymbol{Y}^G$ et les notes partielles. Là encore, on réalise une objectivation mono-prestation, à ceci près que les variables explicatives sont les notes partielles et non plus les critères physiques.

De la même manière que précédemment, on cherche $\boldsymbol{\beta}$ qui maximise la vraisemblance des données complètes. La différence est que l'on a maintenant à notre disposition un niveau d'information supplémentaire, constitué par l'information sur les notes partielles. On verra dans la partie 4.5 si la prise en compte de cette information améliore ou non la modélisation obtenue.

On pose $\boldsymbol{Y} = \left(\boldsymbol{Y}^1, ..., \boldsymbol{Y}^Q\right)$. La vraisemblance des données complètes s'écrit comme précédemment :

$$\begin{aligned} L_n &= p\left(\boldsymbol{Y}_1^G = y_1^G, ..., \boldsymbol{Y}_n^G = y_n^G | \boldsymbol{X}_1 = \boldsymbol{x}_1, ..., \boldsymbol{X}_n = \boldsymbol{x}_n\right) \\ &= \prod_{i=1}^{n} p\left(\boldsymbol{Y}_i^G = y_i^G | \boldsymbol{X}_i = \boldsymbol{x}_i\right) \\ &= \prod_{i=1}^{n} \sum_{\varepsilon \in \{0,1\}^Q} p\left(\boldsymbol{Y}_i^G = y_i^G | \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{Y}_i = \varepsilon\right) p\left(\boldsymbol{Y}_i = \varepsilon | \boldsymbol{X}_i = \boldsymbol{x}_i\right) \end{aligned}$$

car : $\sum_{\varepsilon \in \{0,1\}^Q} p\left(\boldsymbol{Y}_i = \varepsilon\right) = 1, \forall i \in \{1, ..., n\}$.

Le calcul de la vraisemblance n'est pas direct. On peut simplifier ce calcul en posant les hypothèses suivantes.

4.4.2 Hypothèses

Dans la suite de la méthologie de traitement de la multi-prestations, on supposera que :

Hypothèse 4.1 *Conditionnellement aux notes partielles, la note globale est indépendante des variables explicatives, et donc :*

$$p\left(\boldsymbol{Y}_i^G = y_i^G | \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{Y}_i = \varepsilon\right) = p\left(\boldsymbol{Y}_i^G = y_i^G | \boldsymbol{Y}_i = \varepsilon\right) \tag{4.1}$$

Hypothèse 4.2 *Les prestations intermédiaires sont indépendantes entre elles, conditonnellement aux variables explicatives $\boldsymbol{X}_j, j \in \{1, ..., p\}$.*

$$p\left(\boldsymbol{Y}_i = \varepsilon | \boldsymbol{X}_i = \boldsymbol{x}_i\right) = \prod_{q=1}^{Q} p\left(\boldsymbol{Y}_i^q = \varepsilon^q | \boldsymbol{X}_i = \boldsymbol{x}_i\right) \tag{4.2}$$

Ces hypothèses permettent de simplifier le calcul de la vraisemblance. (Le paragraphe 4.4.4 revient sur la signification de ces deux hypothèses. On verra que ce sont des hypothèses raisonnables.)

4.4.3 Calcul de la vraisemblance dans le cas hiérarchique

Sous les deux hypothèses (4.1) et (4.2), l'expression de la vraisemblance devient :

$$\begin{aligned} L_n &= \prod_{i=1}^{n} \sum_{\varepsilon \in \{0,1\}^Q} p\left(\boldsymbol{Y}_i^G = y_i^G | \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{Y}_i = \varepsilon\right) p\left(\boldsymbol{Y}_i = \varepsilon | \boldsymbol{X}_i = \boldsymbol{x}_i\right) \\ &= \prod_{i=1}^{n} \sum_{\varepsilon \in \{0,1\}^Q} p\left(\boldsymbol{Y}_i^G = y_i^G | \boldsymbol{Y}_i = \varepsilon\right) \prod_{q=1}^{Q} p\left(\boldsymbol{Y}_i^q = \varepsilon^q | \boldsymbol{X}_i = \boldsymbol{x}_i\right) \end{aligned} \tag{4.3}$$

– Pour calculer la vraisemblance, on commence d'abord par objectiver chacune des prestations.

$$\widehat{\boldsymbol{\beta}}_{AH}^q = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \exp\left\{ \sum_{i=1}^{n} y_i^q \boldsymbol{x}_i^T \boldsymbol{\beta} - \log\left(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right) \right\}$$

qui correpond à la densité :

$$\prod_{q=1}^{Q} p\left(\boldsymbol{Y}_i^q = \varepsilon^q | \boldsymbol{X}_i = \boldsymbol{x}_i\right) = \prod_{q=1}^{Q} \left(\frac{e^{\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_{AH}^q}}{1 + e^{\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_{AH}^q}} \right)^{\varepsilon^q} \left(\frac{1}{1 + e^{\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_{AH}^q}} \right)^{(1-\varepsilon^q)}$$

– On objective maintenant la note globale à partir des notes partielles. C'est là que réside l'originalité de l'approche hiérarchique.

$$\widehat{\boldsymbol{\beta}}_{AH}^{G} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^{Q}} \exp\left\{\sum_{i=1}^{n} y_i^G \widetilde{\boldsymbol{Y}}_i^T \boldsymbol{\beta} - \log\left(1 + e^{\widetilde{\boldsymbol{Y}}_i^T \boldsymbol{\beta}}\right)\right\}$$

où $\widetilde{\boldsymbol{Y}} = 2\boldsymbol{Y} - 1$ de sorte que $y_i^q \in \{-1, 1\}$ et non $y_i^q \in \{0, 1\}$. [2]
Cette vraisemblance correspond à la densité :

$$\prod_{i=1}^{n} p\left(\boldsymbol{Y}_i^G = y_i^G | \widetilde{\boldsymbol{Y}}_i = \varepsilon\right) = \prod_{i=1}^{n} \left(\frac{e^{\varepsilon^T \widehat{\boldsymbol{\beta}}_{AH}^G}}{1 + e^{\varepsilon^T \widehat{\boldsymbol{\beta}}_{AH}^G}}\right)^{y_i^G} \left(\frac{1}{1 + e^{\varepsilon^T \widehat{\boldsymbol{\beta}}_{AH}^G}}\right)^{(1-y_i^G)}$$

$\widehat{\boldsymbol{\beta}}_{AH}$ est le jeu de coefficients qui maximisent la vraisemblance dans le cas du modèle hiérarchique.

$$\widehat{\boldsymbol{\beta}}_{AH} = \left\{\widehat{\boldsymbol{\beta}}_{AH}^{G}, \widehat{\boldsymbol{\beta}}_{AH}^{1}, ..., \widehat{\boldsymbol{\beta}}_{AH}^{Q}\right\}$$

où $\widehat{\boldsymbol{\beta}}_{AH}^{G} \in \mathbb{R}^{Q}$ et $\widehat{\boldsymbol{\beta}}_{AH}^{q} \in \mathbb{R}^{p}$, $\forall q \in \{1, ..., Q\}$.

Cette approche hiérarchique a un double intérêt. En effet, les ingénieurs ont accès à un modèle pour chaque prestation intermédiaire - ce qui est une information en soi - mais ils disposent également d'une modélisation de la note globale par les notes partielles. Ce modèle permet de prioriser les prestations intermédiaires entre elles. Cette priorisation est même quantifiée (comparaison des composantes du vecteur $\widehat{\boldsymbol{\beta}}_{AH}^{G}$).

[2] L'intérêt de ce recodage est de symétriser le rôle de chacune des deux classes. En effet, les variables explicatives sont ici binaires. Supposons que l'on ait une seule variable explicative $\boldsymbol{Y}$. Si l'on code l'appartenance de cette variable à l'une des classes par 0 et l'appartenance à l'autre classe par 1, alors

$$\mathbb{P}\left(\boldsymbol{Y}_i^G = 1 | \boldsymbol{Y}_i = 0\right) = \mathbb{P}\left(\boldsymbol{Y}_i^G = 1 | \boldsymbol{Y}_i = 0\right) = 0.5$$

et

$$\mathbb{P}\left(\boldsymbol{Y}_i^G = 0 | \boldsymbol{Y}_i = 1\right) = \frac{1}{1 + e^{\boldsymbol{Y}_i^T \boldsymbol{\beta}^G}}, \mathbb{P}\left(\boldsymbol{Y}_i^G = 1 | \boldsymbol{Y}_i = 1\right) = \frac{e^{\boldsymbol{Y}_i^T \boldsymbol{\beta}^G}}{1 + e^{\boldsymbol{Y}_i^T \boldsymbol{\beta}^G}}$$

Les deux classes ne jouent donc pas le même rôle. Le codage par 0 donne à la classe correspondante la signification que la variable explicative n'a aucune influence sur la réponse. Ce codage conviendrait plutôt à la notion de classe « médiocre » ou « moyenne », dans le cas d'une variable pouvant prendre trois valeurs : {-1, 0, 1}. C'est pourquoi on choisit de recoder en {-1, 1}. Dans le cas où un intercept est pris en compte, ce "recentrage" est bien sûr inutile.

4.4.4 Discussion des hypothèses

On vient de voir que, pour pouvoir utiliser cette approche hiérarchique, deux hypothèses doivent être vérifiées. On se propose dans ce paragraphe de revenir sur leur signification concrète.

Hypothèse (4.1) « Conditionnellement aux notes partielles, la note globale est indépendante des variables explicatives ».
Faire cette hypothèse signifie que les notes partielles suffisent à l'explication de la note globale. Ajouter les variables explicatives dans l'explication n'apporte rien de plus car les notes partielles prennent en compte l'intégralité de l'information contenue dans les variables explicatives. Par ailleurs cette hypothèse suppose aussi que les prestations intermédiaires sont en nombre suffisant pour décrire les mesures physiques. En pratique, cela équivaut à dire que le travail préliminaire qui consiste à lister les critères physiques ainsi que les prestations intermédiaires a bien été fait : aucun critère physique n'influe sur la note globale directement sans passer par l'explication d'une note partielle.

Hypothèse (4.2) « Les prestations intermédiaires sont indépendantes entre elles, conditonnellement aux variables explicatives ».
Cette hypothèse signifie que les prestations intermédiaires ne sont pas redondantes. Il est important d'insister sur la notion d'indépendance *conditionnelle*. Il ne faudrait pas mal interpréter cette hypothèse comme requérant l'indépendance entre les prestations intermédiaires. Dire que les prestations intermédiaires sont indépendantes conditionnellement aux variables explicatives signifie que si l'on enlève de l'étude les variables explicatives conditionnantes (ou qu'on les fixe à certaines valeurs), alors les prestations intermédiaires sont indépendantes entre elles. Cette hypothèse suppose notamment que les variables explicatives ont bien été choisie, et qu'elles sont suffisantes pour décrire non seulement les prestations intermédiaires mais également les inter-dépendances entre celles-ci.

4.5 Résultats

Faute de données réelles, nous allons comparer ces deux méthodes, méthode directe et méthode hiérarchique, sur un jeu de données simulées.

Nous décrivons tout d'abord comment nous avons simulé les données. Les performances des deux estimateurs sont ensuite comparées via le tracé de leurs courbes ROC. Nous rappelons rapidement comment l'on construit ces courbes.

4.5.1 Simulations

Dans un cas pratique réel, on dispose des mesures de p variables explicatives, ainsi que de $Q+1$ notes subjectives (Q notes partielles et une note globale). On fixe donc les vecteurs $\left(\boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^Q, \boldsymbol{\beta}^G\right)$ à des valeurs choisies arbitrairement.

Dans notre cas, nous ne sommes pas limités par l'échantillon de données à notre disposition, puisque l'on compare les approches sur des simulations. Ainsi, nous nous permettons d'estimer les modèles hiérarchique et direct sur plusieurs ensembles d'apprentissage différents, par exemple, sur 100 ensembles. On obtient donc 100 estimateurs pour le modèle direct et autant pour le modèle hiérarchique. La performance des estimateurs, elle, est calculée sur un même jeu de donnée de référence, que l'on appelle ensemble de validation. On moyenne ensuite les performances pour chaque approche, directe ou hiérarchique.

Ci-dessous, on décrit comment l'on simule les données pour ensemble d'apprentissage donné.

Les p variables explicatives : On tire, pour chacune des p variables explicatives $\boldsymbol{X}_j$, n réalisations d'une distribution normale de moyenne nulle et de variance unité. On obtient la matrice d'expérience $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_p]$. On notera $\boldsymbol{X}^{appr}$ pour préciser qu'il s'agit du jeu de données servant à l'apprentissage du modèle.

Les Q notes partielles : Pour chacune des Q notes partielles, on se donne un jeu de coefficient $\boldsymbol{\beta}^q$ de taille p, grâce auquel on simule les n valeurs de la note partielle numéro q. Pour cela, on calcule les quantités :

$$\frac{e^{(\boldsymbol{X}^{appr})_i^T \boldsymbol{\beta}^q}}{1 + e^{(\boldsymbol{X}^{appr})_i^T \boldsymbol{\beta}^q}}$$

On prend ensuite n valeurs s_i^q tirées aléatoirement de manière uniforme entre 0 et 1 et on seuille les quantités précédentes pour obtenir les n valeurs de la note partielle $\boldsymbol{Y}^q = (y_i^q)_{i=1...n}$.

$$\begin{aligned} y_i^q &= 1 \text{ si } \frac{e^{(\boldsymbol{X}^{appr})_i^T \boldsymbol{\beta}^q}}{1 + e^{(\boldsymbol{X}^{appr})_i^T \boldsymbol{\beta}^q}} > s_i^q \\ &= 0 \text{ sinon} \end{aligned}$$

De la même manière que pour la matrice d'expérience, on notera $\boldsymbol{Y}^{appr}$ le vecteur des notes partielles, pour préciser qu'il s'agit du jeu de données servant à l'apprentissage du modèle. $\boldsymbol{Y}^{appr} \in \mathcal{M}_{n \times Q}$.

La note globale : De la même manière que pour les notes partielles, on se donne un jeu de coefficient $\boldsymbol{\beta}^G$ de taille p, grâce auquel on simule les n valeurs de

la note globale. Les quantités calculées sont :

$$\frac{e^{\left(\widetilde{\boldsymbol{Y}}^{appr}\right)_i^T \boldsymbol{\beta}^G}}{1+e^{\left(\widetilde{\boldsymbol{Y}}^{appr}\right)_i^T \boldsymbol{\beta}^G}}$$

où pour les mêmes raisons que dans le paragraphe 4.4.3 :
$\widetilde{\boldsymbol{Y}}^{appr} = 2\boldsymbol{Y}^{appr} - 1$ de sorte que $y_i^q \in \{-1, 1\}$. On prend ensuite n valeurs s_i^G tirées aléatoirement de manière uniforme entre 0 et 1 et on seuille les quantités précédentes pour obtenir les n valeurs de la note globale $\boldsymbol{Y}^G = \left(y_i^G\right)_{i=1...n}$:

$$\begin{aligned} y_i^G &= 1 \text{ si } \frac{e^{\left(\widetilde{\boldsymbol{Y}}^{appr}\right)_i^T \boldsymbol{\beta}^G}}{1+e^{\left(\widetilde{\boldsymbol{Y}}^{appr}\right)_i^T \boldsymbol{\beta}^G}} > s_i^G \\ &= 0 \text{ sinon} \end{aligned}$$

De la même manière que pour la matrice d'expérience, on notera $\boldsymbol{Y}^{G_{appr}}$ le vecteur de la note globale, pour préciser qu'il s'agit du jeu de données servant à l'apprentissage du modèle. $\boldsymbol{Y}^{G_{appr}} \in \mathcal{M}_{n\times 1}$.

Comme on l'a vu plus haut, on réalise cette succession d'étapes d'apprentissage 100 fois. Toutefois, on valide les performances des 100 estimateurs obtenus sur un unique jeu de données.

Ensemble de validation : Cet ensemble est construit sur le même principe que l'ensemble d'apprentissage, en prenant les mêmes valeurs des vecteurs $\left(\boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^Q, \boldsymbol{\beta}^G\right)$. On obtient alors les ensembles de validations : $\boldsymbol{X}^{valid}$, $\boldsymbol{Y}^{valid}$ et $\boldsymbol{Y}^{G_{valid}}$.

Une fois que l'on a construit les deux estimateurs $\widehat{\boldsymbol{\beta}}_{AD}$ et $\widehat{\boldsymbol{\beta}}_{AH}$, on calcule sur cet ensemble de données de validation, une estimation $\widetilde{\boldsymbol{Y}}^{valid,q}$ de la probabilité $\mathbb{P}\left(\boldsymbol{Y}^{valid,q} = 1 | \boldsymbol{X}^{valid}\right)$ pour chacune note partielle $q \in \{1, ..., Q\}$. De même, on calcule une estimation $\widetilde{\boldsymbol{Y}}^{G_{valid}}$ de $\mathbb{P}\left(\boldsymbol{Y}^{G_{valid}} = 1 | \boldsymbol{X}^{valid}\right)$.

Ensemble de test : En toute rigueur, il convient de partager le jeu de données en trois ensembles : un ensemble d'apprentissage, un ensemble de test et un ensemble de validation. L'ensemble apprentissage permet d'estimer le modèle. L'ensemble de test sert à régler l'arrêt de l'apprentissage, afin d'éviter le phénomène de sur-apprentissage. L'ensemble de validation, lui, sert à obtenir une estimation non-optimiste de l'erreur. Dans la pratique, il est fréquent d'omettre cet ensemble, en raison du nombre souvent trop petit de données disponibles. En effet, dans le cas de petits jeux de données, on préfère considérer

la quasi-totalité des données pour apprendre le modèle. Il est tout de même recommandé de conserver quelques données pour valider du modèle obtenu.

Dans notre cas, on fait le choix de ne pas considérer d'ensemble de test. Pour comparer les deux approches, cela n'est nullement dommageable, car dans les courbes ROC, tous les seuils possibles sont considérés.

4.5.2 Exemples de frontières obtenues

L'approche directe fournit, comme dans le cas mono-prestation, une frontière linéaire. Dans l'étude d'une prestation complexe qui se divise en plusieurs sous-prestations, l'obtention d'une frontière linéaire peut paraître un peu restrictif. On propose dans ce paragraphe de prendre deux exemples et d'observer les frontières obtenues par chacune des deux approches.

Pour l'étude des frontières obtenues, nous avons simulé des données de la même manière que dans le paragraphe 4.5.1. On considère deux variables explicatives et une note globale à deux modalités. On regarde les résultats obtenus pour le modèle direct et pour le modèle hiérarchique dans deux cas : un premier cas à deux prestations intermédiaires (fig. 4.1) puis un second cas à six prestations intermédiaires (fig. 4.2 et fig. 4.3).

Toutes les figures sont représentées dans le plan des deux variables explicatives. Nous présentons les résultats de la manière suivante, de haut en bas et de gauche à droite :

- Nous représentons tout d'abord l'ensemble des notes globales. Les points rouges correspondent par exemple à la classe "1" et les points bleus à l'autre classe, la classe "0".
- Nous montrons ensuite ces mêmes points de notes globales avec la nappe de probabilités issue du modèle direct.
- Nous représentons ensuite les points de la note globale ainsi qu'une coupe de cette nappe issue du modèle direct à une valeur seuil choisie arbitrairement. On voit alors nettement la forme prise par la séparation. Dans le cas du modèle direct, cette séparation est ici une droite.
- Nous figurons ensuite les points de la note globale avec la nappe de probabilités issue du modèle hiérarchique.
- Pour une meilleure visualisation de cette nappe de probabilités, nous représentons cette nappe seule, non plus par des niveaux de couleurs, mais directement en 3D.
- Nous représentons les points de la note globale ainsi qu'une coupe de la nappe de probabilités issue du modèle hiérarchique à la même valeur seuil que celle choisie pour le modèle direct. On voit alors nettement la forme prise par la séparation. Dans le cas du modèle hiérarchique, cette séparation

est a priori plus complexe qu'une droite et peut prendre des formes très variées (comme le montre l'exemple à six prestations intermédiaires).

– Enfin, on superpose cette coupe de la nappe de probabilités issue du modèle hiérarchique avec chacune des notes partielles. De la même manière que pour la note globale, les points rouges correspondent à la classe "1" et les points bleus à la classe "0". Ces représentations permettent de voir l'influence qu'a chacune des notes partielles sur la forme finale de la nappe de probabilités.

4.5.3 Estimations

Une fois les données simulées, on s'attache à l'estimation de chacun de deux modèles tels que décrits dans les parties 4.3 et 4.4.

Modèle direct : estimation de $\widehat{\boldsymbol{\beta}}_{AD}$.

Dans le cas du lien direct, l'ensemble d'apprentissage se résume aux variables explicatives $\boldsymbol{X}^{appr}$ et à la note globale $\boldsymbol{Y}^{G_{appr}}$. Comme on l'a vu plus haut, l'information contenue dans les notes partielles ne sert pas à l'estimation de $\widehat{\boldsymbol{\beta}}_{AD}$.

Tout d'abord, on construit le chemin de régularisation qui fait le lien entre la note globale et les variables explicatives. On utilise pour cela l'algorithme décrit dans le chapitre 2. On obtient donc une suite croissante de modèles et on utilise le BIC pour sélectionner les variables explicatives à retenir pour la modélisation. On construit enfin le modèle logistique non pénalisé calculé à partir uniquement des variables explicatives sélectionnées. C'est ainsi que l'on obtient $\widehat{\boldsymbol{\beta}}_{AD}$.

Modèle hiérarchique : estimation de $\widehat{\boldsymbol{\beta}}_{AH}$.

En ce qui concerne l'approche hiérarchique, l'ensemble d'apprentissage comprend les variables explicatives, la note globale, mais également les notes partielles $\boldsymbol{Y}^{appr}$. De manière similaire avec l'approche directe, on construit tout d'abord le chemin de régularisation qui fait le lien entre la note globale et les notes partielles. A l'aide du BIC, on sélectionne les notes partielles qui sont significativement explicatives de la note globale. On construit ensuite le modèle logistique non pénalisé calculé à partir uniquement des notes partielles sélectionnées par le BIC.

On s'intéresse seulement dans un second temps à la modélisation des notes partielles en fonction des variables explicatives. En effet, comme on peut le voir dans le calcul de la vraisemblance dans le cas du modèle hiérarchique (voir équation (4.3)), il est inutile de calculer $\widehat{\boldsymbol{\beta}}^{q}_{AH}$ si la note partielle $\boldsymbol{Y}^{q}$ n'a pas été sélectionnée par le BIC pour participer à l'explication de la note

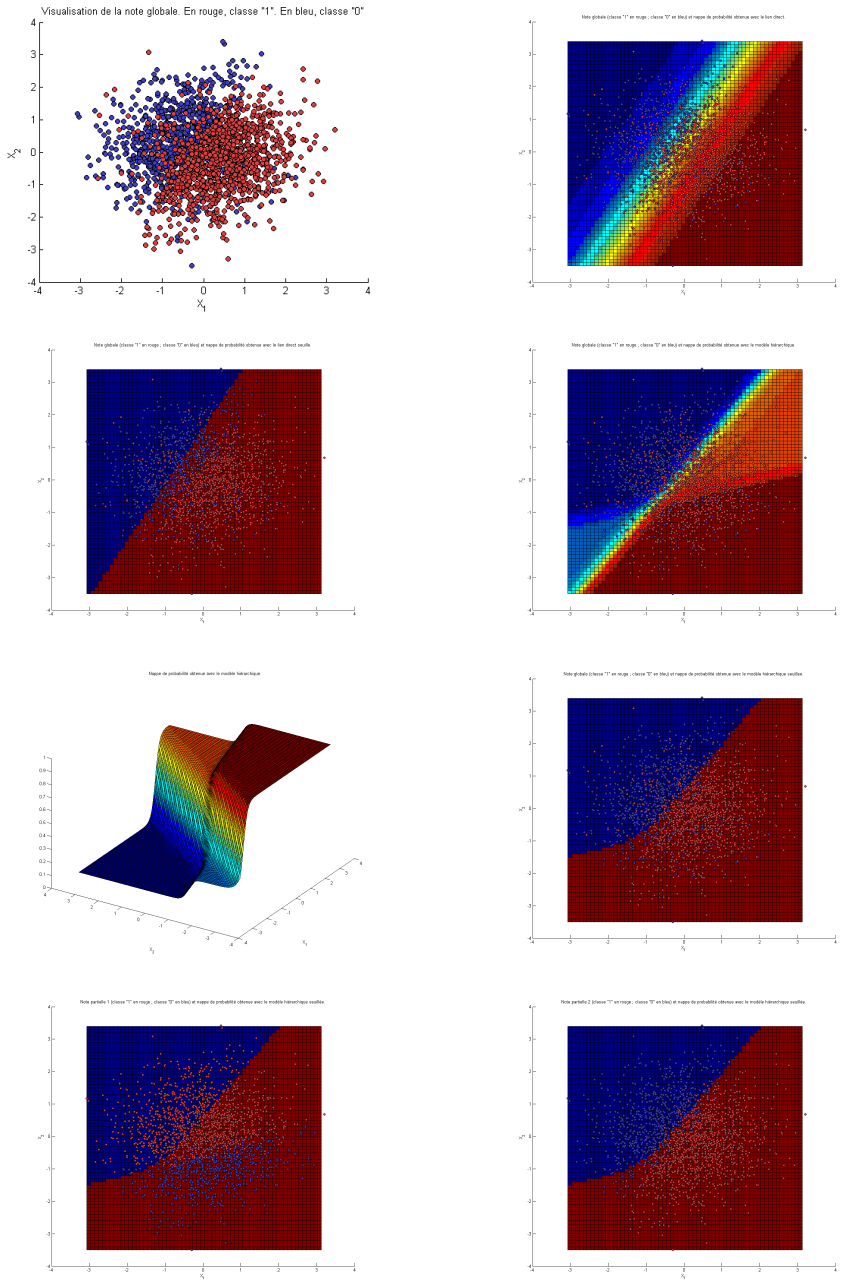

FIG. 4.1 – Cas à deux variables explicatives et deux prestations intermédiaires. Pour la signification de chacune des figures, se reporter au paragraphe 4.5.2.

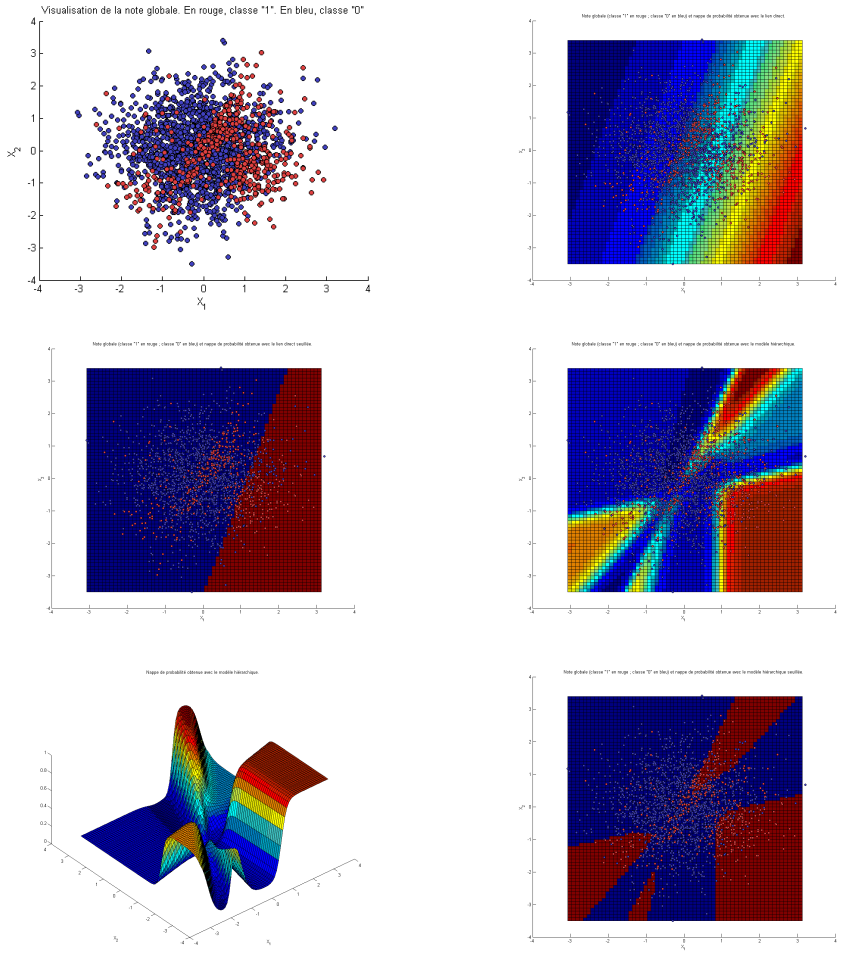

FIG. 4.2 – Cas à deux variables explicatives et six prestations intermédiaires. Pour la signification de chacune des figures, se reporter au paragraphe 4.5.2. Pour les représentations des notes partielles et de leur influence sur la nappe de probabilité issue du modèle hiérarchique, se reporter à la figure fig. 4.3.

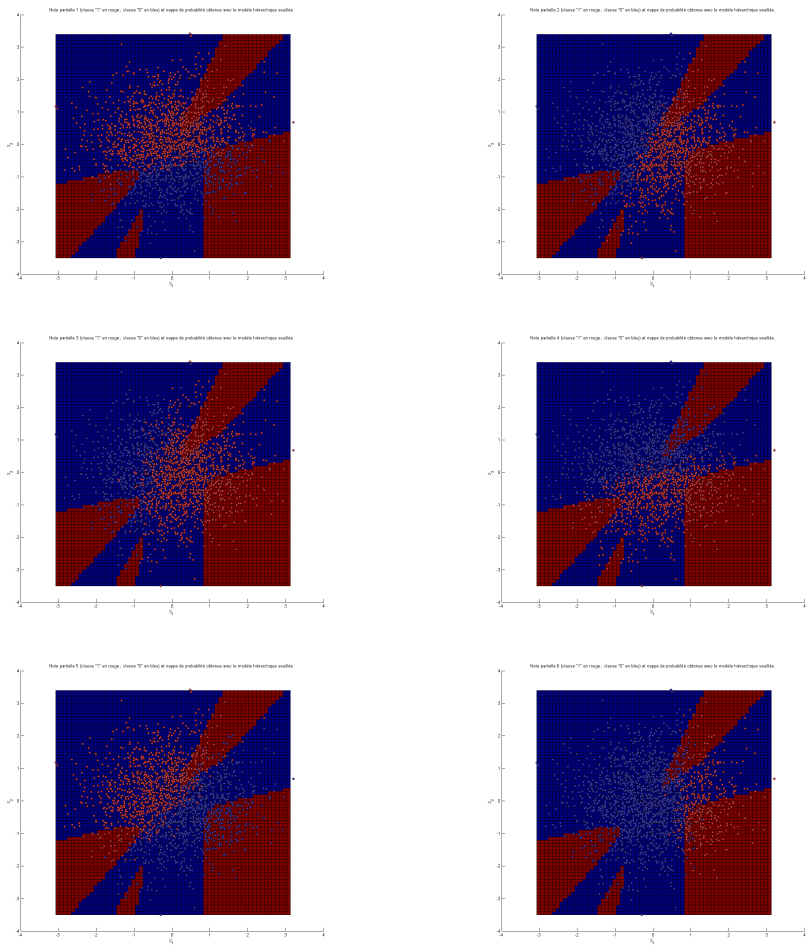

FIG. 4.3 – Coupe de la nappe de probabilité issue du modèle hiérarchique avec chacune des notes partielles pour le cas à deux variables explicatives et six prestations intermédiaires.

globale. Ainsi, pour chacune des notes partielles sélectionnées par le BIC à l'étape précédente, et pour celles-là uniquement, on construit le chemin de régularisation qui fait le lien entre cette note partielle et l'ensemble des variables explicatives. A l'aide du BIC, on sélectionne les variables significativement explicatives de chacune des notes partielles sélectionnées. On construit ensuite les modèles logistiques non pénalisés $\widehat{\boldsymbol{\beta}}^{q}_{AH}$. Ainsi, on obtient $\widehat{\boldsymbol{\beta}}_{AH}$ via la formule donnée par l'équation (4.3).

Le nombre de variables explicatives et/ou de notes partielles retenues par le BIC dans le chemin de régularisation dépend très fortement de l'ensemble d'apprentissage. Le fait de calculer 100 fois les estimateurs atténue cet effet important de variation du nombre de variables explicatives et/ou notes partielles retenues par le BIC.

4.5.4 Rappels sur les courbes ROC

On évalue la performance des estimateurs en traçant leurs courbes ROC (pour Receiver Operating Characteristic). Cette technique a été utilisée historiquement dans le domaine médical. On pourra se référer à l'article de Swets et al. [43] sur l'utilisation des courbes ROC. On rappelle brièvement dans ce paragraphe comment sont construites ces courbes.

Dans notre application industrielle, on veut que le véhicule qui est conforme au cahier des charges reçoive une note subjective positive de la part du client. Notre "hypothèse nulle" est donc : « Le client donne une note subjective positive. »

$$H_0 : \{\boldsymbol{Y}^G = 1\}$$

On construit alors le test suivant :
- Si $\mathbb{P}\left(\boldsymbol{Y}^G = 1 | \boldsymbol{X} = \boldsymbol{x}\right) <$ seuil alors on rejette H_0.
- Si $\mathbb{P}\left(\boldsymbol{Y}^G = 1 | \boldsymbol{X} = \boldsymbol{x}\right) \geq$ seuil alors on accepte H_0.

Il y a deux types d'erreurs :

Risque de première espèce : probabilité que le test rejette à tort l'hypothèse H_0 lorsqu'elle est vraie. Ce risque est souvent noté α.

Risque de deuxième espèce : probabilité que le test accepte à tort l'hypothèse H_0 lorsqu'elle est fausse. Ce risque est souvent noté β.

On fait jouer à ces deux erreurs un rôle asymétrique. Il est en effet plus dommageable d'accepter l'hypothèse nulle alors qu'elle est fausse (risque de second espèce). En effet, en termes industriels, cela correspond au cas de figure où le véhicule est conforme au cahier des charges et pourtant le client l'évalue négativement. Dans la pratique, c'est principalement ce type d'erreur qui préoccupe les ingénieurs.

Les différentes erreurs possibles se résume par la matrice dite de confusion :

réel test	$\boldsymbol{Y}^G$=1	$\boldsymbol{Y}^G = 0$
$\mathbb{P}\left(\boldsymbol{Y}^G = 1 \vert \boldsymbol{X} = \boldsymbol{x}\right) \geq$ seuil	Vrai Positif	Faux Positif
$\mathbb{P}\left(\boldsymbol{Y}^G = 1 \vert \boldsymbol{X} = \boldsymbol{x}\right) <$ seuil	Faux Negatif	Vrai Negatif

où on l'on a :

Vrai Positif (VP) : essai pour lequel le test accepte à raison H_0.

Vrai Negatif (VN) : essai pour lequel le test rejette à raison H_0.

Faux Positif (FP) : essai pour lequel le test accepte à tort H_0.

Faux Negatif (FN) : essai pour lequel le test rejette à tort H_0.

On définit également les quantités suivantes :

Sensibilité : c'est le taux de vrais positifs.

$$\text{sensibilité } = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

Le risque de seconde espèce étant le complément à 1 de cette quantité, le test est d'autant meilleur que la sensibilité est grande.

Spécificité :

$$\text{spécificité } = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

Le complément à 1 de cette quantité est le risque de première espèce. La quantité (1-spécificité) est aussi appelée taux de faux positifs ou encore taux de fausses alarmes.

Une courbe ROC s'obtient en traçant la sensibilité en fonction de la quantité (1-spécificité) pour toutes les valeurs de seuil possibles. Cela est équivalent de dire que la courbe ROC s'obtient en traçant le taux de vrais positifs en fonction du taux de faux positifs.

4.5.5 Comparaison des deux approches

Les performances de chacune des 100 paires d'estimateurs $\left(\widehat{\boldsymbol{\beta}}_{AD}, \widehat{\boldsymbol{\beta}}_{AH}\right)$ obtenues sont comparées en traçant les courbes ROC de chacun des estimateurs. Pour

une comparaison globale des deux approches, directe et hiérarchique, on moyenne les 100 courbes ROC obtenues pour chaque estimateur. Cette technique permet également d'amoindrir les effets aléatoires introduits notamment dans la construction des variables explicatives et des notes partielles, via le seuillage aléatoire.

4.5.5.1 Exemple à 4 notes partielles avec parcimonie

Le premier exemple présenté ici est un exemple que l'on suppose être assez proche d'un jeu de données réelles. En effet, on considère que l'on a :

- 14 variables explicatives,
- 4 notes partielles et
- 1 note globale.

Les 4 notes partielles sont construites de manière parcimonieuse. Cela signifie que pour chaque note partielle, seulement un petit nombre des 14 variables explicatives participe avec un coefficient non nul à la construction de cette note. Certaines variables explicatives (au nombre de 4) n'interviennent dans la construction d'aucune note partielle (coefficient nul). A contrario, une des variables explicatives intervient dans la construction de chacune des notes partielles. Cette variable commune exceptée, les ensembles de variables explicatives impliquées dans la construction d'une note partielle sont disjoints. Si l'on fait le rapport entre le nombre de coefficients à zéro et le nombre total de coefficients, on définit ainsi un taux de parcimonie qui est ici de 76.8%. On note qu'un tel exemple est proche de ce que l'on peut trouver dans la réalité. En effet, les variables explicatives impliquées dans la construction de différentes prestations intermédiaires peuvent provenir de domaines d'expertise radicalement différents. Pour autant, on se doit d'estimer la significativité statistique d'une variable explicative d'un domaine par rapport à une prestation intermédiaire d'un autre domaine. D'où la présence d'un taux de parcimonie - tel qu'on l'a rapidement défini ci-dessus - éventuellement grand.

On représente dans la colonne de gauche de la figure 4.4 les courbes ROC pour trois différentes tailles d'ensembles d'apprentissage. On considère des ensembles d'apprentissage de taille :

- $n_{appr} = 30$,
- $n_{appr} = 50$ et
- $n_{appr} = 100$.

On voit assez nettement que le modèle hiérarchique a de meilleures performances que le modèle direct. En effet, l'écart entre les courbes est toujours en faveur du modèle hiérarchique. On note également que cet écart s'accroît avec la taille de l'ensemble d'apprentissage. Cela signifie que plus le jeu de donnée est petit, plus la différence entre les deux approches s'estompe. Cependant, on aurait pu s'attendre à ce que le modèle direct ait de meilleures performances sur les petits

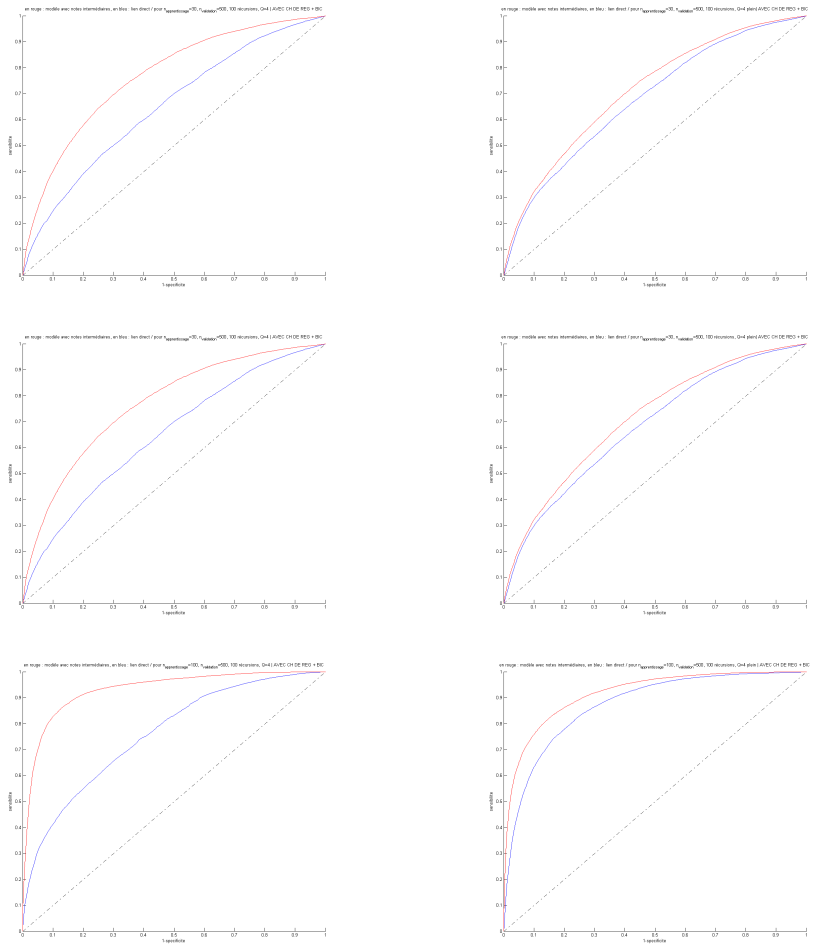

FIG. 4.4 – Courbes ROC pour l'exemple à 4 notes partielles avec parcimonie (colonne de gauche) et sans parcimonie (colonne de droite). Courbes bleues : lien direct. Courbes rouges : modèle hiérarchique. En haut : $n_{appr} = 30$. Au milieu : $n_{appr} = 50$. En bas : $n_{appr} = 100$.

ensembles d'apprentissage, mais ce n'est pas le cas. En effet, les petits ensembles d'apprentissage favorisent le modèle direct au détriment du modèle hiérarchique car ce dernier possède un nombre de coefficients à estimer supérieur au modèle direct. Ainsi le modèle direct possède une variance a priori plus faible que le modèle hiérarchique. Par ailleurs il faut noter également que le jeu de données a été simulé conformément au modèle hiérarchique. Ce qui signifie que le modèle hiérarchique a, lui, un biais a priori plus faible que le modèle direct.

4.5.5.2 Exemple à 4 notes partielles sans parcimonie

Cet exemple est en tout point similaire à l'exemple précédent, à ceci près que le taux de parcimonie est de 0%. En effet, toutes les variables explicatives interviennent dans la construction de toutes les notes partielles. On présente ces résultats dans la colonne de droite de la figure 4.4.

On note que les performances des estimateurs croissent, comme sur l'exemple précédent, avec la taille de l'ensemble d'apprentissage, ce qui n'a rien d'étonnant. Ce qui est plus étonnant est que l'écart entre les performances des deux approches est plus petit, quand bien même le modèle hiérarchique reste meilleur quelle que soit la taille du jeu d'apprentissage.

Si l'on compare l'exemple sans parcimonie et l'exemple avec parcimonie (c'est-à-dire les résultats de la colonne de droite et les résultats de la colonne de gauche), on voit que l'augmentation de la variance affecte davantage le modèle hiérarchique que le modèle direct. Le modèle hiérarchique est meilleur quand les prestations sont parcimonieuses. En revanche, la parcimonie du modèle n'est pas bien exploitée par la méthode d'estimation directe.

4.5.5.3 Exemple à 7 notes partielles dont une seule est non nulle

On regarde enfin le cas où seule une des prestations intermédiaires influe effectivement sur la note globale. Le taux de parcimonie est alors de 85.7%. On est alors dans un cas où l'approche directe est la modélisation *naturelle*. Nous considérons cet exemple pour vérifier si, dans un tel cas, il est aberrant ou non d'appliquer l'approche hiérarchique.

Les résultats sont présentés en fig. 4.5. On note que les courbes sont très semblables. Ce résultat est rassurant car le modèle hiérarchique n'est pas mis en défaut sur cet exemple. En effet, même sur cet exemple dédié au lien direct, le modèle hiérarchique est au moins aussi performant.

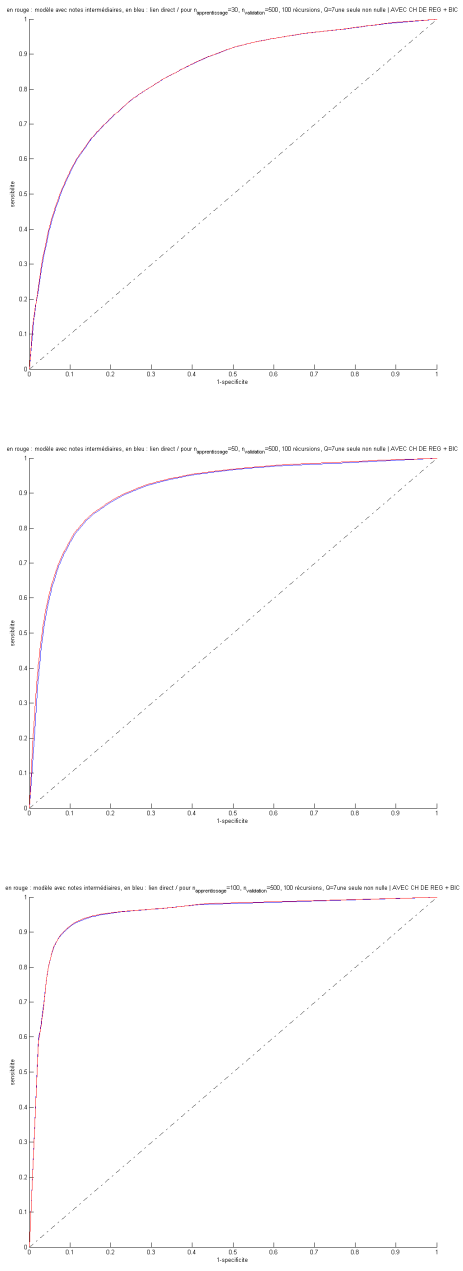

FIG. 4.5 – Courbes ROC pour l'exemple à 7 notes partielles dont une seule est non nulle. Courbes bleues : approche directe. Courbes rouges : approche hiérarchique. En haut : $n_{appr} = 30$. Au milieu : $n_{appr} = 50$. En bas : $n_{appr} = 100$.

4.5.6 Conclusion

On conclut que, dans tous les cas, il est préférable de passer par une approche hiérarchique.

Les résultats montrent en effet qu'il y a un fort intérêt statistique à appliquer une décomposition de la prestation étudiée en plusieurs prestations intermédiaires plutôt que d'appliquer une objectivation directe sur la note globale. De plus l'information sur la décomposition des prestations intermédiaire au sein de la construction de la note globale est une information qui a un sens en elle-même pour les ingénieurs. En effet, ces derniers sont intéressés par la hiérarchisation des prestations intermédiaires les unes par rapport aux autres, dans un objectif d'explication de la prestation globale.

On note aussi que le modèle hiérarchique voit ses performances dépasser largement celles de l'approche directe quand les modèles de prestations intermédiaires sont parcimonieux, ce qui est très souvent le cas dans les applications pratiques.

Chapitre 5

Régression logistique où certaines variables explicatives ne sont pas observées

Comme on l'a vu au chapitre 4, il est intéressant de passer par une description hiérarchique du problème quand une prestation se décompose en plusieurs prestations intermédiaires. Pour appliquer cette approche hiérarchique, nous avons également vu qu'il fallait vérifier deux hypothèses.

Dans cette partie, on propose une méthode pour tester la première hypothèse. La question à laquelle on souhaite répondre est la suivante : comment s'assurer qu'il n'existe pas une variable que l'on n'a pas mesurée et qui influe pourtant sur le jugement de la prestation ?

5.1 Formalisation du problème

On se ramène au cas mono-prestation.

Soit $\boldsymbol{X}$ un vecteur gaussien. On cherche à modéliser la réponse $\boldsymbol{Y}$ à partir des variables explicatives $\boldsymbol{X}$. On suppose maintenant que parmi ces variables explicatives, certaines sont observées et d'autres ne le sont pas. On note $\boldsymbol{X_1}$ l'ensemble des variables explicatives qui sont observées. $\boldsymbol{X_1}$ est de dimension p.

On note $\boldsymbol{U}$ l'ensemble des variables explicatives qui ne sont pas observées. On ne connaît bien sûr pas la dimension de cet ensemble. On suppose que $\boldsymbol{X_1}$ et $\boldsymbol{U}$ sont des vecteurs conjointement gaussiens et que $\boldsymbol{Y}$ suit un modèle logistique avec $\boldsymbol{X_1}$ et $\boldsymbol{U}$ comme variables explicatives.

Le modèle s'écrit donc, pour une observation donnée :

$$\mathbb{P}(\boldsymbol{Y} = 1|\boldsymbol{X}) = \frac{e^{\boldsymbol{X}^T\boldsymbol{\beta}}}{1 + e^{\boldsymbol{X}^T\boldsymbol{\beta}}}$$

avec $\boldsymbol{X} = [\boldsymbol{X_1}^T, \boldsymbol{U}^T]^T$. On pose $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T]^T$ de sorte que l'on a la décomposition suivante :

$$\boldsymbol{X}^T\boldsymbol{\beta} = \boldsymbol{X_1}^T\boldsymbol{\beta_1} + \boldsymbol{U}^T\boldsymbol{\beta_2}$$

On projette maintenant $\boldsymbol{U}$ sur $\boldsymbol{X_1}$. Le projeté qui minimise l'écart quadratique s'écrit, sous l'hypothèse gaussienne :

$$\mathbb{E}[\boldsymbol{U}|\boldsymbol{X_1}] = \boldsymbol{M}\boldsymbol{X_1}$$

d'où :

$$\boldsymbol{U} = \boldsymbol{M}\boldsymbol{X_1} + \boldsymbol{R}$$

avec $\boldsymbol{R}$ gaussien et indépendant de $\boldsymbol{X_1}$.

$$\boldsymbol{X}^T\boldsymbol{\beta} = \underbrace{\boldsymbol{X_1}^T\boldsymbol{\beta_1} + \boldsymbol{X_1}^T\boldsymbol{M}^T\boldsymbol{\beta_2}}_{\boldsymbol{X_1}^T\tilde{\boldsymbol{\beta}}} + \underbrace{\boldsymbol{R}^T\boldsymbol{\beta_2}}_{\tilde{\sigma}\boldsymbol{X_2}}$$

où $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta_1} + \boldsymbol{M}^T\boldsymbol{\beta_2}$. $\boldsymbol{R}^T\boldsymbol{\beta_2} \sim \mathcal{N}(0, \tilde{\sigma}^2)$ ou encore : $\boldsymbol{X_2} \sim \mathcal{N}(0, 1)$, (de dimension 1).

On a opéré une simple reparamétrisation. Le modèle se récrit alors :

$$\mathbb{P}(\boldsymbol{Y} = 1|\boldsymbol{X}) = \frac{e^{\boldsymbol{X_1}^T\tilde{\boldsymbol{\beta}} + \tilde{\sigma}\boldsymbol{X_2}}}{1 + e^{\boldsymbol{X_1}^T\tilde{\boldsymbol{\beta}} + \tilde{\sigma}\boldsymbol{X_2}}} \tag{5.1}$$

où $\boldsymbol{X_1}$ est observé et où $\boldsymbol{X_2}$ est dite variable cachée. De plus, le cas $\tilde{\sigma} = 0$ correspond au cas où toutes les variables explicatives sont contenues dans $\boldsymbol{X_1}$. De la même manière que précédemment, on cherche donc à maximiser la vraisemblance des données complète $L_n(\theta)$ à laquelle on va ajouter une fonction de pénalisation $J(\theta)$:

$$\min_{\theta} \{-\log L_n(\theta) + \lambda J(\theta)\} \ . \tag{5.2}$$

La vraisemblance s'écrit :

$$-\log L_n(\theta) = \sum_{i=1}^{n} -\log p_\theta(y_i, \boldsymbol{x_{1i}}) \ . \tag{5.3}$$

Et donc la fonction de contraste pénalisée s'écrit, comme $\theta = (\boldsymbol{\beta}, \sigma)$:

$$-\log L_n(\boldsymbol{\beta}, \sigma) + \lambda J(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^{n} -\log p_{\boldsymbol{\beta},\sigma}(y_i, \boldsymbol{x_{1i}}) + \lambda J(\boldsymbol{\beta}, \sigma) \ . \tag{5.4}$$

La fonction $J(\theta)$ est convexe en ses paramètres. En revanche, du fait de la présence d'une variable cachée, la fonction $(-\log L_n(\theta))$ ne l'est pas a priori et par voie de conséquence, $(-\log L_n(\theta) + \lambda J(\theta))$ n'est pas convexe non plus (voir un exemple en fig. 5.1).

Ainsi pour optimiser la fonction de contraste pénalisée, on ne peut pas bénéficier des techniques rapides d'optimisation convexe existantes. Pour réaliser proprement cette optimisation, il est classique d'appliquer un algorithme EM.

5.2 Algorithme EM

L'algorithme Espérance-Maximisation, (en anglais *Expectation-Maximization algorithm*), souvent noté EM, est introduit par Dempster et al. dans [11]. Le but de cet algorithme est en effet de calculer un maximum de vraisemblance lorsque le modèle dépend de variables cachées.

Le but final est de minimiser en θ la quantité :

$$-\log L_n(\theta) + \lambda J(\theta)$$

où $-\log L_n(\theta)$ a été définie dans l'équation (5.3).

On rappelle tout d'abord rapidement en quoi consiste l'algorithme EM dans sa version sans pénalisation.

5.2.1 Principe général

On définit tout d'abord, pour deux vecteurs de paramètres θ et θ', la fonction $Q(\theta, \theta')$ grâce à une espérance conditionnée par les variables observées :

$$Q(\theta, \theta') = \sum_{i=1}^{n} \mathbb{E}_{\theta'} \left[-\log p_\theta(y_i, \boldsymbol{x_{1i}}, \boldsymbol{x_{2i}}) | y_i, \boldsymbol{x_{1i}} \right] \tag{5.5}$$

On procède de la manière suivante :

Initialisation : L'initialisation de l'algorithme EM est un problème en soi, et est loin d'être résolu. On suppose que l'on sait initialiser l'algorithme par *bonne valeur*. L'initialisation au hasard n'est pas nécessairement le meilleur des choix, notamment parce que l'algorithme EM n'assure qu'une convergence vers un optimum local de la fonction $(-\log L_n)$ (voir [11] pour les détails). Afin d'exposer le principe général de l'algorithme EM, on suppose donc qu'on part d'une valeur proche de l'optimum global de $(-\log L_n(\theta))$, qu'on note $\theta^{(0)}$. On initialise les paramètres courants par $\theta^{(c)} = \theta^{(0)}$.

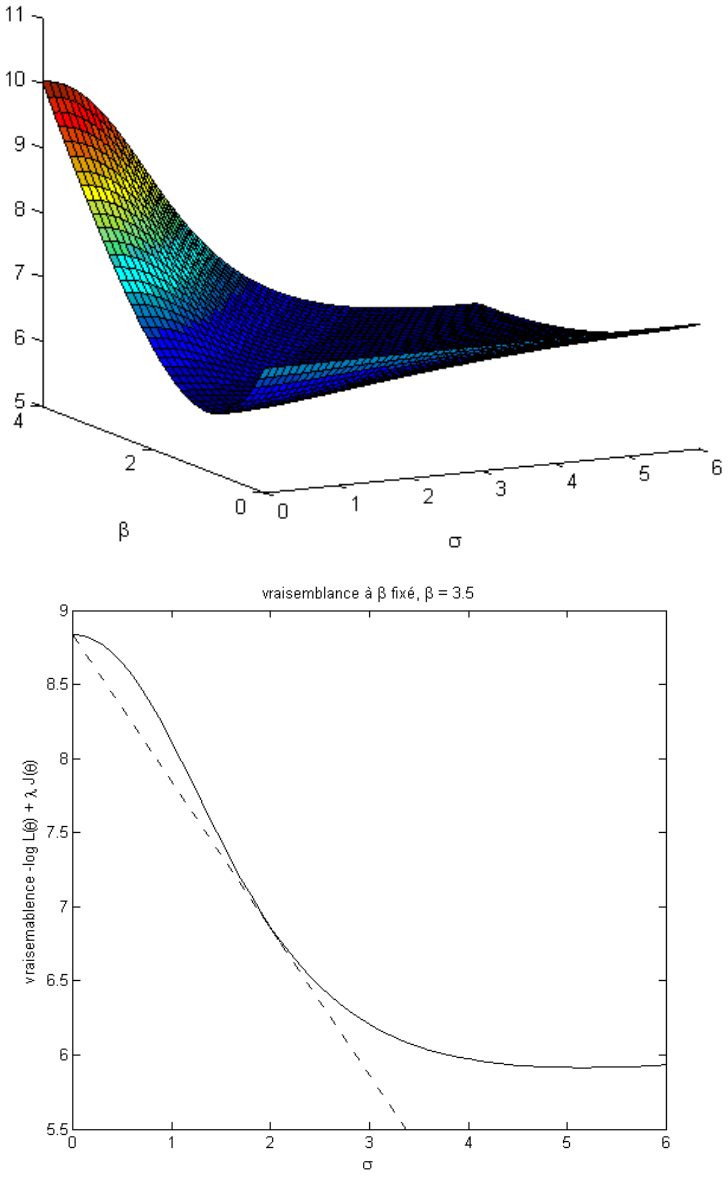

FIG. 5.1 – On a considéré un modèle à une seule variable observée. σ est l'écart-type de la variable cachée. En haut : on représente la vraisemblance pour une grille de valeurs pour ces deux paramètres. En bas : on représente une coupe de cette vraisemblance pour $\beta = 3.5$ fixé.

Etape E : Il s'agit de l'étape d'évaluation de l'espérance. On calcule : $-\log L_n\left(\theta^{(c)}\right)$. Tant que l'algorithme n'a pas convergé, on calcule $Q\left(\theta,\theta^{(c)}\right)$, pour θ quelconque, les paramètres courants $\theta^{(c)}$ étant connus.

Etape M : Il s'agit de l'étape de maximisation.

$$\theta^{new} = \arg\min_{\theta\in\Theta} Q\left(\theta,\theta^{(c)}\right)$$

$\theta^{(c)} = \theta^{new}$ et on reprend à l'étape E.

Condition d'arrêt : On dit que l'algorithme converge lorsque la valeur $\left(-\log L_n\left(\theta^{(c)}\right)\right)$ ne varie plus. [1]

L'algorithme décrit ci-dessus est justifié par la propriété suivante :

Propriété 5.1 *Quel que soit un vecteur de paramètres θ, on a :*

$$Q\left(\theta,\theta^{(c)}\right) \leq Q\left(\theta^{(c)},\theta^{(c)}\right) \implies -\log L\left(\theta\right) \leq -\log L_n\left(\theta^{(c)}\right)$$

En effet, par définition, on a :

$$Q\left(\theta,\theta^{(c)}\right) = \sum_{i=1}^{n} \mathbb{E}_{\theta'}\left[-\log p_\theta(y_i, \boldsymbol{x_{1i}}, \boldsymbol{x_{2i}})|y_i, \boldsymbol{x_{1i}}\right]$$

Pour alléger les notations, on considère un i donné. On retrouvera la sommation sur i à la fin.

$$\begin{aligned}
& \mathbb{E}_{\theta'}\left[-\log p_\theta(y, \boldsymbol{x_1}, \boldsymbol{x_2})|y, \boldsymbol{x_1}\right] \\
= \; & \mathbb{E}_{\theta'}\left[\left(-\log p_\theta(\boldsymbol{x_2}|y, \boldsymbol{x_1}) - \log p_\theta(y, \boldsymbol{x_1})\right)|y, \boldsymbol{x_1}\right] \\
= \; & \mathbb{E}_{\theta'}\left[-\log p_\theta(\boldsymbol{x_2}|y, \boldsymbol{x_1})|y, \boldsymbol{x_1}\right] - \log p_\theta(y, \boldsymbol{x_1}) \\
= \; & \mathbb{E}_{\theta'}\left[-\log \frac{p_\theta(\boldsymbol{x_2}|y, \boldsymbol{x_1})}{p_{\theta^{(c)}}(\boldsymbol{x_2}|y, \boldsymbol{x_1})}|y, \boldsymbol{x_1}\right] \\
& -\mathbb{E}_{\theta'}\left[\log p_{\theta^{(c)}}(\boldsymbol{x_2}|y, \boldsymbol{x_1})|y, \boldsymbol{x_1}\right] - \log p_\theta(y, \boldsymbol{x_1}) \\
= \; & \mathbb{E}_{\theta'}\left[\log \frac{p_{\theta^{(c)}}(\boldsymbol{x_2}|y, \boldsymbol{x_1})}{p_\theta(\boldsymbol{x_2}|y, \boldsymbol{x_1})}|y, \boldsymbol{x_1}\right] \\
& -\mathbb{E}_{\theta'}\left[\log p_{\theta^{(c)}}(\boldsymbol{x_2}|y, \boldsymbol{x_1})|y, \boldsymbol{x_1}\right] - \log p_\theta(y, \boldsymbol{x_1}) \\
= \; & K\left(p_{\theta^{(c)}}\left(.|y, \boldsymbol{x_1}\right)||p_\theta\left(.|y, \boldsymbol{x_1}\right)\right) \\
& +\mathbb{E}_{\theta'}\left[-\log p_{\theta^{(c)}}(\boldsymbol{x_2}|y, \boldsymbol{x_1})|y, \boldsymbol{x_1}\right] - \log p_\theta(y, \boldsymbol{x_1})
\end{aligned}$$

[1] La règle d'arrêt peut être, par exemple, de la forme : si pendant 5 itérations, $\left(-\log L_n\left(\theta^{(c)}\right)\right)$ ne décroît pas de plus de 1% alors s'arrêter.

où $K\left(p_{\theta^{(c)}}\left(.|y,\boldsymbol{x_1}\right)||p_\theta\left(.|y,\boldsymbol{x_1}\right)\right)$ est la distance de Kullback-Leibner entre ces deux distributions.

D'autre part, en réécrivant

$$\mathbb{E}_{\theta'}\left[-\log p_{\theta^{(c)}}(\boldsymbol{x_2}|y,\boldsymbol{x_1})|y,\boldsymbol{x_1}\right]=\mathbb{E}_{\theta'}\left[-\log p_{\theta^{(c)}}(y,\boldsymbol{x_1},\boldsymbol{x_2})|y,\boldsymbol{x_1}\right]+\log p_{\theta^{(c)}}(y,\boldsymbol{x_1})$$

on obtient, en sommant à nouveau sur i :

$$\begin{aligned} Q\left(\theta,\theta^{(c)}\right) &= nK\left(p_{\theta^{(c)}}\left(.|\boldsymbol{Y},\boldsymbol{X_1}\right)||p_\theta\left(.|\boldsymbol{Y},\boldsymbol{X_1}\right)\right) \\ &\quad +Q\left(\theta^{(c)},\theta^{(c)}\right) \\ &\quad +\sum_{i=1}^{n}-\log p_\theta(y_i,\boldsymbol{x_{1i}})-\sum_{i=1}^{n}-\log p_{\theta^{(c)}}(y_i,\boldsymbol{x_{1i}})\ , \end{aligned}$$

expression dans laquelle on reconnaît l'expression de $(-\log L_n(\theta))$ et $\left(-\log L_n\left(\theta^{(c)}\right)\right)$.

La propriété de positivité de la distance de Kullback-Leibler :

$$K\left(p_{\theta^{(c)}}\left(.|y,x_1\right)||p_\theta\left(.|y,x_1\right)\right)\geq 0$$

nous donne finalement :

$$Q\left(\theta,\theta^{(c)}\right)-Q\left(\theta^{(c)},\theta^{(c)}\right)\geq -\log L_n(\theta)-\left(-\log L_n\left(\theta^{(c)}\right)\right)\ . \qquad (5.6)$$

D'où la propriété 5.1 :

$$Q\left(\theta,\theta^{(c)}\right)\leq Q\left(\theta^{(c)},\theta^{(c)}\right)\implies -\log L\left(\theta\right)\leq -\log L_n\left(\theta^{(c)}\right)\ .$$

On s'assure ainsi qu'à chaque itération de l'algorithme EM, on augmente la vraisemblance (en pratique, comme on l'a vu, on diminue la fonction de contraste , en l'occurence ici la fonction de contraste est l'opposé de la log-vraisemblance). En effet, on a le résultat suivant.

Corollaire 1 *D'après la définition de* $\theta_{new}=\arg\min_{\theta\in\Theta}Q\left(\theta,\theta^{(c)}\right)$*, on a*

$$Q\left(\theta^{new},\theta^{(c)}\right)\leq Q\left(\theta^{(c)},\theta^{(c)}\right)\ .$$

Donc il vient :

$$-\log L\left(\theta^{new}\right)\leq -\log L_n\left(\theta^{(c)}\right)\ .$$

5.2.2 Algorithme EM pénalisé

On ajoute maintenant à la fonction de vraisemblance la fonction de pénalisation $J(\theta)$. Le résultat de l'algorithme EM sur l'optimisation de la fonction $-\log L_n(\theta)+\lambda J(\theta)$ ne sont pas modifiés. En effet, il suffit d'ajouter la même fonction de pénalisation à la fonction auxiliaire $Q(\theta,\theta')$ et de remarquer que :

$$Q\left(\theta,\theta^{(c)}\right)-Q\left(\theta^{(c)},\theta^{(c)}\right)\geq-\log L_n(\theta)-\left(-\log L_n\left(\theta^{(c)}\right)\right)$$

implique

$$\begin{aligned}&Q\left(\theta,\theta^{(c)}\right)+\lambda J(\theta)-\left(Q\left(\theta^{(c)},\theta^{(c)}\right)+\lambda J\left(\theta^{(c)}\right)\right)\\&\geq-\log L_n(\theta)+\lambda J(\theta)-\left(-\log L_n\left(\theta^{(c)}\right)+\lambda J\left(\theta^{(c)}\right)\right)\ .\end{aligned}$$

Ainsi l'ajout d'une fonction de pénalisation ne modifie rien au comportement de l'algorithme.

Pour imposer une certaine parcimonie au vecteur $\boldsymbol{\beta}$, on a choisi de pénaliser la vraisemblance à l'aide de la norme L_1 du vecteur de paramètres. La variable non observée doit être prise en compte dans la pénalisation. D'après la forme du modèle logistique en présence d'une variable cachée donnée par l'équation (5.1), la fonction $J(\theta)$ prend naturellement la forme suivante :

$$J(\boldsymbol{\beta},\sigma)=\sum_{j=1}^{p}|\beta_j|+\sigma \tag{5.7}$$

5.3 Mise en pratique de l'algorithme EM

Dans ce paragraphe, nous allons détailler les expressions nécessaires à la mise en pratique de l'algorithme EM pénalisé.

On explicite tout d'abord l'expression de $Q(\theta,\theta')$ dans le cas de la régression logistique binaire, puis l'expression de la fonction de contraste pénalisée, qui contrôle la convergence de l'algorithme. Grâce à l'expression que l'on obtient de $Q(\theta,\theta')$, on peut maintenant implémenter l'algorithme EM, comme suit :

Initialisation : (cf Principe général 5.2.1)

Etape E : calcul de $Q(\theta,\theta')+\lambda J(\theta)$.

Etape M : optimisation en θ de l'expression $Q(\theta, \theta') + \lambda J(\theta)$, à θ' fixé :

$$\theta^{new} = \arg\min_{\theta} Q(\theta, \theta') + \lambda J(\theta)$$

Ici $\theta = (\boldsymbol{\beta}, \sigma)$. Lorsqu'on établit l'expression de $Q(\theta, \theta')$, il apparaît qu'il est possible d'optimiser séparément $Q(\theta, \theta') + \lambda J(\theta)$ en $\boldsymbol{\beta}$ et en σ. On note également que l'optimisation en σ peut se faire de manière analytique.

Condition d'arrêt : on utilise le même critère que dans le Principe général du paragraphe 5.2.1, en substituant $-\log L_n\left(\theta^{(c)}\right)$ par $-\log L_n\left(\theta^{(c)}\right) + \lambda J\left(\theta^{(c)}\right)$.

5.3.1 Expression de la fonction Q dans le cas de la régression logistique binaire

Dans ce paragraphe, on explicite la forme que prend la fonction auxiliaire Q dans le cas de la régression logistique binaire.

$$Q(\theta, \theta') = \sum_{i=1}^{n} \mathbb{E}_{\theta'}\left[-\log p_\theta(y_i, \boldsymbol{x_{1i}}, \boldsymbol{x_{2i}}) | y_i, \boldsymbol{x_{1i}}\right]$$

Pour un i donné, on a :

$$\begin{aligned} & \mathbb{E}_{\theta'}\left[-\log p_\theta(y, \boldsymbol{x_1}, \boldsymbol{x_2}) | y, \boldsymbol{x_1}\right] \\ = & \int -\log p_\theta(y, \boldsymbol{x_1}, \boldsymbol{x_2}) p_{\theta'}(\boldsymbol{x_2} | y, \boldsymbol{x_1}) d\boldsymbol{x_2} \end{aligned}$$

On ne connaît pas la densité jointe de $\boldsymbol{Y}$, $\boldsymbol{X_1}$ et $\boldsymbol{X_2}$. En revanche, on connaît la densité de $\boldsymbol{Y}$ conditionnellement à $\boldsymbol{X_1}$ et $\boldsymbol{X_2}$. C'est pourquoi on cherche à s'y ramener.
Notons qu'on a supposé que la variable non observée $\boldsymbol{X_2}$ est indépendante des autres variables explicatives $\boldsymbol{X_1}$ On a également le fait que $\theta = (\boldsymbol{\beta}, \sigma)$. Comme la densité de $\boldsymbol{X_1}$ ne dépend pas de θ et que la densité de $\boldsymbol{X_2}$ ne dépend que de σ et pas de $\boldsymbol{\beta}$, on a alors :
d'une part :

$$-\log p_\theta(y, \boldsymbol{x_1}, \boldsymbol{x_2}) = -\log p_\theta(y | \boldsymbol{x_1}, \boldsymbol{x_2}) - \log p(\boldsymbol{x_1}) - \log p_\sigma(\boldsymbol{x_2})$$

et d'autre part :

$$\begin{aligned} p_{\theta'}(\boldsymbol{x_2} | y, \boldsymbol{x_1}) &= \frac{p_{\theta'}(y, \boldsymbol{x_1}, \boldsymbol{x_2})}{p_{\theta'}(y, \boldsymbol{x_1})} \\ &= \frac{p_{\theta'}(y, \boldsymbol{x_1}, \boldsymbol{x_2})}{\int p_{\theta'}(y, \boldsymbol{x_1}, \boldsymbol{x_2}) d\boldsymbol{x_2}} \\ &= \frac{p_{\theta'}(y | \boldsymbol{x_1}, \boldsymbol{x_2}) p_(\boldsymbol{x_1}) p_{\sigma'}(\boldsymbol{x_2})}{\int p_{\theta'}(y | \boldsymbol{x_1}, \boldsymbol{x_2}) p_(\boldsymbol{x_1}) p_{\sigma'}(\boldsymbol{x_2}) d\boldsymbol{x_2}} \end{aligned}$$

On oubliera par la suite la densité de probabilité de $\boldsymbol{X_1}$ qui n'entre pas en jeu dans l'optimisation. La densité de $\boldsymbol{X_2}$ est supposée normale $\mathcal{N}(0, \sigma^2)$.

$$p_\sigma(\boldsymbol{x_2}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\boldsymbol{x_2}^2}{2\sigma^2}} \tag{5.8}$$

On a :

$$p_{\theta'}(y|\boldsymbol{x_1}, \boldsymbol{x_2}) = \left(\frac{e^{\boldsymbol{x_1}'\boldsymbol{\beta'}+\boldsymbol{x_2}}}{1+e^{\boldsymbol{x_1}'\boldsymbol{\beta'}+\boldsymbol{x_2}}}\right)^y \left(\frac{1}{1+e^{\boldsymbol{x_1}'\boldsymbol{\beta'}+\boldsymbol{x_2}}}\right)^{1-y}$$

Après simplification, on obtient l'expression suivante :

$$-\log p_\theta(y|\boldsymbol{x_1}, \boldsymbol{x_2}) = -y(\boldsymbol{x_1}'\boldsymbol{\beta} + \boldsymbol{x_2}) + \log\left(1 + e^{\boldsymbol{x_1}'\boldsymbol{\beta}+\boldsymbol{x_2}}\right)$$

Pour alléger les calculs, on remarque que l'on est amené à calculer des quantités de la forme :

$$\begin{aligned} &\int \frac{(\boldsymbol{x_2})^q\, e^{-\frac{\boldsymbol{x_2}^2}{2\sigma'^2}}}{\sqrt{2\pi\sigma'^2}} \left(\frac{e^{\boldsymbol{x_1}'\boldsymbol{\beta'}+\boldsymbol{x_2}}}{1+e^{\boldsymbol{x_1}'\boldsymbol{\beta'}+\boldsymbol{x_2}}}\right)^y \left(\frac{1}{1+e^{\boldsymbol{x_1}'\boldsymbol{\beta'}+\boldsymbol{x_2}}}\right)^{1-y} d\boldsymbol{x_2} \\ = \quad & \mathbb{E}\left[\frac{u(\sigma')^q}{1+a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right] \end{aligned} \tag{5.9}$$

avec :
- $u(\sigma') \sim \mathcal{N}(0, \sigma'^2)$
- $q \in \{0, 1, 2\}$
- si $y = 1 : a(\boldsymbol{\beta'}) = e^{-\boldsymbol{x_1}'\boldsymbol{\beta'}}$ et $\alpha = -1$
- si $y = 0 : a(\boldsymbol{\beta'}) = e^{\boldsymbol{x_1}'\boldsymbol{\beta'}}$ et $\alpha = 1$

ainsi que des quantités de la forme :

$$\begin{aligned} &\int \frac{\log\left(1 + e^{\boldsymbol{x_1}'\boldsymbol{\beta}+\boldsymbol{x_2}}\right) e^{-\frac{\boldsymbol{x_2}^2}{2\sigma'^2}}}{\sqrt{2\pi\sigma'^2}} \left(\frac{e^{\boldsymbol{x_1}'\boldsymbol{\beta'}+\boldsymbol{x_2}}}{1+e^{\boldsymbol{x_1}'\boldsymbol{\beta'}+\boldsymbol{x_2}}}\right)^y \left(\frac{1}{1+e^{\boldsymbol{x_1}'\boldsymbol{\beta'}+\boldsymbol{x_2}}}\right)^{1-y} d\boldsymbol{x_2} \\ = \quad & \mathbb{E}\left[\frac{\log\left(1 + b(\boldsymbol{\beta})e^{u(\sigma')}\right)}{1+a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right] \end{aligned}$$

avec :
- $u(\sigma') \sim \mathcal{N}(0, \sigma'^2)$
- $b(\boldsymbol{\beta}) = e^{\boldsymbol{x_1}'\boldsymbol{\beta}}$
- si $y = 1 : a(\boldsymbol{\beta'}) = e^{-\boldsymbol{x_1}'\boldsymbol{\beta'}}$ et $\alpha = -1$
- si $y = 0 : a(\boldsymbol{\beta'}) = e^{\boldsymbol{x_1}'\boldsymbol{\beta'}}$ et $\alpha = 1$

On obtient l'expression de $Q(\theta, \theta')$ suivante :

$$\begin{aligned} Q(\theta,\theta') &= \sum_{i=1}^{n}\left(-y_i \boldsymbol{x_{1i}}'\boldsymbol{\beta} + \frac{1}{2}\log\sigma^2\right) \\ &+ \sum_{i=1}^{n}\left(\mathbb{E}\left[\frac{1}{1+a(\boldsymbol{\beta'})_i e^{\alpha_i u(\sigma')}}\right]\right)^{-1}\left\{-y_i\mathbb{E}\left[\frac{u(\sigma')}{1+a(\boldsymbol{\beta'})_i e^{\alpha_i u(\sigma')}}\right]\right. \\ &\left. + \frac{1}{2\sigma^2}\mathbb{E}\left[\frac{u(\sigma')^2}{1+a(\boldsymbol{\beta'})_i e^{\alpha_i u(\sigma')}}\right] + \mathbb{E}\left[\frac{\log\left(1+b(\boldsymbol{\beta})_i e^{u(\sigma')}\right)}{1+a(\boldsymbol{\beta'})_i e^{\alpha_i u(\sigma')}}\right]\right\} \end{aligned}$$

avec :
- $u(\sigma') \sim \mathcal{N}(0, \sigma'^2)$
- $b(\boldsymbol{\beta})_i = e^{\boldsymbol{x_{1i}}'\boldsymbol{\beta}}$
- si $y_i = 1 : a(\boldsymbol{\beta'})_i = e^{-\boldsymbol{x_{1i}}'\boldsymbol{\beta'}}$ et $\alpha_i = -1$
- si $y_i = 0 : a(\boldsymbol{\beta'})_i = e^{\boldsymbol{x_{1i}}'\boldsymbol{\beta'}}$ et $\alpha_i = 1$

Notons dès maintenant que, pour l'optimisation de $Q(\theta, \theta')$ en θ à θ' fixé, il est inutile de calculer tous les termes. En effet :

$$-y_i\mathbb{E}\left[\frac{u(\sigma')}{1+a(\boldsymbol{\beta'})_i e^{\alpha_i u(\sigma')}}\right] \quad \text{et} \quad \frac{1}{2\sigma^2}\mathbb{E}\left[\frac{u(\sigma')^2}{1+a(\boldsymbol{\beta'})_i e^{\alpha_i u(\sigma')}}\right]$$

ne dépendent pas de θ.

On note également que l'optimisation en σ peut se faire de manière indépendante de l'optimisation en $\boldsymbol{\beta}$ et ce, de manière analytique.

Dans la version pénalisée de l'algorithme EM, on contrôle la convergence non plus via la fonction de contraste $-\log L_n(\theta)$ mais via la fonction de contraste pénalisée qui a l'expression suivante : $-\log L_n(\theta) + \lambda J(\theta)$.

5.3.2 Calcul de la vraisemblance pénalisée

Comme on l'a vu, c'est la fonction de contraste pénalisée qui contrôle la convergence de l'algorithme EM. On détaille dans ce paragraphe comment on le calcule.

$$-\log L_n(\theta) = \sum_{i=1}^{n} -\log p_\theta(y_i, \boldsymbol{x_{1i}})$$

Le calcul de $-\log L_n(\theta)$ s'effectue donc de la manière suivante :

$$\begin{aligned} -\log L_n(\theta) &= \sum_{i=1}^{n} -\log \int p_\theta(y_i|\boldsymbol{x_{1i}}, \boldsymbol{x_{2i}}) p_\theta(\boldsymbol{x_{1i}}, \boldsymbol{x_{2i}}) d\boldsymbol{x_{2i}} \\ &= \sum_{i=1}^{n} -\log \int p_\theta(y_i|\boldsymbol{x_{1i}}, \boldsymbol{x_{2i}}) p_{\boldsymbol{\beta}}(\boldsymbol{x_{1i}}) p_\sigma(\boldsymbol{x_{2i}}) d\boldsymbol{x_{2i}} \end{aligned}$$

De même que pour l'expression de $Q(\theta, \theta')$, la densité de $\boldsymbol{X_1}$ n'entre pas en jeu dans l'algorithme d'optimisation. On connaît la densité de $\boldsymbol{Y}$ conditionnellement à $\boldsymbol{X_1}$ et $\boldsymbol{X_2}$. On a donc :

$$-\log L_n(\theta) =$$
$$\sum_{i=1}^{n} -\log \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\boldsymbol{x_{2i}}^2}{2\sigma^2}} \left(\frac{e^{\boldsymbol{\beta x_{1i}} + \boldsymbol{x_{2i}}}}{1 + e^{\boldsymbol{\beta x_{1i}} + \boldsymbol{x_{2i}}}} \right)^{y_i} \left(\frac{1}{1 + e^{\boldsymbol{\beta x_{1i}} + \boldsymbol{x_{2i}}}} \right)^{1-y_i} d\boldsymbol{x_{2i}}$$

et avec les notations adoptées en (5.9) :

$$-\log L_n(\theta) + \lambda J(\theta) = \sum_{i=1}^{n} -\log \mathbb{E}\left[\frac{1}{1 + a(\boldsymbol{\beta})_i e^{\alpha_i u(\sigma)}} \right] + \lambda \left(||\boldsymbol{\beta}||_1 + \sigma \right)$$

avec :
- $u(\sigma) \sim \mathcal{N}(0, \sigma^2)$
- si $y_i = 1 : a(\boldsymbol{\beta})_i = e^{-\boldsymbol{x_{1i}}'\boldsymbol{\beta}}$ et $\alpha_i = -1$
- si $y_i = 0 : a(\boldsymbol{\beta})_i = e^{\boldsymbol{x_{1i}}'\boldsymbol{\beta}}$ et $\alpha_i = 1$

5.3.3 Approximations

On rappelle que l'optimisation en σ peut être traitée analytiquement. C'est l'objet du paragraphe suivant. En revanche, l'optimisation en $\boldsymbol{\beta}$ est numérique et nécessite le calcul des espérances suivantes :

$$\mathbb{E}\left[\frac{1}{1 + a(\boldsymbol{\beta'}) e^{\alpha u(\sigma')}} \right], \mathbb{E}\left[\frac{u(\sigma')}{1 + a(\boldsymbol{\beta'}) e^{\alpha u(\sigma')}} \right], \mathbb{E}\left[\frac{u(\sigma')^2}{1 + a(\boldsymbol{\beta'}) e^{\alpha u(\sigma')}} \right]$$

et

$$\mathbb{E}\left[\frac{\log\left(1 + b(\boldsymbol{\beta}) e^{u(\sigma')}\right)}{1 + a(\boldsymbol{\beta'}) e^{\alpha u(\sigma')}} \right]$$

pour $u(\sigma') \sim \mathcal{N}(0, \sigma'^2)$, $\alpha \in \{-1, 1\}$, $a(\boldsymbol{\beta'}) \in \mathbb{R}$ et $b(\boldsymbol{\beta}) \in \mathbb{R}$.

Nous utilisons des procédures numériques d'intégration pour calculer ces espérances sauf pour certaines valeurs des paramètres pour lesquelles des approximations sont obtenues assez facilement.

5.3.3.1 Premier type d'approximations

Les premières approximations concernent les quantités de la forme :

$$\mathbb{E}\left[\frac{1}{1+a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right]$$

avec :
- $u(\sigma') \sim \mathcal{N}(0, \sigma'^2)$
- si $y = 1 : a(\boldsymbol{\beta'}) = e^{-\boldsymbol{x_1}'\boldsymbol{\beta'}}$ et $\alpha = -1$
- si $y = 0 : a(\boldsymbol{\beta'}) = e^{\boldsymbol{x_1}'\boldsymbol{\beta'}}$ et $\alpha = 1$

On fait les approximations suivantes :

- si $a(\boldsymbol{\beta'})e^{2\sigma'} \ll 1$ $\left(a(\boldsymbol{\beta'})e^{2\sigma'} < \varepsilon \text{ avec par exemple } \varepsilon = 10^{-4}\right)$ alors :

$$\mathbb{E}\left[\frac{1}{1+a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right] \simeq 1$$

- si $a(\boldsymbol{\beta'})e^{-2\sigma'} \gg 1$ $\left(a(\boldsymbol{\beta'})e^{-2\sigma'} < \varepsilon^{-1} \text{ avec par exemple } \varepsilon = 10^{-4}\right)$ alors :

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{1+a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right] &\simeq \frac{1}{a(\boldsymbol{\beta'})}\mathbb{E}\left[e^{-\alpha u(\sigma')}\right] \\
&= \frac{1}{a(\boldsymbol{\beta'})}\int \frac{1}{\sqrt{2\pi\sigma'^2}} e^{-\alpha u - \frac{u^2}{2\sigma'^2}} du \\
&= \frac{1}{a(\boldsymbol{\beta'})}\int \frac{1}{\sqrt{2\pi\sigma'^2}} e^{-\left(\frac{u}{\sqrt{2\sigma'^2}}+\frac{\alpha\sigma'}{2}\right)^2+\frac{\sigma'^2}{2}} du
\end{aligned}$$

On pose $\frac{v}{\sqrt{2}} = \frac{u}{\sqrt{2\sigma'^2}} + \frac{\alpha\sigma'}{2}$, avec $dv = \frac{du}{\sigma'}$, et on a :

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{1+a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right] &\simeq \frac{e^{\frac{\sigma'^2}{2}}}{a(\boldsymbol{\beta'})}\int \frac{1}{\sqrt{2\pi\sigma'^2}} e^{-\frac{v^2}{2}} \sigma' dv \\
&= \frac{e^{\frac{\sigma'^2}{2}}}{a(\boldsymbol{\beta'})}
\end{aligned}$$

5.3.3.2 Deuxième type d'approximations

Dans l'expression de $Q(\theta, \theta')$, on utilise également des approximations pour les quantités de la forme :

$$\mathbb{E}\left[\frac{\log\left(1+b(\boldsymbol{\beta})e^{u(\sigma')}\right)}{1+a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right]$$

avec :
- $u(\sigma') \sim \mathcal{N}(0, \sigma'^2)$
- $b(\boldsymbol{\beta}) = e^{\boldsymbol{x_1}'\boldsymbol{\beta}}$
- si $y = 1 : a(\boldsymbol{\beta'}) = e^{-\boldsymbol{x_1}'\boldsymbol{\beta'}}$ et $\alpha = -1$
- si $y = 0 : a(\boldsymbol{\beta'}) = e^{\boldsymbol{x_1}'\boldsymbol{\beta'}}$ et $\alpha = 1$

On fait les approximations suivantes :
- si $b(\boldsymbol{\beta})e^{2\sigma'} \ll 1$ $\left(b(\boldsymbol{\beta})e^{2\sigma'} < \varepsilon \text{ avec par exemple } \varepsilon = 10^{-4}\right)$ alors :

$$\mathbb{E}\left[\frac{\log\left(1 + b(\boldsymbol{\beta})e^{u(\sigma')}\right)}{1 + a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right] \simeq 0$$

- si $b(\boldsymbol{\beta})e^{-2\sigma'} \gg 1$ $\left(b(\boldsymbol{\beta})e^{-2\sigma'} > \varepsilon^{-1} \text{ avec par exemple } \varepsilon = 10^{-4}\right)$ alors :

$$\begin{aligned}\mathbb{E}\left[\frac{\log\left(1 + b(\boldsymbol{\beta})e^{u(\sigma')}\right)}{1 + a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right] \simeq\ & \log\left[b(\boldsymbol{\beta})\right]\mathbb{E}\left[\frac{1}{1 + a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right] \\ & + \mathbb{E}\left[\frac{u(\sigma')}{1 + a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right]\end{aligned}$$

5.3.3.3 Troisième type d'approximations

On peut également utiliser des approximations pour les quantités de la forme :

$$\mathbb{E}\left[\frac{u(\sigma')}{1 + a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right]$$

En effet, on fait les approximations suivantes :
- si $a(\boldsymbol{\beta'})e^{2\sigma'} \ll 1$ $\left(a(\boldsymbol{\beta'})e^{2\sigma'} < \varepsilon \text{ avec par exemple } \varepsilon = 10^{-4}\right)$ alors :

$$\mathbb{E}\left[\frac{u(\sigma')}{1 + a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right] \simeq \mathbb{E}[u(\sigma')] = 0$$

– si $a(\boldsymbol{\beta'})e^{-2\sigma'} \gg 1$ $\left(a(\boldsymbol{\beta'})e^{-2\sigma'} < \varepsilon^{-1}\right.$ avec par exemple $\left.\varepsilon = 10^{-4}\right)$ alors :

$$\begin{aligned}
\mathbb{E}\left[\frac{u(\sigma')}{1+a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right] &\simeq \frac{1}{a(\boldsymbol{\beta'})}\mathbb{E}\left[u(\sigma')e^{-\alpha u(\sigma')}\right] \\
&= \frac{1}{a(\boldsymbol{\beta'})}\int \frac{1}{\sqrt{2\pi\sigma'^2}} u e^{-\alpha u - \frac{u^2}{2\sigma'^2}} du \\
&= \frac{1}{a(\boldsymbol{\beta'})}\int \frac{1}{\sqrt{2\pi\sigma'^2}} u e^{-\left(\frac{u}{\sqrt{2\sigma'^2}}+\frac{\alpha\sigma'}{2}\right)^2+\frac{\sigma'^2}{2}} du
\end{aligned}$$

On pose $\frac{v}{\sqrt{2}} = \frac{u}{\sqrt{2\sigma'^2}} + \frac{\alpha\sigma'}{2}$, avec $dv = \frac{du}{\sigma'}$. On a également $u = \sigma'(v - \alpha\sigma')$. On obtient :

$$\begin{aligned}
\mathbb{E}\left[\frac{u(\sigma')}{1+a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right] &\simeq \frac{e^{\frac{\sigma'^2}{2}}}{a(\boldsymbol{\beta'})}\left(-\alpha\sigma'^2 + \int \frac{1}{\sqrt{2\pi\sigma'^2}}\sigma' v e^{-\frac{v^2}{2}}\sigma' dv\right) \\
&= -\alpha\sigma'^2 \frac{e^{\frac{\sigma'^2}{2}}}{a(\boldsymbol{\beta'})}
\end{aligned}$$

car la fonction $v \mapsto v e^{-\frac{v^2}{2}}$ est impaire sur $\mathbb{R}$.

5.3.3.4 Quatrième type d'approximations

Pour finir, on anticipe sur le paragraphe qui suit, en proposant également des approximations pour les quantités de la forme :

$$\mathbb{E}\left[\frac{u(\sigma')^2}{1+a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right]$$

Ces quantités entrent en jue dans l'optimisation en σ. On fait les approximations suivantes :

– si $a(\boldsymbol{\beta'})e^{2\sigma'} \ll 1$ $\left(a(\boldsymbol{\beta'})e^{2\sigma'} < \varepsilon\right.$ avec par exemple $\left.\varepsilon = 10^{-4}\right)$ alors :

$$\mathbb{E}\left[\frac{u(\sigma')^2}{1+a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right] \simeq \mathbb{E}[u(\sigma')^2] = \sigma'^2$$

En effet :

$$\mathbb{E}[u(\sigma')^2] = \int \frac{1}{\sqrt{2\pi\sigma'^2}} u^2 e^{-\frac{u^2}{2\sigma'^2}} du$$

On réalise une intégration par partie, en posant :

$$r'(u) = u e^{-\frac{u^2}{2\sigma'^2}} \text{ et } s(u) = u$$

On a alors :

$$r(u) = -{\sigma'}^2 e^{-\frac{u^2}{2{\sigma'}^2}} \text{ et } s'(u) = 1$$

On a donc :

$$\begin{aligned}
\mathbb{E}[u(\sigma')^2] &= \int \frac{1}{\sqrt{2\pi{\sigma'}^2}} u^2 e^{-\frac{u^2}{2{\sigma'}^2}} du \\
&= \left[\frac{1}{\sqrt{2\pi{\sigma'}^2}} r(u)s(u)\right]_{-\infty}^{+\infty} - \int \frac{1}{\sqrt{2\pi{\sigma'}^2}} r(u)s'(u)du \\
&= 0 - \int \frac{1}{\sqrt{2\pi{\sigma'}^2}} \left(-{\sigma'}^2 e^{-\frac{u^2}{2{\sigma'}^2}}\right) du \\
&= {\sigma'}^2
\end{aligned}$$

– si $a(\boldsymbol{\beta'})e^{-2\sigma'} \gg 1$ $\left(a(\boldsymbol{\beta'})e^{-2\sigma'} < \varepsilon^{-1} \text{ avec par exemple } \varepsilon = 10^{-4}\right)$ alors :

$$\begin{aligned}
\mathbb{E}\left[\frac{u(\sigma')^2}{1 + a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right] &\simeq \frac{1}{a(\boldsymbol{\beta'})} \mathbb{E}\left[u(\sigma')^2 e^{-\alpha u(\sigma')}\right] \\
&= \frac{1}{a(\boldsymbol{\beta'})} \int \frac{1}{\sqrt{2\pi{\sigma'}^2}} u^2 e^{-\alpha u - \frac{u^2}{2{\sigma'}^2}} du \\
&= \frac{1}{a(\boldsymbol{\beta'})} \int \frac{1}{\sqrt{2\pi{\sigma'}^2}} u^2 e^{-\left(\frac{u}{\sqrt{2{\sigma'}^2}} + \frac{\alpha\sigma'}{2}\right)^2 + \frac{{\sigma'}^2}{2}} du
\end{aligned}$$

On pose $\frac{v}{\sqrt{2}} = \frac{u}{\sqrt{2{\sigma'}^2}} + \frac{\alpha\sigma'}{2}$, soit encore $v = \frac{u}{\sigma'} + \alpha\sigma'$. Il vient : $dv = \frac{du}{\sigma'}$. On a également $u = \sigma'(v - \alpha\sigma')$. On obtient :

$$\begin{aligned}
\mathbb{E}\left[\frac{u(\sigma')^2}{1 + a(\boldsymbol{\beta'})e^{\alpha u(\sigma')}}\right] &\simeq \frac{1}{a(\boldsymbol{\beta'})} \int \frac{1}{\sqrt{2\pi{\sigma'}^2}} {\sigma'}^2 (v - \alpha\sigma')^2 e^{-\frac{v^2}{2}} \sigma' dv \\
&= \frac{1}{a(\boldsymbol{\beta'})} \left\{ \int \frac{1}{\sqrt{2\pi}} {\sigma'}^2 v^2 e^{-\frac{v^2}{2}} dv \right. \\
&\quad - \left(2\alpha{\sigma'}^3\right) \int \frac{1}{\sqrt{2\pi}} v e^{-\frac{v^2}{2}} dv \\
&\quad \left. + \left(\alpha^2{\sigma'}^4\right) \int \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} dv \right\} \\
&= \frac{{\sigma'}^2\left(1 + {\sigma'}^2\right)}{a(\boldsymbol{\beta'})}
\end{aligned}$$

car, comme $\alpha = \pm 1$, $\alpha^2 = 1$.

5.3.4 Résolution analytique de l'optimisation en σ

La minimisation de $Q((\boldsymbol{\beta}, \sigma), \theta') + \lambda J(\boldsymbol{\beta}, \sigma)$ en σ à $\boldsymbol{\beta}$ et θ' fixés peut se faire de manière analytique. En effet, si l'on fixe à $\boldsymbol{\beta}$ et θ', on a :

$$Q((\boldsymbol{\beta}, \sigma), \theta') + \lambda \left(\sum_{j=1}^{p} |\beta_j| + \sigma \right) = \sum_{i=1}^{n} \frac{1}{2} \log \sigma^2 + \frac{A}{2\sigma^2} + \lambda\sigma + C^{te}$$

où C^{te} est une constante ne dépendant pas de σ
et $A = \sum_{i=1}^{n} \mathbb{E}\left[\frac{u(\sigma')^2}{1+a(\boldsymbol{\beta'})_i e^{\alpha_i u(\sigma')}} \right] \mathbb{E}\left[\frac{1}{1+a(\boldsymbol{\beta'})_i e^{\alpha_i u(\sigma')}} \right]^{-1}$. Notons que $A > 0$ pour $\sigma > 0$ et que $\sigma = 0 \implies A = 0$, car alors $\boldsymbol{X_2} = 0$ presque sûrement.

σ étant une variance, on définit la fonction suivante pour $\sigma > 0$:

$$f(\sigma) = n \log \sigma + \frac{A}{2\sigma^2} + \lambda\sigma$$

On cherche alors :

$$\sigma^{new} = \arg\min_{\sigma > 0} f(\sigma)$$

5.3.4.1 Cas particulier

Si $\lambda = 0$, alors l'expression se simplifie et on a ($\sigma > 0$) :

$$f'(\sigma) = 0 \quad \iff \quad \frac{n}{\sigma} - \frac{A}{\sigma^3} = 0$$

On a donc :

$$\sigma^{new} = \sqrt{\frac{A}{n}} \tag{5.10}$$

On suppose dans la suite que $\lambda > 0$.

5.3.4.2 Etude de variations

On calcule pour quelles valeurs de σ sa dérivée s'annule.

$$f'(\sigma) = 0 \iff \frac{n}{\sigma} - \frac{A}{\sigma^3} + \lambda = 0$$

Pour l'étude des zéros de la dérivée, on définit la fonction g sur $\mathbb{R}$ tout entier.

$$g(\sigma) = \sigma^3 f'(\sigma) = \lambda\sigma^3 + n\sigma^2 - A$$

Tout d'abord, $\sigma = 0$ n'est pas un zéro de g. Donc on peut affirmer que les zéros de g sont exactement les zéros de f'.

On étudie les variations de g sur $\mathbb{R}$ tout en sachant qu'on s'intéresse aux zéros strictement positifs.

$$\begin{aligned} g(\sigma) &= \lambda\sigma^3 + n\sigma^2 - A \\ g'(\sigma) &= 3\lambda\sigma^2 + 2n\sigma \\ g''(\sigma) &= 6\lambda\sigma + 2n \end{aligned}$$

On a donc :

σ	$-\infty$		$-\frac{2n}{6\lambda}$		$+\infty$
$g''(\sigma)$		$-$		$+$	
$g'(\sigma)$	$+\infty$	$\searrow$	$-\frac{n^2}{3\lambda}$	$\nearrow$	$+\infty$

et :

σ	$-\infty$		$-\frac{2n}{3\lambda}$		0		$+\infty$
$g'(\sigma)$		$+$		$-$		$+$	
$g(\sigma)$	$-\infty$	$\nearrow$	$\frac{4n^3}{27\lambda^2} - A$	$\searrow$	$-A$	$\nearrow$	$+\infty$

Comme $A > 0$, g a exactement un zéro positif. De plus, on note que :

– si $\frac{4n^3}{27\lambda^2} - A < 0$, l'équation $g(\sigma) = 0$ possède une unique solution strictement positive sur $\mathbb{R}$.

– si $\frac{4n^3}{27\lambda^2} - A = 0$, l'équation $g(\sigma) = 0$ possède trois solutions dont une solution strictement positive et une solution double strictement négative.

– si $\frac{4n^3}{27\lambda^2} - A > 0$, l'équation $g(\sigma) = 0$ possède trois solutions dont une solution strictement positive et deux solutions strictement négatives.

Résolvons l'équation : $g(\sigma) = \lambda\sigma^3 + n\sigma^2 - A = 0$ On note tout d'abord que si $\lambda = 0$ alors :

$$\sigma^{new} = \sqrt{A/n} \tag{5.11}$$

Supposons maintenant que $\lambda \neq 0$.

$$g(\sigma) = 0 \iff \sigma^3 + \frac{n}{\lambda}\sigma^2 - \frac{A}{\lambda} = 0$$

On fait le changement de variables suivant :

$$\sigma = z - \frac{n}{3\lambda} \tag{5.12}$$

On obtient alors :

$$z^3 + pz + q = 0 \tag{5.13}$$

avec $(p, q) = \left(-\frac{n^2}{3\lambda^2}, \frac{2n^3}{27\lambda^3} - \frac{A}{\lambda}\right)$.

5.3.4.3 Méthode de Cardan

Pour résoudre cette équation, on utilise la méthode de Cardan de résolution d'équations du troisième degré. On pose alors : $z = u + v$.

On introduit une variable supplémentaire. En procédant de la sorte, on augmente artificiellement le nombre de degrés de liberté. On se laisse ainsi la possibilité d'imposer une relation entre les variables, afin de revenir au nombre de degrés de liberté initial. Il vient :

$$u^3 + v^3 + (u + v)(3uv + p) + q = 0$$

On choisit donc d'imposer :

$$3uv + p = 0 \tag{5.14}$$

On a alors :

$$\begin{aligned} z^3 + pz + q = 0 \quad &\iff \quad \left\{ \begin{array}{l} u^3 + v^3 + q = 0 \\ 3uv + p = 0 \end{array} \right. \\ &\iff \quad \left\{ \begin{array}{l} u^3 + v^3 + q = 0 \\ u^3 v^3 = -p^3/27 \end{array} \right. \end{aligned}$$

à condition que u et v soient les racines cubiques de u^3 et v^3 telles que $uv = -p/3$.

On connaît la somme et le produit de u^3 et v^3. Ce sont donc les deux racines de l'équation du second degré suivante :

$$t^2 + qt - \frac{p^3}{27} = 0 \tag{5.15}$$

Le déterminant vaut : $\Delta = q^2 + \frac{4p^3}{27} = \frac{A}{\lambda^2}\left(A - \frac{4n^3}{27\lambda^2}\right)$.

– si $\Delta > 0$, alors l'équation (5.15) a deux solutions réelles :

$$(t_1, t_2) = \left(\frac{-q - \sqrt{\Delta}}{2}, \frac{-q + \sqrt{\Delta}}{2}\right)$$

– si $\Delta < 0$, alors l'équation (5.15) a deux solutions complexes conjuguées :

$$(t_1, t_2) = \left(\frac{-q - i\sqrt{-\Delta}}{2}, \frac{-q + i\sqrt{-\Delta}}{2}\right)$$

– si $\Delta = 0$, alors l'équation (5.15) a une solution réelle double :

$$t_1 = t_2 = -\frac{q}{2}$$

Si u et v sont les racines cubiques de t_1 et t_2 telles que $uv = -p/3$, alors les racines de l'équation (5.13) sont :

$$\left(u + v; uj + vj^2; uj^2 + vj\right) \text{ avec } j = e^{\frac{2i\pi}{3}}$$

On remarque que, comme λ et A sont des quantités stritement positives, le signe de Δ est le même que celui de $\left(A - \frac{4n^3}{27\lambda^2}\right)$, qui est une valeur déjà mise en évidence dans l'étude des variations de la fonction $g(\sigma)$.

On récapitule ici :

– Le cas $\Delta > 0$ correspond au cas où $\frac{4n^3}{27\lambda^2} - A < 0$.
L'équation $g(\sigma) = 0$ possède une unique solution strictement positive.
t_1 et t_2 sont des valeurs réelles. Donc si u et v sont les racines cubiques réelles respectives de t_1 et t_2 alors $uv = -p/3$ est vérifiée. Et on a bien une racine réelle unique : $u + v$ et deux racines complexes conjuguées $uj + vj^2$ et $uj^2 + vj$. Donc, après la translation introduite dans l'équation (5.12) :

$$\sigma^{new} = u + v - \frac{n}{3\lambda}$$

– Le cas $\Delta < 0$ correspond au cas où $\frac{4n^3}{27\lambda^2} - A > 0$.
L'équation $g(\sigma) = 0$ possède trois solutions dont une solution strictement

positive et deux solutions strictement négatives. t_1 et t_2 sont des valeurs complexes conjuguées. On prend les racines cubiques de t_1 et t_2, respectivement u et v telles que $uv = -p/3$ est vérifiée. Alors $u+v$, $uj+vj^2$ et uj^2+vj sont toutes trois réelles. Une seule est strictement positive. Ainsi, après la translation :

$$\sigma^{new} = \max\left(u+v, uj+vj^2, uj^2+vj\right) - \frac{n}{3\lambda} \text{ avec } j = e^{\frac{2i\pi}{3}}$$

– Le cas $\Delta = 0$ correspond au cas où $\frac{4n^3}{27\lambda^2} - A = 0$.
Pour ce cas, on a une expression explicite de la solution.

$$\Delta = \frac{A}{\lambda^2}\left(A - \frac{4n^3}{27\lambda^2}\right) = 0 \iff A = \frac{4n^3}{27\lambda^2}$$

$$q = \frac{2n^3}{27\lambda^3} - \frac{A}{\lambda} = \frac{2n^3}{27\lambda^3} - \frac{4n^3}{27\lambda^3} = -\frac{2n^3}{27\lambda^3} \neq 0$$

La solution de l'équation $g(\sigma) = 0$ n'est donc pas triple, mais on a bien trois solutions dont une solution strictement positive et une solution double strictement négative.

$$t_1 = t_2 = -\frac{q}{2} = \frac{n^3}{27\lambda^3} \tag{5.16}$$

On a donc l'expression exacte des racines cubiques :

$$(u, v) = (n/3\lambda, n/3\lambda) \tag{5.17}$$

On vérifie que l'on a bien :

$$uv = (n/3\lambda)^2 = -p/3 \tag{5.18}$$

Et finalement, on a, en tenant compte de la propriété des racines cubiques de l'unité $\left(j + j^2 = -1\right)$:

$$\left(u+v; uj+vj^2; uj^2+vj\right) = (2n/3\lambda; -n/3\lambda; -n/3\lambda) \tag{5.19}$$

L'unique solution strictement positive est donc, après translation :

$$\sigma^{new} = \frac{n}{3\lambda}$$

5.4 Chemin de régularisation en présence de variables cachées

Ce qui a été décrit ci-dessus permet de minimiser la fonction de contraste à laquelle on ajoute la fonction de pénalisation telle qu'elle est exprimée dans l'équation (5.2) pour un λ donné.

Pour ce λ, la fonction de contraste pénalisée est calculé grâce à l'algorithme EM décrit dans la partie 5.2, en prenant en compte la présence de variables cachées (cas où $\sigma > 0$). On minimise par ailleurs la fonction de contraste pénalisée en l'absence de variables cachées (cas où $\sigma = 0$ où l'on optimise simplement en $\boldsymbol{\beta}$, comme c'est le cas dans la partie 2). On compare les deux quantités obtenues. Si la quantité obtenue avec $\sigma > 0$ est plus petite que celle avec $\sigma = 0$, alors le modèle avec variables cachées devient plus explicatif que le modèle sans variables cachées. L'algorithme que l'on propose va donc consister à faire varier λ pour obtenir un chemin de régularisation et à positionner ensuite sur celui-ci le moment où la quantité obtenue avec $\sigma > 0$ devient plus petite que celle avec $\sigma = 0$.

Le problème majeur de l'algorithme EM est son initialisation. Dans le cas sans variables cachées, on sait où commencer la construction du chemin de régularisation. En effet, on sait que pour $\lambda > \lambda_{max}$, la pénalisation force tous les coefficients $(\beta_j)_{j=1,\dots,p}$ à valoir zéro. L'introduction de variables cachées ne modifie pas ce phénomène : il existe bien un $\widetilde{\lambda}_{max}$ au-delà duquel tous les coefficients $(\beta_j)_{j=1,\dots,p}$ et σ sont contraints à zéro. Cependant, on ne dispose plus d'expression analytique pour cette borne. On propose donc de procéder comme suit :

- On part de $\lambda = \lambda^0_{max}$ où λ^0_{max} est calculé en l'absence de variables cachées. Pour cette valeur de λ, on prend $\boldsymbol{\beta}^{init} = 0$. Initialiser σ par $\sigma^{init} = 0$ n'est pas judicieux. En effet, l'algorithme EM n'est utilisé que dans le cas où il y a des variables cachées. Comme il s'agit de régler l'initialisation d'un paramètre monodimensionnel, on se donne la possibilité de calculer la fonction de contraste pénalisée pour une grille de, par exemple, 10 valeurs $(\sigma_k)_{k=1,\dots,10}$ et de prendre pour σ^{init} celui qui donne la plus petite valeur :

$$\sigma^{init} = \arg\min_k -\log L_n(\boldsymbol{\beta}^{init}, \sigma_k) + \lambda J(\boldsymbol{\beta}^{init}, \sigma_k)$$

- si la fonction de contraste pénalisée avec variables cachées est plus grande que la quantité sans variables cachées dès cette première étape, alors on augmente λ d'un $\delta\lambda$, jusqu'à se trouver dans le cas inverse. On sera alors dans le cas où le modèle sans variables cachées explique mieux que le modèle avec variables cachées. La valeur atteinte par λ est alors notée $\widetilde{\lambda}_{max}$. Cette notation est légitime car pour $\lambda > \widetilde{\lambda}_{max}$ tous les coefficients sont à zéros (σ compris).

– on peut alors débuter l'algorithme, en récrivant :

$$\lambda_{max} = \max(\lambda^0_{max}, \widetilde{\lambda}_{max})$$

Une fois λ_{max} déterminé, le chemin de régularisation s'obtient en faisant décroître λ par pas de $\delta\lambda$ de λ_{max} jusqu'à 0. A chaque étape, on calcule les fonctions de contraste pénalisées pour $\sigma = 0$ et pour $\sigma \neq 0$. A chaque étape, la question de l'initialisation de l'EM se pose. Nous proposons deux stratégies d'initialisation.

Initialisation itérative Cette stratégie consiste à initialiser les paramètres $\boldsymbol{\beta}^{init}_{(k)}$ et pour $\sigma^{init}_{(k)}$ de l'étape k de l'EM par les valeurs finales obtenues à l'étape précédente.

Initialisation aux vraies valeurs des paramètres Cette stratégie est possible dans notre cas, du fait que nous travaillons sur des données simulées. On a donc accès aux vraies valeurs pour ces paramètres. Cette stratégie nous permet d'avoir un comparatif auquel confronter la première stratégie d'initialisation.

5.5 Résultats

Nous présentons le résultat obtenu pour chacune des deux stratégies d'initialisation sur un jeu de données simulées comportant 30 essais. La variable réponse a été calculée sur 9 variables explicatives. L'ensemble des variables observées compte 5 de ces 9 variables dont une correspond à un coefficient β_j nul. La taille des données est limitée par les temps de calcul qui sont assez importants. La convergence de l'algorithme EM, pour chaque λ, est réglé par un nombre fixe d'itérations, à savoir 35 itérations, sur cet exemple. On a vérifié que la fonction de contraste pénalisée converge effectivement au cours de ces 35 itérations.

On présente en figure 5.2 le chemin de régularisation obtenu sur ce jeu de données. Le chemin de régularisation de la figure 5.2 présente l'évolution des β_j en fonction de λ.

La figure 5.3 présente les fonctions de contraste pénalisées dans les trois cas suivants :

courbe rouge : sans variables cachées,

courbe noire : avec variables cachées et initialisation itérative,

courbe bleue : avec variables cachées et initialisation aux vraies valeurs des paramètres.

On note tout d'abord que notre stratégie d'initialisation dite itérative est meilleure que l'initialisation aux vraies valeurs des paramètres, dès que λ n'est pas dans un voisinage de 0. Initialiser aux vraies valeurs n'aide pas l'algorithme EM à trouver le

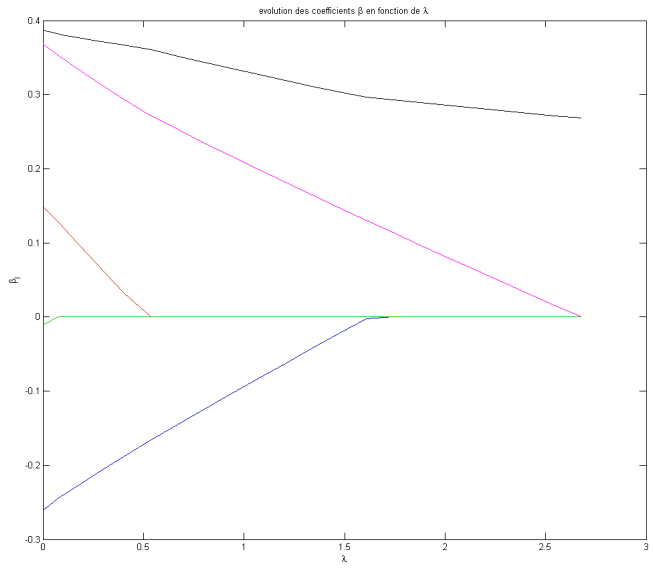

FIG. 5.2 – Chemin de régularisation correspondant à l'exemple contenant des variables cachées.

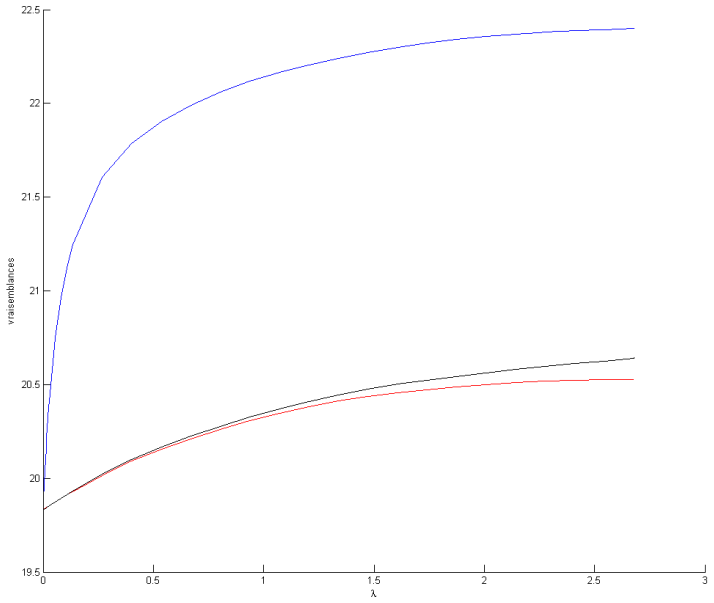

FIG. 5.3 – Fonctions de contraste pénalisées sans variables cachées (courbe rouge), avec variables cachées et initialisation itérative (courbe noire) et avec variables cachées et initialisation aux vraies valeurs des paramètres (courbe bleue).

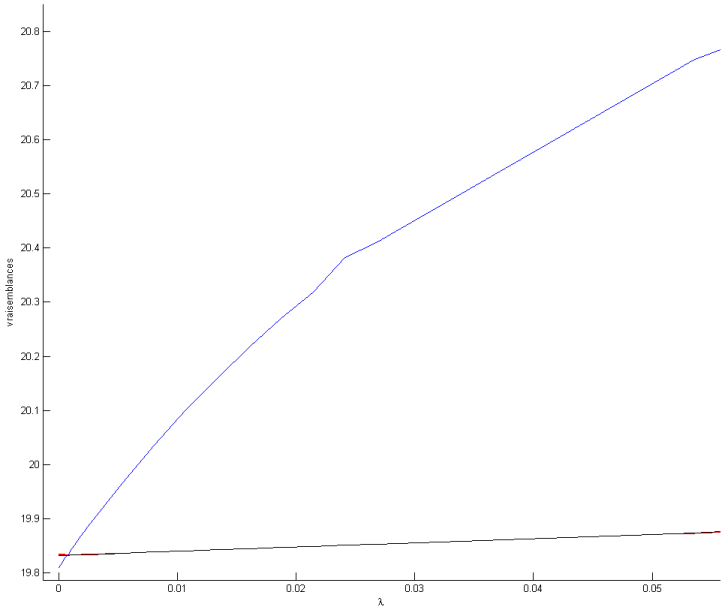

FIG. 5.4 – Zoom au voisinage de zéro de la figure 5.3. Les courbes sont toujours les fonctions de contraste pénalisées sans variables cachées (courbe rouge), avec variables cachées et initialisation itérative (courbe noire) et avec variables cachées et initialisation aux vraies valeurs des paramètres (courbe bleue).

minimum global de la fonction de contraste pénalisée : l'algorithme EM converge dans ce cas vers des minima locaux.

L'autre résultat intéressant s'observe au voisinage de 0. On peut voir sur la figure 5.4 que c'est cette fois l'initialisation aux vraies valeurs des paramètres qui donne le meilleur score. Cela signifie notamment que la stratégie proposée dite itérative n'est pas judicieuse si l'on cherche une "bonne valeur" à donner aux paramètres dans un problème de type estimation non-pénalisée.

Enfin, on peut noter que le modèle avec variables cachées devient plus explicatif que le modèle sans variables cachées pour λ proche de zéro, et ce quelle que soit la stratégie d'initialisation adoptée. Le fait que la variable cachée sorte après toutes les autres variables, y compris la variable ayant un coefficient β_j nul, peut paraître surprenant. On interprète ce résultat en notant que nous avons pénalisé de la même manière les variables observées et les variables cachées dans l'équation(5.2). En réalité, il est probable que les coefficients β_j aient une variance asymptotique bien plus faible que le paramètre σ de la variable cachée, ce qui devrait influencer

fortement le chemin de régularisation. Il serait intéressant d'utiliser les résultats asymptotiques de la partie 3 pour proposer une pondération plus adaptée entre les pénalisations des coefficients correspondants aux variables observées et ceux correspondants aux variables cachées.

Notons enfin que ces résultats asymptotiques devraient en théorie permettre de construire un test de l'hypothèse $H_0 : \{\sigma = 0\}$ contre $H_1 : \{\sigma > 0\}$ en utilisant le chemin de régularisation. Néanmoins, à l'heure actuelle, d'un point de vue pratique, ce type de méthode nécessite avant toute chose de trouver une manière efficace d'initialiser l'algorithme EM.

Chapitre 6

Annexes

6.1 Article paru le journal CSBIGS

CS-BIGS 1(2): 164-172
http://www.bentley.edu/csbigs/vol1-2/germain.pdf

Pampering the Client: Calibrating Vehicle Parts to Satisfy Customers

Jean-François Germain
Ecole Nationale Supérieure des Télécommunications, France

We present in this paper a statistical methodology to address the following industrial problem. Car manufacturers have to calibrate their vehicles in order to reach a level of quality which is acceptable to the customer. We consider here the specific case of a gear-box. Our study relies on a dataset consisting of evaluations by 507 testers of 28 configurations, each described by 12 physical parameters. We suggest a procedure for selecting and calibrating the physical parameters which have an impact on the evaluations. Our procedure consists of two steps. We first compute the regularization path of an L_1 – penalized logistic likelihood from which we extract an increasing sequence of models. In the second step of our procedure, we apply the BIC criterion to select a model in the sequence obtained in the first step. We provide a simple numerical procedure for this approach and discuss its application to the data. This article is accessible to readers with at least an intermediate knowledge of statistics; previous exposure to logistic regression and the principles of model selection would be useful, although not strictly necessary.

Description of our industrial case

One of the most important activities for a car manufacturer is to calibrate vehicle parts. The calibration of some components is driven by levels of customer satisfaction regarding issues such as drivability, habitability, acoustics, ergonomics, etc.

However, this implies that engineers know how to calibrate physical parameters in order to reach a given quality level which is satisfactory enough for the customer. The question is of how to link those qualitative customer evaluations and the quantitative physical design characteristics of the vehicle.

To address this question, we will present a new complete methodology, which handles subjective evaluation and design parameters together. Our modeling of customers' appreciation yields:

- the selection of design parameters that are explanatory, which means that they have an impact on customers' subjective evaluation.
- Some help with the calibration of the selected design parameters.

The industrial application field is the evaluation of gear-boxes. The dataset we tested the methodology on is described below.

Data

The dataset consists of two types of data:

Subjective data are subjective evaluations. Testers, selected for their high sensitivity to issues related to shifting gears, evaluated several different gear-boxes. The goal of the study is to determine what is or is not acceptable for the customer. For the sake of confidentiality we simplify this evaluation procedure as follows: each evaluation is one or zero.

Objective data are physical design parameters. Twenty-eight different gear-boxes are chosen: they are representative enough to allow for a robust analysis of subjective evaluations. All testers evaluate each gear-box several times, so that we globally handle a dataset of 507 observations.

Gear-boxes for which testers proceed with subjective evaluations are all measured in the same way. As they attempt to link subjective evaluations and physical parameters together, engineering experts know or have an idea of which design parameters are of interest. Those experts selected and extracted twelve potentially explanatory parameters from each measured signal.

The dataset used in the analysis is available in the accompanying MS Excel file. A row is one of the 507 samples. A column is one of the twelve design parameters. The data have been centered and scaled.

Methodology based on logistic regression

Let us now introduce a statistical model. Let $\mathbf{Y}$ be the vector of subjective evaluations. As seen above, each response y_i is one or zero. We have:

$$\mathbf{Y} = (y_i)_{i=1\ldots n} \text{ with } y_i \in \{0,1\}, \forall i \in [1,n].$$

Our goal is the modeling of $\mathbf{Y}$ on the basis of the p measured explanatory variables X_1, X_2, …, X_p where $\mathbf{X_j} = [x_{1j}, x_{2j}, ..., x_{nj}]^T$. Let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p]$ be the design matrix. The response has only two different ratings: we are in the field of binary classification.

As mentioned above, engineers first want to know which variables explain the separation between the two classes most efficiently. A model must then be provided to engineers, so that they can set target values on the selected explanatory variables in order to reach a given quality level. The model must be simple enough. Therefore we consider generalized linear models, excluding the use of kernels.

Our first task is to select a set of explanatory variables. We will keep in mind the importance of the variable selection aspects in the methodology we want to develop.

Popular approaches to binary classifications are for instance classification and regression trees (CART), discriminant analysis (DA) or support vector machines (SVM). CART, introduced by Breiman et al. (1984), can help with variable selection. However, in practice, this procedure can be sensitive to noise on data. Model selection in the context of a discriminant analysis is not easy because of the large number of models to consider. Finally, support vector machines are not tailored to model selection.

For purposes of model selection, it is crucial to rely on a likelihood criterion. We will base our method on the logistic regression likelihood, which, in the context of binary classification, has the following expression:

$$L_n(\boldsymbol{\beta}) = \exp\left\{\sum_{i=1}^{n} y_i x_i \boldsymbol{\beta} - \log\left(1 + e^{x_i \boldsymbol{\beta}}\right)\right\} \qquad (1)$$

where $\boldsymbol{\beta}$ is the regression parameter.

Model selection

A two step approach

Since in our context a model can be considered as a set of variables, the number of models grows exponentially with respect to the number of explanatory variables. To deal with model selection, one often relies on penalization functions. First contributors in this direction were Akaike (1973) who introduced a penalized log-likelihood for density estimation and Mallow (1973) who proposed a penalized least square regression. We can also quote the work of Birgé and Massart (2004) where powerful theoretical results are derived in a general Gaussian framework. In this work the collection of models is also quite general. An important conclusion of these theoretical results is that the larger the collection of models, the stronger the penalization term. Moreover, the stronger the penalization term, the larger the prediction error of the estimated model.

It is more favorable to perform model selection with a smaller number of models. Therefore we suggest a two steps approach. We propose to:

a) organize explanatory variables into a hierarchy
b) apply a penalized likelihood method to the sequence of nested models.

The sequence of nested models comes from step a). The first model contains only the first explanatory variable in the hierarchy and each following model consists of the current model with the addition of the next variable in the hierarchy.

Regularization path

We want to organize explanatory variables into a hierarchy, in order to know which model with only one variable is best, which model with two variables is best, etc. Maintaining a control on the model size is a goal we keep in mind in addition to the other goal of the performance of the binary classification.

There is indeed a trade-off between the model size and the error rate. The idea is to obtain a model which describes the data well enough while having a suitably low number of explanatory variables. Those two goals are indeed opposite: the best predictor with all the variables leads to the lowest error rate, but this best predictor is specific to the data and does not adapt itself well to new observations.

The model size is defined as the number of variables appearing in the linear combination in (1), and can be interpreted as the L_0 norm of the regression parameter. Direct handling of the L_0 norm penalization causes heavy algorithmic difficulties and computation delays. We prefer here to use the L_1 norm, which equals the sum of the absolute values of the regression coefficients, and is a good compromise between the L_0 norm and the L_2 norm – the Euclidian norm. We prefer the L_1 norm to the L_2 norm because the L_1 norm is closer to the L_0 norm than the L_2 norm which defines the Ridge regression. The reason for choosing the L_1 norm rather than the L_0 norm is the convexity of the L_1 – penalized problem, contrary to the L_0 – penalized one which is not convex. In this approach, the "model size" is thus evaluated by the L_1 norm of the regression parameter.

We focus on the trade-off between effective classification and controlled model size. We can describe this trade-off as follows:

- The error rate is controlled by the likelihood $L_n(\boldsymbol{\beta})$
- The model size is controlled by $\|\boldsymbol{\beta}\|_1$

Combining these two terms, we define the expression:

$$\boldsymbol{\beta}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbf{R}^p}{\arg\min}\left\{-\log L_n(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1\right\} \qquad (2)$$

where λ is a regularization parameter. This parameter sets the relative importance of each of the two antagonistic goals.

This penalized approach – in the case of standard linear regression – was introduced as LASSO by Tibshirani (1996) and well-studied in the literature since. We can quote the very interesting contribution by Efron et al. (2004). The authors present an algorithm called LAR (for Least Angle Regression) which computes a LASSO solution (for L_1-penalized least squares regression). He also obtains a Stagewise Regression solution by slightly modifying the LAR algorithm. Rosset et al. (2004) establish conditions on both cost and penalty functions in order to have a piecewise linear regularization path. Under those conditions, the entire path can easily be calculated from only a few points. Keerthi and Shevade (2006) suggest an approximation of the logistic regression loss function by a piecewise quadratic function.

Methods have also been proposed for choosing λ from data. For instance Zou et al. (2004) prove that the number of non-zero coefficients is an unbiased estimator of the number of degrees of freedom.

Asymptotic results have also been established. For a fixed p, Knight and Fu (2004) prove in a more general setting (namely least squares penalized by $\sum_{j=1}^{p}\left|\beta_j\right|^{\gamma}$) than the LASSO formulation that there exists –asymptotically with n – a mass of probability at 0 when the variable is not in the true model. Zhao and Yu (2006) establish that LASSO selects the true model consistently under a condition called the "Irrepresentable Condition". Those statements hold in the large p setting as n gets large.
Our approach is different since we focus on building a hierarchy of explanatory variables.

We are interested in the *order of appearance*, which is the order in which the explanatory variables enter the model as the model size increases. We assume that an explanatory variable is more important if it appears early in the model linear combination. However this requires a lot of care, as will be discussed later in the paper.

One way to obtain the explanatory variable hierarchy is to determine the *regularization path*, defined as the mapping $\lambda \mapsto \boldsymbol{\beta}(\lambda)$, where $\boldsymbol{\beta}(\lambda)$ is defined by (2). It can be computed by the LAR in the case of LASSO (Efron et al., 2004).

In the case of the penalized logistic regression procedure, an algorithm presented by Park and Hastie (2006)

computes an approximation to the path which is inspired by the LAR algorithm.
Since we focus only on the order of appearance, we do not need to compute the whole path. We propose later a different algorithm whose goal is to determine this order.

Model selection with BIC

A regularization path procedure organizes explanatory variables into a hierarchy, as shown in Figure 1. We denote by *active set* the current set of explanatory variables in the model. The sequence of active sets is displayed in Figure 1.

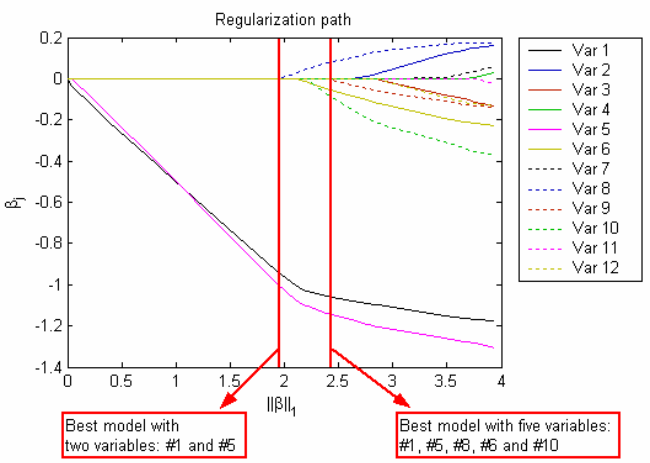

Figure 1. The regularization path identifies the best model with one variable, the best model with two variables, etc. Reading from left to right, explanatory variables are organized by importance into a hierarchy.

The regularization path results in an increasing sequence of models, as follows. Reading the figure vertically, at the very left we have $\|\boldsymbol{\beta}\|_1 = 0$, which corresponds to $\lambda \to \infty$. Then $\|\boldsymbol{\beta}\|_1$ increases as λ decreases until the very right of the figure, where $\lambda = 0$ corresponds to the maximal value of $\|\boldsymbol{\beta}\|_1$. The coefficient vector for $\lambda = 0$ is the result of the non-penalized logistic regression. Between $\lambda = 0$ and $\lambda \to \infty$, we have all the intermediate models.

In a second step, we consider the Bayesian Information Criterion of Schwarz (1978) for selecting the optimal model. If we denote by $\{M_k\}_{k=1\ldots m}$ the sequence of models, this criterion equals:

$$\left[BIC(M_k), \boldsymbol{\beta}_{BIC}(M_k)\right] = \inf_{\boldsymbol{\beta} \in V_k} \{-2\log L_n(\boldsymbol{\beta})\} + \|M_k\|_0 \log(n) \quad (3)$$

where n is the sample size of sample, V_k is the subspace of $\mathbf{R}^p$ with zeros at coordinates that do not appear in the current model M_k, and $\|M_k\|_0$ is the number of variables in the model M_k.

In our procedures, we recommend selecting the model that yields the minimal value of the BIC sequence. This way we ensure good prediction performance in the sense that we avoid over-fitting. Figure 2 displays the BIC sequence corresponding to the models sequence obtained from Figure 1.

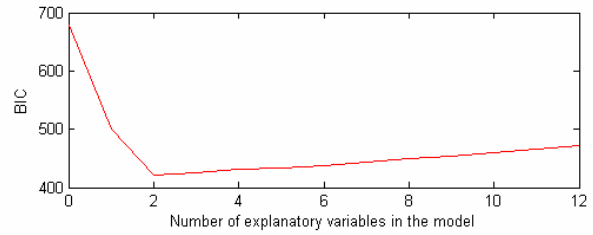

Figure 2. BIC curve.

Numerical aspects

In this section, we present an algorithm for obtaining the sequence of actives sets. An active set A_λ is defined as the set of explanatory variables corresponding to non-vanishing coordinates of β_j for a given regularization parameter $\lambda \geq 0$.

At a given λ, in the setting of L_1-penalized logistic regression, the calculation of a specific active set is a convex optimization problem.

By (1) and (2), for a given $\lambda \geq 0$, we need to compute:

$$\boldsymbol{\beta}(\lambda) = \arg\min_{\boldsymbol{\beta} \in \mathbf{R}^p} \left\{ \sum_{i=1}^{n} [-y_i x_i \boldsymbol{\beta} + \log\left(1 + e^{x_i \boldsymbol{\beta}}\right)] + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (4)$$

This criterion is clearly convex, but the L_1 norm on coefficients is problematic for differentiation near the axes. That is why we rewrite this criterion as the following:

$$\boldsymbol{\beta}(\lambda) = \arg\min_{\boldsymbol{\beta} \in \mathbf{R}^p} \left\{ \sum_{i=1}^{n} [-y_i x_i \boldsymbol{\beta} + \log\left(1 + e^{x_i \boldsymbol{\beta}}\right)] + \lambda \sum_{j=1}^{p} u_j \right\}$$
$$s.t. \quad -u_j \leq \beta_j \leq u_j \quad (5)$$

This new criterion searches for the same solution as (4) and is C^∞, with linear constraints.
Such an optimization can be performed using standard numerical procedures. In our application, we have used the Matlab function *fmincon* to solve this optimization problem.

A very simple way (close to the one presented by Park and Hastie, 2006) to calculate the path consists in a stepwise optimization. Beginning with $\lambda = 0$, λ is increased by $\delta\lambda$ step by step. At each step $\boldsymbol{\beta}(\lambda)$ is calculated along with the optimization function. The algorithm ends when all coordinates of $\boldsymbol{\beta}(\lambda)$ are zero (for λ large enough, all coordinates are indeed zero).

The main problem is the choice of step length, which must be small enough to insure that all interesting points of the path are visited. We present an intuitive algorithm, which is easy to develop and reduces the number of optimizations to perform.

Algorithm idea: instead of calculating $\boldsymbol{\beta}(\lambda)$ at each step, by increasing λ by steps of $\partial\lambda$, we suggest exploring the range of $\lambda \in [0,+\infty[$ in a dichotomist way.

Considering equation (4), we first note that there exists $\lambda_{\max}$ such that $\forall \lambda > \lambda_{\max} : \boldsymbol{\beta} \equiv 0$. The constraint on the size of the coefficient tends to infinity and thus forces every coefficient to be zero. We consider the value of $\lambda_{\max}$ given in Park and Hastie (2006), namely $\lambda_{\max} = \max_{j\in\{1,\dots,p\}} \left| X'_j (y - \bar{y}) \right|$. We just need an upper bound on λ.

We base our stopping condition on the current active set, that is, we decide to perform new levels of dichotomy if the difference between the two active sets is more than one in terms of cardinality. If we denote the active set for a given λ by A_λ, the stopping condition can be written as:
$COND(A_{\lambda_1}, A_{\lambda_2}) : \exists!\, j \in \{1,\dots,p\}, \exists B \subset A_{\lambda_2} : A_{\lambda_1} = \{j\} \cup B$.
Our algorithm is described in more detail, with comments in italics, in the Appendix.

Application to the industrial case

In this section, we apply our algorithm to the data we presented earlier. We recall that variables are centered and scaled.

Interpretation of the regularization path

Our algorithm performs the regularization path presented in Figure 1. For ease of interpretation, we present the evolution of the coefficients of explanatory variables as functions of $\|\boldsymbol{\beta}\|_1$.

We can summarize the interpretation of this result as follows. As we release the constraint, variables #1 and #5 enter the model nearly together. For an industrial application, we shall say that there is no statistical reason to consider one of those without the second one. This is the first piece of information. The BIC curve presented in Figure 2 contributes to the second main part of the interpretation, that is: both variables #1 and #5 are sufficient to explain the physical phenomenon we model.

Those results concur with the engineers' intuitive analysis of the physical phenomenon: variable #1 was already in the specifications and engineers had the feeling that variable #5 could bring some more information. Our statistical analysis proves them to be right.

Sensitivity analysis

In this step of the interpretation, we have found which variables have an impact on subjective evaluations by customers. Having estimated the model and the regression parameter β, we have in hand the probability $\mathrm{P}(Y = 1 | X = x)$ where *Y* is the gear-box rating and *x* is the value of the physical variables.

Our prediction is a probability between 0 and 1. But the variable *Y* we wish to predict is binary. So in order to decide the value of *Y*, knowing that $X = x$, we have to choose from which threshold probability we decide that *Y=1*.

In statistical words, we say that we build the following test:

$$\hat{Y}(x) = \begin{cases} 1 & \text{if } \mathrm{P}(Y = 1 | X = x) > c \\ 0 & \text{otherwise} \end{cases},$$

where *c* is called a *probability threshold*. This test decides whether the physical parameter *x* calibrates an acceptable gear-box.

Our job is also to determine the value of the probability threshold. Specifications will follow. We can for instance set the threshold to a high value to harden specifications: only very good gear-boxes will be validated.

The choice of the probability threshold is driven by the error rate. There are two types of error rate: the false

negative one and the false positive one. The false positive error rate is $P(\hat{Y}=1|Y=0)$ and the false negative error rate is $P(\hat{Y}=0|Y=1)$.

From an industrial point of view, a false positive observation corresponds for instance to a validated gear-box (a recast-as-"1" sample) which is criticized by the customer (the customer's subjective evaluation is "0"). For a car manufacturer, such a false positive observation causes the most prejudicial type of error.

Figure 3 summarizes these trade-offs graphically. The x-axis plots differences between a given value of the model linear combination and the value of the linear combination which corresponds to an estimated probability $P(\hat{Y}=1)$ of .8. These differences decrease from left to right. The smooth curve in Figure 3 represents the estimated probabilities $P(\hat{Y}=1)$ for each value of the differences (note that a zero difference does correspond to a .8 estimated probability). On the other hand, the jagged line represents the false positive rates for threshold probabilities corresponding to differences on the x-axis. We can see that the .8 threshold probability corresponds to a false positive rate of about 8%.

Note that the choice of a threshold probability of .5 would imply a false positive rate of near 20%, which is too high. Overall we feel that the choice of .8 for a probability threshold is a reasonable compromise.

For another description of the evolution of error rates, we display the ROC (Receiver Operating Characteristic) curve in Figure 4.

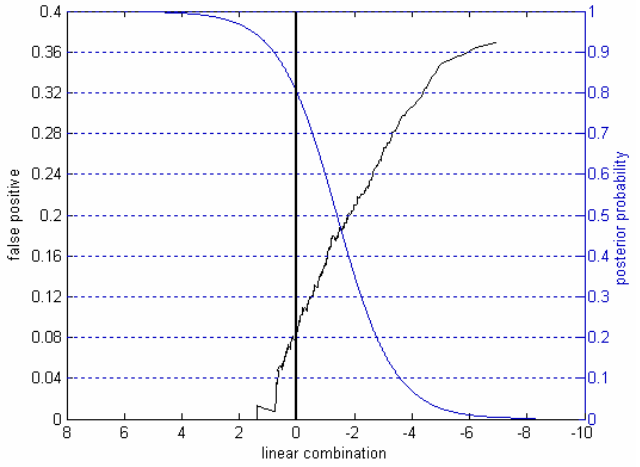

Figure 3. Posteriori probability and false positive error rate as functions of the value of the linear combination (distance from the separating line). Vertical line corresponds to the separating line in Figure 5.

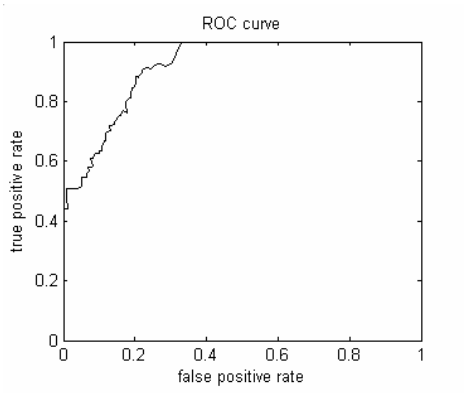

Figure 4. ROC curve.

We also plot the complete sample in the space {variable #1, variable #5} in order to visualize the separating line corresponding to the probability threshold of .8.

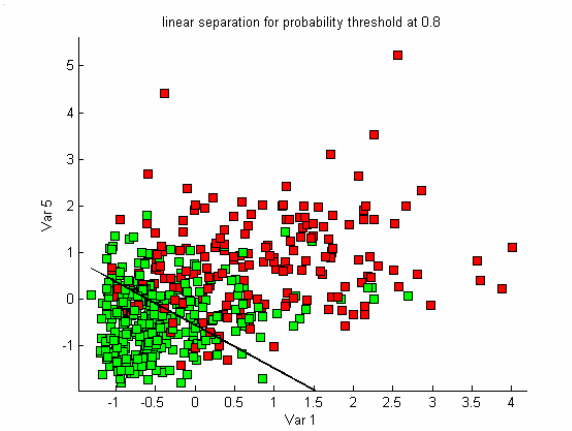

Figure 5. Position of the separating line corresponding to a probability threshold of 0.8. Observations with Y=1 are green. Observations with Y=0 are red. Observations below the line have $\hat{Y}=1$; observations above the line have $\hat{Y}=0$.

For a complete industrial analysis of those results and recommendations, we have to point out that those figures can also help with design. Indeed, the links between values of linear combinations, separating line positions and probability thresholds highlighted above can help to calibrate a new gear-box. Engineers have to decide on a specific target zone in the space {variable #1, variable #5}. Assuming that one of the variables has already been calibrated and cannot be modified anymore, the target zone could nevertheless be reached thanks to the other variable. Moreover, engineers would have information about the gain or loss in probability for each value of the variables.

Discussion

We assumed above that the importance of explanatory variables is related to the order in which the variables

enter the model as the controlled model size gets larger. We assumed that an explanatory variable is more important if it appears early in the model linear combination.

In the figure presented in this paper, the order of the two selected variables changes along the path. This is not a problem because they are selected together: we consider those two variables as a group. Variables maintain a constant order during the path if the following circumstances occur: a variable which enters the model maintains a magnitude

- larger than the following variables
- and at the same time smaller than that of the active variables (variables which are already in the model).

In practice, this is not the case. Links between explanatory variables such as nearly linear relations may have an impact on the stability of the order of variables. This issue becomes even more critical as soon as $p > n$.

We show in Figure 6 the following example. We consider the same dataset as previously, but to which we add only three more explanatory variables. They are measurements of physical parameters similar to the twelve first variables.

The BIC leads us to considering two variables in the model. But which two variables? In this setting, the two first variables to enter the model ('Var 1' and 'Var 5') are not the same two variables as the two maximum absolute magnitude variables at the end of the path. Our approach seems to fail because variable #5 is the second variable into our defined hierarchy but is the third variable in absolute magnitude at the end of the path.

Greenshtein and Ritov (2006) give bounds on the L_1-norm and on the number of variables as n and p simultaneously get large, with $p >> n$. These model selection procedures tend to select much larger variable sets. The first variables that enter the model are obviously of some importance in the explanation of the response but this procedure may have to be moderated: there might be a trade off to discover between those two approaches.

The path presented in Figure 6 is ambiguous. In our industrial application, we discussed this phenomenon with industrial experts. Considering the physical interpretation of each design parameter, variables #5, #13 and #14 appeared deeply linked. Another physical link appeared between variables #1 and #15.

Experts decided to consider only one variable, representative of its group. We kept variables #1 and #5. Experts made their choice according to the physical meaning of each group and have selected the variable which is easiest to interpret.

Once the variables #13, #14 and #15 are removed from data, the path appears to be clearer and the selection of variables set is not ambiguous any more: the two variables first entering the model remain the two variables with the largest absolute magnitude along the path.

This step turned out to be crucial for a good understanding of the problem. Discussions with experts enabled us to find a regularization path that can be easily interpreted. It would be interesting to have an automatic method to clarify the regularization path. This is an open issue.

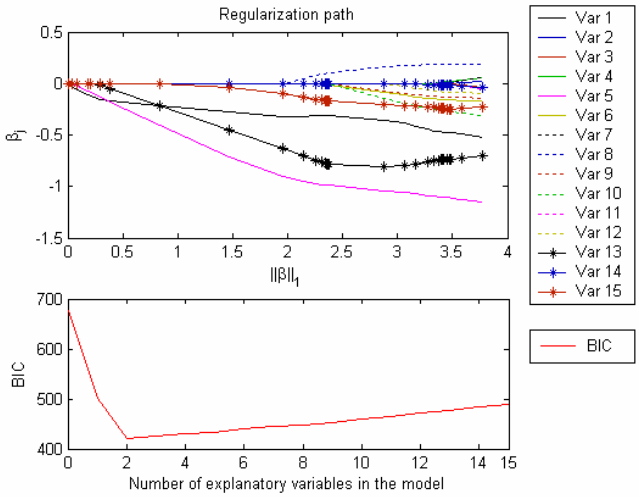

Figure 6. Regularization path and BIC curve on the augmented industrial dataset. We note that variable #1 and variable #5 are the two first design parameters to enter the model but they are not the two with the greatest magnitude at the end of the path (model resulting from the non-penalized logistic regression).

Conclusion

We have considered the problem of binary classification using a logistic regression model. Our main objective was to select a few explanatory variables. We suggested

- first to select a sequence of models by ordering the explanatory variables as an output of the regularization path of the L_1 – penalized likelihood.
- second to apply the BIC criterion to select a model in this sequence.

Our ordering method can be summarized as follows:
"The sooner a variable enters the regularization path as the penalty decreases, the more explanatory it is." This

interpretation of the regularization path is ad hoc and thus needs some care in practice. As we applied this approach to our industrial case, we relied on experts to first clean up the regularization path so that the selected variables also correspond to those having large regression coefficients order of magnitude in the regularization path.

Acknowledgments

This methodology was developed in the framework of a CIFRE-PhD in the Research Center RENAULT. I want to thank François Roueff[1] for his precious help with this work, and Marc Chauvet for his language corrections.

Correspondence: germain@enst.fr

[1] Télécom Paris, CNRS LTCI, URA820, 75634 Paris Cedex 13, France.

REFERENCES

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium on Information Theory*, P. N. Petrov and F. Csaki, editors, 267-281.

Birgé, L., and Massart, P. 2001. Gaussian model selection. *Journal of the European Mathematical Society* 3(3):203-268.

Breiman L., Friedman J. H., Olshen R. A., and Stone C. J. 1984. *Classification and regression trees*. Wadsworth International, Belmont, California.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. 2004. Least angle regression. *Annals of Statistics* 32:407-499.

Greenshtein, E., and Ritov, Y. 2004. Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli* 10:971-988.

Keerthi, S., and Shevade, S. 2006. A fast tracking algorithm for generalized LARS/LASSO. Submitted to *IEEE transactions on Neural Networks*.

Knight, K., and Fu, W. 2000. Asymptotics for LASSO-type estimators. *Annals of Statistics* 28(5):1356-1378.

Mallows, C. L. 1973. Some comments on Cp. *Technometrics* 15:661-675.

Park, M. Y., and Hastie, T. 2006. L_1 regularization path algorithm for generalized linear models. *Technical report*. Stanford University, Stanford.

Rosset, S., and Zhu, J. 2004. Piecewise linear regularized solution paths. *Annals of Statistics*, to appear.

Schwarz, G. 1978. Estimating the Dimension of a Model. *Annals of Statistics* 6(2):461-464.

Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society B* **58**:229-243.

Zhao, P., and Yu, B. 2006. On Model Selection Consistency of Lasso. *Technical Report 702*, Statistics Department, UC Berkeley (to appear in *Journal of Machine Learning Research*).

Zou, H., Hastie, T., and Tibshirani, R. 2004. On the degrees of freedom of the lasso. *Technical report*, Department of Statistics, Stanford University.

Appendix: Description of Algorithm:

1) Initialization: calculation of $\lambda_{\max}$.

$\lambda_{\max} = \max_{j\in\{1,\dots,p\}} \left| X'_j (y - \bar{y}) \right|$

set $(\lambda_1, \lambda_2) \leftarrow (0, \lambda_{\max})$

calculate A_{λ_1} and A_{λ_2}

2) While $not\{COND(A_{\lambda_1}, A_{\lambda_2})\}$ We are going to cut the λ range until the two active sets A_{λ_1} and A_{λ_2} are equal or differ by exactly one variable.

a) $\mu = \frac{\lambda_1 + \lambda_2}{2}$. We split the λ range in its middle.

b) calculate A_μ

c) while $not\{COND(A_\mu, A_{\lambda_2})\}$, do

i) if $A_\mu = A_{\lambda_2}$, do

(1) $(\lambda_1, \mu, \lambda_2) \leftarrow \left(\lambda_1, \frac{\lambda_1 + \lambda_2}{2}, \mu\right)$ We dichotomize this other side of the λ range.

(2) calculate the new A_μ

(3) calculate the new A_{λ_2} In fact, this calculation is not performed because we already have calculated the new A_{λ_2} : it is A_μ .

ii) end if

iii) if $\left|A_\mu - A_{\lambda_2}\right| \geq 2$, do

(1) $(\lambda_1, \mu, \lambda_2) \leftarrow \left(\mu, \frac{\lambda_1 + \lambda_2}{2}, \lambda_2\right)$ We have not split enough. We keep dichotomizing this side.

(2) calculate the new A_μ

(3) calculate the new A_{λ_2} Same remark as in 2) c) i) (3): the new A_{λ_2} is already calculated: it is A_{λ_2} .

iv) end if

d) end while

e) $\lambda_1 \leftarrow \max\{\lambda \in \Lambda : \lambda < \mu \text{ and } not\{COND(A_\lambda, A_\mu)\}\}$

Let Λ be the set of λ already calculated. During the algorithm, calculations of active sets give much information. We can improve the algorithm by looking for the more judicious λ_1 to consider.

f) $\lambda_2 \leftarrow \min\{\lambda \in \Lambda : A_\lambda = A_\mu\}$ The more judicious λ_2 to consider is the lowest value of λ for which $A_\lambda = A_\mu$.

3) end while

4) end

The algorithm summarized:

1) $\lambda_{\max} = \max_{j\in\{1,\dots,p\}} \left| X'_j (y - \bar{y}) \right|$

set $(\lambda_1, \lambda_2) \leftarrow (0, \lambda_{\max})$

calculate A_{λ_1} and A_{λ_2}

2) while $not\{COND(A_{\lambda_1}, A_{\lambda_2})\}$ do

a) $\mu = \frac{\lambda_1 + \lambda_2}{2}$

b) calculate A_μ

c) while $not\{COND(A_\mu, A_{\lambda_2})\}$, do

i) if $A_\mu = A_{\lambda_2}$, do

(1) $(\lambda_1, \mu, \lambda_2) \leftarrow \left(\lambda_1, \frac{\lambda_1 + \lambda_2}{2}, \mu\right)$

(2) calculate A_μ

ii) if $\left|A_\mu - A_{\lambda_2}\right| \geq 2$, do

(1) $(\lambda_1, \mu, \lambda_2) \leftarrow \left(\mu, \frac{\lambda_1 + \lambda_2}{2}, \lambda_2\right)$

(2) calculate A_μ

d) $\lambda_1 \leftarrow \max\{\lambda \in \Lambda : \lambda < \mu \text{ and } not\{COND(A_\lambda, A_\mu)\}\}$

e) $\lambda_2 \leftarrow \min\{\lambda \in \Lambda : A_\lambda = A_\mu\}$

6.2 Description des critères physiques potentiellement explicatifs dans le cas de la prestation « Accroc-Croquement »

Fc : effort de crabotage maximal La mesure de l'effort maximal en phase de crabotage (voir fig. 6.2) est ici pris en compte indépendamment de l'effort maximal en phase de synchronisation. Ce critère premet d'identifier le pic le plus sévère mais n'apporte pas d'information sur la forme du croquement.

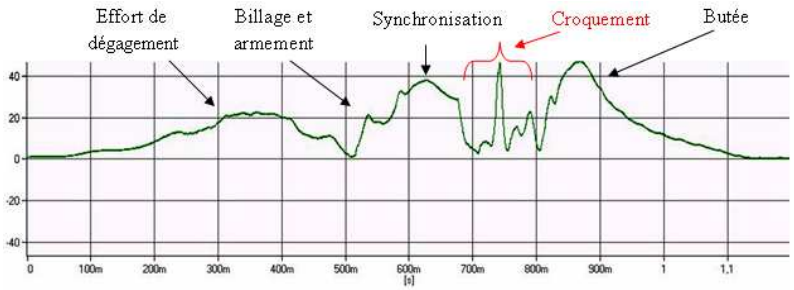

FIG. 6.1 – Courbe d'effort ressenti au pommeau en fonction du temps lors d'un passage de vitesse

Fc/Fs Fc/Fs est le rapport entre l'effort de crabotage et l'effort maximal dans la phase de synchronisation (voir fig. 6.2).
Lors d'un passage, les efforts de dégagement, de synchronisation et de butée font partie du retour d'information compris par le client alors que l'effort de crabotage, s'il est trop important, est lui perçu comme une perturbation gênante (voir fig. 6.1). L'effort de dégagement permet de quitter la vitesse qui était enclenchée, l'effort de synchronisation indique que le nouveau rapport de boîte de vitesse commence à s'enclencher et l'effort de butée est simplement l'effort subi indiquant qu'on est au bout de la course du levier et que la vitesse est bien enclenchée. Le crabotage est la mise en contact du pignon fou de l'axe secondaire (après avoir été ralenti - ou accéléré par l'action au manchon baladeur) avec le pignon fixe correspondant sur l'axe primaire. Les pièces sont en phase et le crabotage consiste à les rapprocher jusqu'au mouvement solidaire. Ces pièces sont munies de dents hélicoïdales et bien que les pièces tournent à la même vitesse, le rapprochement peut provoquer un choc dû au non-alignement des dents.
On peut noter tout de même que ce critère physique ne prend pas en compte les multichocs, ni la forme du pic de croquement.

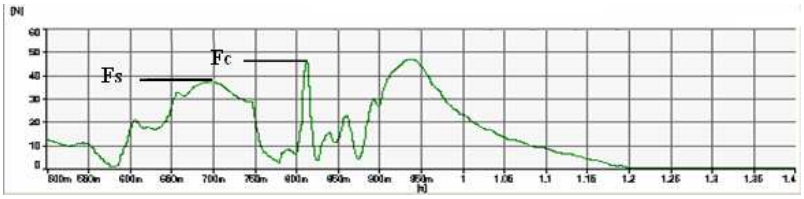

FIG. 6.2 – Courbe d'effort ressenti au pommeau en fonction du temps lors d'un passage de vitesse

Nombre de pics Ce critère physique a été proposé par un fournisseur, dans le cadre du développement d'un synchroniseur double cône. Il s'agit d'une analyse visuelle qui permet de hiérarchiser les différentes courbes de déplacement en fonction de leur degré de criticité pour le client. Ce critère permet de caractériser la forme la courbe de déplacement.

Recul Les phénomènes d'accroc et de croquement ont pour effet l'appartition d'efforts ressentis par l'utilisateur au moment du crabotage, mais également de ralentissements, d'arrêts, voire de reculs dans le déplacement du pommeau lors du passage.
Le critère physique correspondant à ces perturbations touchant le déplacement de la main du conducteur consiste donc à examiner l'amplitude du recul maximal au pommeau.
Ainsi, même sans recul, le déplacement maximal pendant l'effort de crabotage donnera un aperçu de la fluidité du passage (fig. 6.3).

- passage fluide : recul > 0
- temps d'arrêt : recul ≈ 0
- passage avec recul : recul < 0

Angle de recul En considérant que l'amplitude du recul n'est pas suffisante pour caractériser la violence du choc de croquement, l'angle de recul permet d'apporter cette information supplémentaire.
Ce critère physique consiste à prendre en compte les pentes du déplacement du pommeau avant et après le choc afin de quantifier leurs variations (voir fig. 6.4).
Ainsi l'angle entre ces deux tangentes apporte une information sur la brutalité du recul à la main du conducteur. Un angle très fermé est synonyme de passage critiqué.

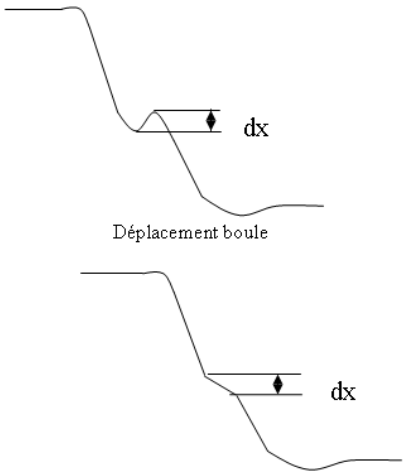

FIG. 6.3 – Déplacement de la boule du pommeau depuis la fin de synchronisation jusqu'à l'engagement complet de la vitesse, avec recul (courbe du haut) et sans recul (courbe du bas)

Énergie cinétique Lors du phénomène de croquement, le manchon baladeur absorbe une partie du choc principal sous la forme d'un déplacement (qui est souvent un recul). L'énergie cinétique ainsi absorbée par le baladeur est restituée par la commande interne, jusqu'à l'utilisateur via le pommeau. Le critère physique correspondant est le suivant :

$$E_c = \frac{1}{2} m_{eq\_baladeur} . V^2_{recul\_pommeau} . \xi$$

avec $\xi \in \{-1, 1\}$. $V_{recul\_pommeau}$ est la vitesse du pommeau. $m_{eq\_baladeur}$ est la masse équivalente du pommeau.

Impulsion de passage Pour ce critère physique, le passage est pris en compte en entier. L'impulsion de passage permet d'identifier le niveau d'effort global du passage (voir fig. 6.5). L'inconvénient est que tous les éléments pouvant dégrader la fluidité du passage sont fondus dans ce critère physique. L'impulsion de passage de cible en particulier pas le croquement seul et reste très sensible aux forts efforts de synchronisation, ce qui n'implique pas le ressenti d'une gêne.

Impulsion de crabotage L'impulsion de crabotage est définie comme l'intégrale de la courbe d'effort mesurée au pommeau, durant la phase de crabotage. Les chocs

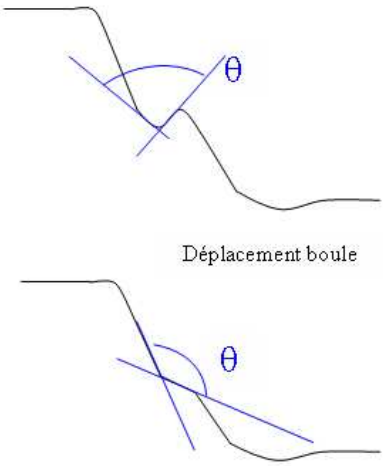

FIG. 6.4 – Angle de recul lors d'un déplacement de la boule du pommeau avec recul (courbe du haut) et sans recul (courbe du bas)

multiples sont pris en compte ainsi que l'ampleur du croquement mais peut correspondre à des profils d'effort très différents.

Impulsion de recul Ce critère physique est développé à partir d'une courbe théorique « idéale » de déplacement du pommeau. Il consiste à calculer l'aire séparant les courbes réelle et théorique durant les phases de crabotage, zone d'apparition du croquement. La courbe dite « idéale » est construite sur la base d'un polynôme de degré deux calculé sur trois points caractéristiques du déplacement du pommeau : du point de début de la fin de la synchronisation, en passant par le point d'inflexion de la courbe dû au crabotage et enfin au point de fin de passage. Ces trois points sont les plus identifiables sur une courbe de déplacement en dehors du croquement. La fin de synchronisation est repérée par l'avancement du manchon baladeur après l'arrêt dû à la synchronisation, le point d'inflexion de crabotage est le résultat du ralentissement du pommeau à l'approche du pignon fou par le manchon baladeur et le point de fin de passage consiste en la fin de la course. Une impulsion faible correspond à un bon passage, elle est la représentation de l'écart à la courbe « idéale ».

Puissance de recul Ce critère physique est un exemple de composé à partir des autres critères. La puissance de recul ajoute au précédent l'information de l'ampli-

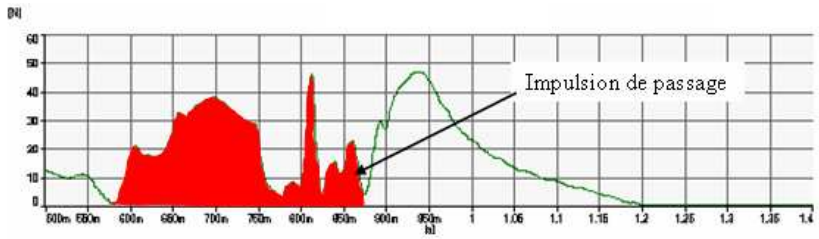

FIG. 6.5 – L'impulsion de passage donne une idée de l'effort global fourni pendant le passage

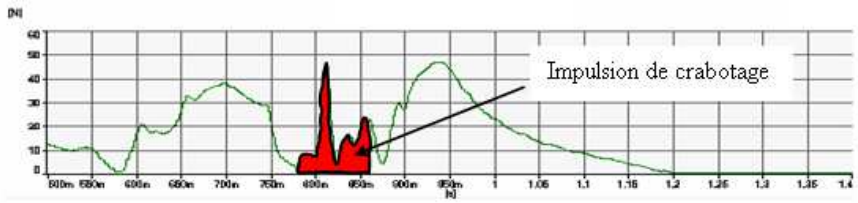

FIG. 6.6 – L'impulsion de crabotage cible l'effort global fourni uniquement lors de la phase de crabotage

tude de l'effort de croquement maximal Fc. La puissance de recul s'écrit :

$$P_{recul} = Fc.V_{recul\_pommeau}.\xi$$

RMS1 Les trois critères physiques vibratoires RMS1, RMS2 et RMS3 sont tous les trois calculés à partir du signal temporel du déplacement du pommeau lors du passage. RMS1 est l'énergie spectrale du signal entre les fréquences 20Hz et 100Hz.

RMS2 RMS2 est l'énergie spectrale du signal entre les fréquences 20Hz et 1000Hz.

RMS3 RMS3 est l'énergie spectrale du signal entre les fréquences 100Hz et 1000Hz.

Temps de passage La prise en compte du temps de passage permet de prendre ce temps en compte pour compléter d'autres informations, via une combinaison de critères. Le temps de passage en lui-même discrimine les passages rapides des

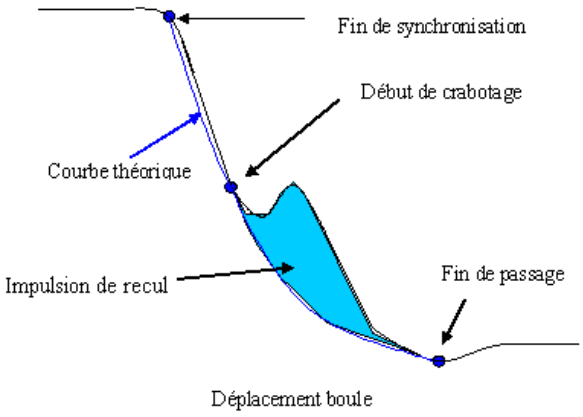

FIG. 6.7 – Impulsion de recul lors d'un déplacement avec recul

passages lents, ces derniers étant plus favorables à l'apparition du croquement. Pour la validation d'une boîte de vitesse, la valeur inscrite au cahier des charges pour le temps de passage est de 0.35s $\pm$ 0.15s.

Température Tout comme l'effort maximal en phase de crabotage, la température peut être combinée à d'autres critères physiques pour créer de nouveaux critères qui permettent une explication plus globale de la prestation. La température joue sur la viscosité de l'huile. La phase de synchronisation est donc dégradée à froid. L'expérience montre que le phénomène de croquement est effectivement plus présent à froid.

6.3 Article : Week Convergence of the Regularization Path in Penalized M-Estimation

L'article reproduit ci-après a été soumis aurpès de la revue *The Annals of Statistics*.

WEAK CONVERGENCE OF THE REGULARIZATION PATH IN PENALIZED M-ESTIMATION

JEAN-FRANÇOIS GERMAIN AND FRANÇOIS ROUEFF

RENAULT DREAM-DTAA and Institut TELECOM, TELECOM ParisTech, LTCI CNRS

ABSTRACT. We consider an estimator $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ defined as the element $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ minimizing a contrast process $\Lambda_n(\boldsymbol{\phi}, \mathbf{t})$ for each $\mathbf{t}$. We give some general results for deriving the weak convergence of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ in the space of bounded functions, where, for each $\mathbf{t}$, $\boldsymbol{\beta}(\mathbf{t})$ is the $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ minimizing the limit of $\Lambda_n(\boldsymbol{\phi}, \mathbf{t})$ as $n \to \infty$. These results are applied in the context of penalized M-estimation, that is, when $\Lambda_n(\boldsymbol{\phi}, \mathbf{t}) = M_n(\boldsymbol{\phi}) + \mathbf{t} J_n(\boldsymbol{\phi})$, where M_n is a usual contrast process and J_n a penalty such as the ℓ^1 norm or the squared ℓ^2 norm. The function $\widehat{\boldsymbol{\beta}}_n$ is then called a *regularization path*. For instance we show that the central limit theorem established for the lasso estimator in [11] continues to hold in a functional sense for the regularization path. Other examples include various possible contrast processes for M_n such as those considered in [14]. To illustrate these results in the lasso case, we propose a test statistic based on the regularization path whose asymptotic distribution is known under the null hypothesis $H_0 : \boldsymbol{\beta} = 0$. The performance of the test is assessed on synthetic data.

1. INTRODUCTION

Let us consider a real-valued contrast process $\{M_n(\boldsymbol{\phi}),\ \boldsymbol{\phi} \in \boldsymbol{\Phi}\}$ based on an observed sample of size n and a contrast function M defined on the same parameter set $\boldsymbol{\Phi}$ and minimized at the point $\boldsymbol{\beta}$. A penalized estimator with weight $\mathbf{t} \geq 0$ is defined as the minimizer of the contrast process

$$\Lambda_n(\boldsymbol{\phi}, \mathbf{t}) = M_n(\boldsymbol{\phi}) + \mathbf{t}\ J_n(\boldsymbol{\phi}), \quad \boldsymbol{\phi} \in \boldsymbol{\Phi}\ , \tag{1}$$

where J_n is a non-negative function defined on $\boldsymbol{\Phi}$, not depending on the observations but possibly on n, mainly to allow some convenient normalization.

The use of penalties is popular for ill-posed problems and model selection, among which the ridge regression (see [8]) and the lasso (see [16]) are emblematic examples. In these two examples the contrast process M_n is the least-square criterion and the penalty function J_n is the squared ℓ^2 norm and the ℓ^1 norm, respectively. Consistency and central limit theorems are established in [11] precisely in the case where M_n is the least-square criterion and J_n is in a family of penalties including both the squared ℓ^2 norm and the ℓ^1 norm. They show that, when the penalty is conveniently normalized, the penalized mean square estimator is

Date: August 29, 2008.

1991 *Mathematics Subject Classification.* Primary 62J07, 62F12, 60F17 Secondary: 62J05, 60F05, 62E20.

Key words and phrases. lasso, penalized M-estimation, regularization path, weak convergence, Argmax theorem.

Corresponding author: F. Roueff, Institut TELECOM, TELECOM ParisTech, LTCI CNRS.

no longer asymptotically normal. Instead, its asymptotic distribution is defined as the minimizer of penalized quadratic form applied to a Gaussian vector (see *e.g.* [11, Theorem 2]). Their asymptotic results hold as the number n of observations tends to infinity and for a fixed finite-dimensional model. Quite different results have been established when the dimension of the model increases with n, see [6, 19, 3, 1] and the references therein. These results provide interesting properties of the lasso for model selection or prediction purposes in the context of sparse models. Although specific normalizations of the penalty (different from those required in [11]) are prescribed in these theoretical results, there exist numerous heuristic ways for choosing the penalty weight $\mathbf{t}$ in practice. The first step is to minimize $\Lambda_n(\boldsymbol{\phi}, \mathbf{t})$ in (1) on $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ for a collection of non-negative weights $\mathbf{t}$, resulting in a collection of estimators $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$, which is called the *regularization path* (or the *solution path*). The Least Angle Regression (LAR) technique introduced by Efron et al. in [4] provides, in most cases, the entire path, computed with the complexity of a linear regression. In a second step, some criterion is used to select $\mathbf{t}$, see *e.g.* [20] where AIC and BIC procedures are proposed for the lasso. Because the whole path is used by the practitioner, we think that it is crucial to examine whether the convergence of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}) - \boldsymbol{\beta})$, established in [11] for one fixed $\mathbf{t}$, continues to hold in a functional sense and, if it is the case, to determine the limit distribution. The goal of this paper is twofold. First we show that, under the same assumptions as in [11], the convergence holds in the space of locally bounded functions. Second we extend this result to more general contrast processes M_n such as generalized linear models (GLM) or least amplitude deviation (LAD). As an illustration we propose a test statistic computed on the lasso regularization path and determine its asymptotic distribution under the null hypothesis $H_0 : \boldsymbol{\beta} = 0$.

Let us specify the asymptotic behavior of the lasso regularization path under the corresponding assumptions. Consider the linear model

$$y_k = \mathbf{x}_k^T \boldsymbol{\beta} + \varepsilon_k, \quad k = 1, 2, \dots \tag{2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown parameter, (y_k) is a sequence of real-valued observations, $(\mathbf{x}_k)$ is the sequence of regression vectors and (ε_k) is a strong white noise with variance σ^2. For any $\mathbf{t} \geq 0$, the lasso estimator $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ minimizes the penalized contrast process $\Lambda_n(\boldsymbol{\phi}, \mathbf{t})$ on $\boldsymbol{\phi} \in \mathbb{R}^p$, where

$$\Lambda_n(\boldsymbol{\phi}, \mathbf{t}) = \frac{1}{n} \sum_{k=1}^{n} (y_k - \mathbf{x}_k^T \boldsymbol{\phi})^2 + \mathbf{t} \lambda_n \sum_{i=1}^{p} |\phi_i| \,, \tag{3}$$

which is a specific form of (1). Denote $\mathbf{X}_n = [\mathbf{x}_1, ..., \mathbf{x}_n]^T$. We consider the following assumptions, for consistency and central limit theorem, respectively. The assumptions are the same as in [11].

Assumption 1.

(i) $C_n = n^{-1} \mathbf{X}_n^T \mathbf{X}_n \to C$, where C is a positive-definite matrix;
(ii) $\lambda_n \to 0$.

Assumption 2.

(i) Assumption 1-(i) holds;
(ii) $\max_{1 \leq k \leq n} \|\mathbf{x}_k\|^2 = o(n)$;
(iii) $\lambda_n = n^{-1/2}$.

Assumptions 1-(i) and 2-(ii) are the classical assumptions for the asymptotic behavior of least squares estimators. The other assumptions provide the appropriate way of normalizing the ℓ^1 penalty.

Theorem 1. *Under Assumption 1, for any $L > 0$, $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ converges in probability to $\boldsymbol{\beta}$ uniformly in $\mathbf{t} \in [0, L]$, that is*

$$\sup_{\mathbf{t}\in[0,L]} \|\widehat{\boldsymbol{\beta}}_n(\mathbf{t}) - \boldsymbol{\beta}\| \xrightarrow{P} 0 \,. \tag{4}$$

We now define the limit process of the lasso regularization path, appropriately centered and normalized. Let $U \sim \mathcal{N}(0, \sigma^2 C)$. For any $\mathbf{t} \geq 0$, we define $\widehat{\boldsymbol{u}}(\mathbf{t})$ as the point $\boldsymbol{\phi} \in \mathbb{R}^p$ which minimizes

$$\mathbb{L}(\boldsymbol{\phi}, \mathbf{t}) = -2U^T\boldsymbol{\phi} + \boldsymbol{\phi}^T C \boldsymbol{\phi} + \mathbf{t}\left[\sum_{j=1}^{p} \phi_j \operatorname{sgn}(\beta_j)\, \mathbb{1}_{\{\beta_j \neq 0\}} + |\phi_j| \mathbb{1}_{\{\beta_j = 0\}}\right] . \tag{5}$$

It is easy to show that this defines $\widehat{\boldsymbol{u}}(\mathbf{t})$ uniquely for all $\mathbf{t} \geq 0$ (see the proof of Theorem 2). The distribution of $\widehat{\boldsymbol{u}}$ as a function is not explicit but is not more complicated than its marginal distributions already described in [11], since the whole path is described as a deterministic function of the random variable (r.v.) U. An interesting property of $\widehat{\boldsymbol{u}}(\mathbf{t})$ is that, with probability 1, the set of its components that vanish for $\mathbf{t}$ large enough is given by the set of zero components of the true parameter $\boldsymbol{\beta}$.

Theorem 2. *Under Assumption 2, for any $L > 0$*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \rightsquigarrow \widehat{\boldsymbol{u}} \;\; \textit{in } \ell^\infty([0, L], p) \,, \tag{6}$$

where $\rightsquigarrow$ denotes the weak convergence and $\ell^\infty([0, L], p)$ the space of bounded $[0, L] \to \mathbb{R}^p$ functions.

Remark 1. The fact that the convergence (6) holds on a compact $[0, L]$ is not only a technical restriction. Indeed, the convergence clearly does not hold on $\ell^\infty(\mathbb{R}_+, p)$. To see why, observe that, by the definition of $\widehat{\boldsymbol{u}}$, its coordinates corresponding to non-vanishing β_j are unbounded as $\mathbf{t} \to \infty$. In contrast, the left-hand side of (6) is bounded since, for any n, there is a large enough $\mathbf{t}$ for which $\widehat{\boldsymbol{\beta}}_n(\mathbf{t}) = 0$. Note that this also implies that $\sup_{\mathbf{t}\in\mathbb{R}_+} \|\widehat{\boldsymbol{\beta}}_n(\mathbf{t}) - \boldsymbol{\beta}\| \geq \|\boldsymbol{\beta}\|$, and thus that the consistence (4) does not hold if $[0, L]$ is replaced by $\mathbb{R}_+$.

The proofs of Theorem 1 and Theorem 2 are applications of some general results on the consistency of convex penalized M-estimators and on the weak convergence of Argmin's depending on a parameter $\mathbf{t}$ (the so called Argmin theorem in the following). More general penalized contrast processes will also be considered. Such extensions are of interest since the lasso regularization path has been extended to the case where M_n is different from the least-square criterion. In [13], Hastie and Park propose a fast numerical algorithm for determining the regularization path when M_n is a regression function based on a negated log-likelihood of the canonical exponential family. In [5], a fast algorithm based on a dichotomy is proposed to explore the range of $\mathbf{t}$'s in the specific case of logistic regression penalized by the ℓ^1 norm.

The paper is organized as follows. Section 2 contains a result on the uniform consistency of M-estimators depending on a parameter (Proposition 1) and an application of this result

for penalized M-estimation (Theorem 3). In Section 3, a similar result is given under additional convexity assumptions. Section 4 provides a functional Argmin theorem (Theorem 5) applying to contrast processes depending on a parameter. In Section 5 we provide a central limit theorem (CLT) for M-estimators depending on a parameter (Theorem 6) and in Section 6, a CLT for penalized M-estimators (Theorem 7). We apply these consistency and CLT results for the lasso estimator in Section 7, which contains the proofs of Theorems 1 and 2, and an application to statistical hypothesis testing based on the regularization path. Other examples are given in Section 8, including the ℓ^1-penalized general linear model (GLM) introduced in [13] and the penalized least absolute deviation (LAD).

2. Consistency of penalized M-estimators

Standard results on the consistency of M-estimators (see *e.g.* [17, Theorem 5.7]) roughly say that if $\widehat{\boldsymbol{\beta}}_n$ is a sequence of minimizers of M_n on $\boldsymbol{\Phi}$, M_n tends to M with some uniformity and $\boldsymbol{\beta}$ is an isolated minimum of M on $\boldsymbol{\Phi}$, then $\widehat{\boldsymbol{\beta}}_n$ converges to $\boldsymbol{\beta}$ in probability. We will use the following set of conditions which are slightly weaker than the classical ones.

Assumption 3. There exists $\boldsymbol{\beta} \in \boldsymbol{\Phi}$ such that

(i) $\sup_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \{M(\boldsymbol{\phi}) - M_n(\boldsymbol{\phi})\}_+ \xrightarrow{P} 0$, where $a_+ = \max(0, a)$ for any $a \in \mathbb{R}$;

(ii) $M_n(\boldsymbol{\beta}) \xrightarrow{P} M(\boldsymbol{\beta})$;

(iii) for all $\epsilon > 0$, $\inf\{M(\boldsymbol{\phi}) \ : \ \boldsymbol{\phi} \in \boldsymbol{\Phi},\ d(\boldsymbol{\phi}, \boldsymbol{\beta}) \geq \epsilon\} > M(\boldsymbol{\beta})$,

where d is a metric endowing the metric space $\boldsymbol{\Phi}$.

Let us briefly comment these assumptions. Conditions (i) and (ii) are generally replaced by the stronger uniform convergence condition $\sup_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} |M(\boldsymbol{\phi}) - M_n(\boldsymbol{\phi})| \xrightarrow{P} 0$. These weaker conditions are for instance useful when $\boldsymbol{\Phi}$ is non-compact since it is then sufficient to show the uniform convergence on a compact subset and provide a lower bound of M_n out of this compact. Condition (iii) is the standard condition which defines $\boldsymbol{\beta}$ as the (unique) isolated minimum of the limit contrast function.

We will show that, under Assumption 3, provided that $J_n(\beta)$ tends to 0, the minimizer $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ of $\Lambda_n(\boldsymbol{\phi}, \mathbf{t})$ converges to $\boldsymbol{\beta}(\mathbf{t})$, *locally uniformly* in $\mathbf{t}$. To avoid making measurability assumptions on the path $\mathbf{t} \mapsto \widehat{\boldsymbol{\beta}}_n(\mathbf{t})$, we need to work with outer probability to extend the probability to possibly non-measurable sets. Given a probability space $(\Omega, \mathcal{F}, P)$, we denote by P^* the outer probability defined on the subsets of Ω by

$$P^*(A) = \inf\{P(B) : \ B \in \mathcal{F} \text{ with } A \subset B\}, \quad A \subseteq \Omega\ .$$

We say that a sequence (Y_n) of real-valued maps defined on Ω converges in P^*–probability to 0 and denote $Y_n \xrightarrow{P^*} 0$ if, for any $\epsilon > 0$, $P^*(\{|Y_n| \geq \epsilon\}) \to 0$. Here $\{|Y_n| \geq \epsilon\}$ is the usual short-hand notation for the subset $\{\omega \in \Omega \ : \ |Y_n(\omega)| \geq \epsilon\}$. When Y_n is measurable as a map taking values in $\mathbb{R}$ endowed with the Borel σ-field, this is equivalent to the usual convergence in probability.

Theorem 3. *Suppose that Assumption 3 holds for some $\boldsymbol{\beta} \in \boldsymbol{\Phi}$, M defined on $\boldsymbol{\Phi}$ and $\{M_n(\boldsymbol{\phi}),\ \boldsymbol{\phi} \in \boldsymbol{\Phi}\}$, a sequence of real-valued processes. Let (J_n) be a sequence of non-negative functions defined on $\boldsymbol{\Phi}$ such that $J_n(\boldsymbol{\beta}) \to 0$. Let $L \geq 0$ and suppose that we have*

a $\mathbf{\Phi}$-valued process $\{\widehat{\boldsymbol{\beta}}_n(\mathbf{t}),\ \mathbf{t} \geq 0\}$ such that

$$\sup_{\mathbf{t}\in[0,L]} \left\{ \Lambda_n(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}), \mathbf{t}) - \Lambda_n(\boldsymbol{\beta}, \mathbf{t}) \right\}_+ \xrightarrow{P^*} 0 \ , \tag{7}$$

where Λ_n is defined by (1). Then $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ converges to $\boldsymbol{\beta}$ uniformly in $\mathbf{t} \in [0, L]$, in P^*–probability, that is,

$$\sup_{\mathbf{t}\in[0,L]} d(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}), \boldsymbol{\beta}) \xrightarrow{P^*} 0 \ . \tag{8}$$

Remark 2. In statistical applications the contrast function M in Assumption 3 depends on the unknown distribution of the contrast process M_n and thus $\boldsymbol{\beta}$ is an unknown point of $\mathbf{\Phi}$. In particular, the convergence condition $J_n(\boldsymbol{\beta}) \to 0$ has to be verified for any $\boldsymbol{\beta} \in \mathbf{\Phi}$ (but not uniformly in $\boldsymbol{\beta}$) and it simply amounts to correctly normalize the penalty J_n as $n \to \infty$.

Remark 3. The same result holds if the convergence in P-probability in Assumption 3-(i) is replaced by a convergence in P^*-probability. However, in applications, the smoothness properties of $\boldsymbol{\phi} \mapsto M_n(\mathbf{\Phi})$ and $\boldsymbol{\phi} \mapsto M(\mathbf{\Phi})$ usually imply that $\sup_{\boldsymbol{\phi}\in\mathbf{\Phi}} \{M(\boldsymbol{\phi}) - M_n(\boldsymbol{\phi})\}_+$ is a measurable function.

Remark 4. The fact that the outer probability P^* appears in (7) does not bring real difficulties in applications. Indeed Condition (7) follows from the definition of $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ as a near minimizer of $\Lambda_n(\cdot, \mathbf{t})$, that is, if $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ satisfies

$$\Lambda_n(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}), \mathbf{t}) \leq \inf_{\boldsymbol{\phi}\in\mathbf{\Phi}} \Lambda_n(\boldsymbol{\phi}; \mathbf{t}) + u_n \ ,$$

with $u_n = o_P(1)$ not depending on $\mathbf{t}$, *e.g.* $u_n = 0$ (perfect minimizer) or $u_n = n^{-1}$ (near minimizer). The numerical computation of a near minimizer is a difficult task in general, in particular in the presence of several local minima. We will focus on convexity assumptions in Section 3, which cover many cases of interest and which usually allow tractable numerical procedures to compute $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ for any $\mathbf{t}$.

Remark 5. Although $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ is an r.v. for any $\mathbf{t}$, in general the map $\sup_{\mathbf{t}\in[0,L]} \|\widehat{\boldsymbol{\beta}}_n(\mathbf{t}) - \boldsymbol{\beta}\|$ defined on Ω is not measurable (it is in some particular cases, for instance if the map $\mathbf{t} \mapsto \widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ is continuous). This is where the outer probability is useful. Nevertheless, for any $\mathbf{t} \geq 0$, the event $\{\|\widehat{\boldsymbol{\beta}}_n(\mathbf{t}) - \boldsymbol{\beta}\| \geq \epsilon\}$ is (usually) measurable, and its probability is less than the left-hand side of Eq. (8); hence, for any $\mathbf{t} \geq 0$, $\widehat{\boldsymbol{\beta}}_n(\mathbf{t}) \xrightarrow{P} \boldsymbol{\beta}(\mathbf{t})$.

Remark 6. For $L = 0$ in (8), we get a standard result on the consistency of M-estimators (without penalty). It is important to notice that the consistency of penalized M-estimators is obtained for free, in the sense that no additional assumption on M_n or M is required.

Theorem 3 is obtained by applying the following general result on M-estimators depending on a parameter $\mathbf{t} \in \mathrm{T}$.

Proposition 1. *Let $\mathbf{\Phi}$ be a subset of a metric space endowed with the metric d and T be any set. Let Λ be a real-valued function defined on $\mathbf{\Phi} \times \mathrm{T}$, $\{\Lambda_n(\boldsymbol{\phi}, \mathbf{t}),\ \boldsymbol{\phi} \in \mathbf{\Phi}, \mathbf{t} \in \mathrm{T}\}$ be a sequence of real-valued processes, $\boldsymbol{\beta}$ be a $\mathrm{T} \to \mathbf{\Phi}$ map and $\{\widehat{\boldsymbol{\beta}}_n(\mathbf{t}),\ \mathbf{t} \in \mathrm{T}\}$ be a sequence of $\mathbf{\Phi}$-valued processes such that*

(i) $\sup_{\boldsymbol{\phi}\in\mathbf{\Phi}} \sup_{\mathbf{t}\in\mathrm{T}} \{\Lambda(\boldsymbol{\phi}; \mathbf{t}) - \Lambda_n(\boldsymbol{\phi}, \mathbf{t})\}_+ \xrightarrow{P^*} 0$;

(ii) $\sup_{\mathbf{t}\in\mathrm{T}} |\Lambda_n(\boldsymbol{\beta}(\mathbf{t}),\mathbf{t}) - \Lambda(\boldsymbol{\beta}(\mathbf{t}),\mathbf{t})| \xrightarrow{P^*} 0$*;*

(iii) For all $\epsilon > 0$*,*

$$\inf_{\mathbf{t}\in\mathrm{T}} \left[\inf\{\Lambda(\boldsymbol{\phi};\mathbf{t}) \ : \ \boldsymbol{\phi}\in\boldsymbol{\Phi},\ d(\boldsymbol{\phi},\boldsymbol{\beta}(\mathbf{t}))\geq\epsilon\} - \Lambda(\boldsymbol{\beta}(\mathbf{t}),\mathbf{t})\right] > 0\ ;$$

(iv) $\sup_{\mathbf{t}\in\mathrm{T}} \left\{\Lambda_n(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}),\mathbf{t}) - \Lambda_n(\boldsymbol{\beta}(\mathbf{t}),\mathbf{t})\right\}_+ \xrightarrow{P^*} 0.$

Then, $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ *converges to* $\boldsymbol{\beta}(\mathbf{t})$ *uniformly in* $\mathbf{t}\in\mathrm{T}$*, in* P^**–probability, that is,*

$$\sup_{\mathbf{t}\in\mathrm{T}} d(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}),\boldsymbol{\beta}(\mathbf{t})) \xrightarrow{P^*} 0\ . \tag{9}$$

Proof. Let $\epsilon > 0$ and define

$$\alpha = \inf_{\mathbf{t}\in\mathrm{T}} \left[\inf_{d(\boldsymbol{\phi},\boldsymbol{\beta})\geq\epsilon/2} \Lambda(\boldsymbol{\phi};\mathbf{t}) - \Lambda(\boldsymbol{\beta}(\mathbf{t}),\mathbf{t})\right]\ .$$

By (iii), we have $\alpha > 0$. Denote

$$A_n = \left\{\sup_{\mathbf{t}\in\mathrm{T}} d(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}),\boldsymbol{\beta}(\mathbf{t})) \geq \epsilon\right\} \subseteq \Omega\ .$$

For all $\omega\in A_n$, there exists $\mathbf{t}\in\mathrm{T}$ such that $d(\widehat{\boldsymbol{\beta}}_n(\omega,\mathbf{t}),\boldsymbol{\beta}(\mathbf{t})) \geq \epsilon/2$, and thus for which $\Lambda(\widehat{\boldsymbol{\beta}}_n(\omega,\mathbf{t}),\mathbf{t}) - \Lambda(\boldsymbol{\beta}(\mathbf{t}),\mathbf{t}) \geq \alpha$. Hence, for all $\omega\in A_n$, we have

$$\sup_{\mathbf{t}\in\mathrm{T}} \left[\Lambda(\widehat{\boldsymbol{\beta}}_n(\omega,\mathbf{t}),\mathbf{t}) - \Lambda(\boldsymbol{\beta}(\mathbf{t}),\mathbf{t})\right] \geq \alpha\ .$$

Now we write, for any $\mathbf{t}_0\in\mathrm{T}$,

$$\begin{aligned}
\Lambda(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}_0),\mathbf{t}_0) - \Lambda(\boldsymbol{\beta}(\mathbf{t}_0),\mathbf{t}_0) &= \left\{\Lambda(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}_0),\mathbf{t}_0) - \Lambda_n(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}_0),\mathbf{t}_0)\right\} \\
&\quad + \left\{\Lambda_n(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}_0),\mathbf{t}_0) - \Lambda_n(\boldsymbol{\beta}(\mathbf{t}_0),\mathbf{t}_0)\right\} \\
&\quad + \{\Lambda_n(\boldsymbol{\beta}(\mathbf{t}_0),\mathbf{t}_0) - \Lambda(\boldsymbol{\beta}(\mathbf{t}_0),\mathbf{t}_0)\} \\
&\leq \sup_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \sup_{\mathbf{t}\in\mathrm{T}} \{\Lambda(\boldsymbol{\phi};\mathbf{t}) - \Lambda_n(\boldsymbol{\phi},\mathbf{t})\}_+ \\
&\quad + \sup_{\mathbf{t}\in\mathrm{T}} \left\{\Lambda_n(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}),\mathbf{t}) - \Lambda_n(\boldsymbol{\beta}(\mathbf{t}),\mathbf{t})\right\}_+ \\
&\quad + \sup_{\mathbf{t}\in\mathrm{T}} |\Lambda_n(\boldsymbol{\beta}(\mathbf{t}),\mathbf{t}) - \Lambda(\boldsymbol{\beta}(\mathbf{t}),\mathbf{t})|\ .
\end{aligned}$$

Taking the sup in $\mathbf{t}_0\in\mathrm{T}$ we obtain that $A_n \subseteq A_n^{(1)}\cup A_n^{(2)}\cup A_n^{(3)}$, where $A_n^{(1)} = \{\sup_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \sup_{\mathbf{t}\in\mathrm{T}} \{\Lambda(\boldsymbol{\phi};\mathbf{t}) - \Lambda_n(\boldsymbol{\phi},\mathbf{t})\}_+ \geq \alpha/3\}$, and where $A_n^{(2)}$ and $A_n^{(3)}$ are defined accordingly by using the last 2 lines of the last display. Applying $P^*(A_n) \leq P^*(A_n^{(1)}) + P^*(A_n^{(2)}) + P^*(A_n^{(3)})$, (i), (ii) and (iv), we thus get (9), which achieves the proof. □

Proof of Theorem 3. We apply Proposition 1 with $\mathrm{T} = [0,L]$, Λ_n defined by (1), $\Lambda(\boldsymbol{\phi},\mathbf{t}) = M(\boldsymbol{\phi})$ and $\boldsymbol{\beta}(\mathbf{t}) = \boldsymbol{\beta}$ for all $\mathbf{t}$. Let us check the conditions in Proposition 1. Since J_n is non-negative,

$$\{\Lambda(\boldsymbol{\phi};\mathbf{t}) - \Lambda_n(\boldsymbol{\phi};\mathbf{t})\}_+ \leq \{M(\boldsymbol{\phi}) - M_n(\boldsymbol{\phi})\}_+\ ,$$

and Condition (i) follows from Assumption 3-(i). Condition (ii) follows from Assumption 3-(ii) and $J_n(\beta) \to 0$. Conditions (iii) and (iv) directly follow from Assumption 3-(iii) and Eq. (7), respectively. Hence (8) follows from (9) with $\mathrm{T} = [0, L]$. □

3. Consistency in the convex case

In this section, we consider the following assumption.

Assumption 4 (convexity assumption)**.** $\boldsymbol{\Phi}$ is a convex subset of an Euclidean space endowed with the norm $\|\cdot\|$ and M_n is a convex real-valued function on $\boldsymbol{\Phi}$ almost surely. Let $V \subseteq \boldsymbol{\Phi}$ be a neighborhood of the point $\boldsymbol{\beta}$ and Δ be a strictly convex real-valued function defined on V such that

(i) for any $\boldsymbol{\phi} \in V$, $M_n(\boldsymbol{\phi}) \xrightarrow{P} \Delta(\boldsymbol{\phi})$;
(ii) $\Delta(\boldsymbol{\phi}) \geq \Delta(\boldsymbol{\beta})$ for all $\boldsymbol{\phi} \in V$.

Convex M-estimation is considered in [7] and somewhat simplified in [12]. In the following result the convexity assumption is twofold. First it implies Assumption 3. Second, if the penalization J_n is strictly convex, then the minimization of (1) has a unique solution with probability tending to 1 and this solution is continuous in $\mathbf{t}$, which allows to replace the outer probability in (8) by a standard probability. Convexity is also useful in practice since $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ can be computed using standard numerical procedure for convex optimization (see [2]).

Theorem 4. *Suppose that Assumption 4 holds. Let (J_n) be a sequence of non-negative functions defined on $\boldsymbol{\Phi}$ such that $J_n(\boldsymbol{\beta}) \to 0$ and define Λ_n as in (1). Then the 3 following assertions hold.*

(a) For any $L \geq 0$, if we have a $\boldsymbol{\Phi}$-valued process $\{\widehat{\boldsymbol{\beta}}_n(\mathbf{t}),\ \mathbf{t} \geq 0\}$ satisfying (7), $\boldsymbol{\beta}_n(\mathbf{t})$ converges to $\boldsymbol{\beta}$ uniformly in $\mathbf{t} \in [0, L]$, in P^*–probability, that is, (8) holds.

(b) If J_n is strictly convex on $\boldsymbol{\Phi}$, then it is always possible to define a deterministic non-negative sequence (L_n) with $L_n \to \infty$, a sequence (A_n) of events in $\mathcal{F}$ with $P(A_n) \to 1$, and, for each n, a collection $\{\widehat{\boldsymbol{\beta}}_n(\mathbf{t}),\ \mathbf{t} \geq 0\}$ of r.v.'s satisfying the two following properties.

(b1) For all $\mathbf{t} \in [0, L_n]$ and $\omega \in A_n$, $\Lambda_n(\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t}), \mathbf{t})$ is a minimum of $\Lambda_n(\boldsymbol{\phi}, \mathbf{t})$ on $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ and this minimum is unique for $\mathbf{t} > 0$.

(b2) For all $\omega \in \Omega$, $\widehat{\boldsymbol{\beta}}_n(\omega, \cdot)$ is a continuous function on $(0, L_n]$ and on (L_n, ∞).

As consequences, (7) holds for any $L > 0$ and the uniform convergence (8) holds in P–probability, that is,

$$\sup_{\mathbf{t} \in [0, L]} \|\widehat{\boldsymbol{\beta}}_n(\mathbf{t}) - \boldsymbol{\beta}\| \xrightarrow{P} 0\ . \tag{10}$$

(c) If M_n is strictly convex on $\boldsymbol{\Phi}$ for all n, then the conclusions of (b) hold with Properties (b1) and (b2) strengthened as follows.

(c1) For all $\mathbf{t} \in [0, L_n]$ and $\omega \in A_n$, $\Lambda_n(\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t}), \mathbf{t})$ is the unique minimum of $\Lambda_n(\boldsymbol{\phi}, \mathbf{t})$ on $\boldsymbol{\phi} \in \boldsymbol{\Phi}$.

(c2) For all $\omega \in \Omega$, $\widehat{\boldsymbol{\beta}}_n(\omega, \cdot)$ is a continuous function on $[0, L_n]$ and on (L_n, ∞).

Proof. Let $\epsilon > 0$ and denote by $B' = \{\boldsymbol{\phi} : \|\boldsymbol{\phi} - \boldsymbol{\beta}\| \leq 2\epsilon\}$ and $B = \{\boldsymbol{\phi} : \|\boldsymbol{\phi} - \boldsymbol{\beta}\| \leq \epsilon\}$ the balls centered at $\boldsymbol{\beta}$ with radii 2ϵ and ϵ. We choose ϵ small enough so that $B' \subseteq V$. We

first show that Assumption 3 holds for M defined on $\mathbf{\Phi}$ by

$$M(\boldsymbol{\phi}) = \begin{cases} \Delta(\boldsymbol{\phi}) & \text{if } \boldsymbol{\phi} \in B \ , \\ \Delta(\boldsymbol{\beta}) + \alpha/2 & \text{otherwise} \ , \end{cases} \tag{11}$$

where

$$\alpha = \inf_{\boldsymbol{\phi} \in B' \setminus B} \Delta(\boldsymbol{\phi}) - \Delta(\boldsymbol{\beta}) > 0 \ . \tag{12}$$

The positiveness of α follows from the strict convexity of Δ and Assumption 4-(ii). Assumption 3-(ii) follows from Assumption 4-(i). Assumption 3-(iii) follows from the strict convexity of Δ, Assumption 4-(ii) and the definition of M in (11). It only remains to prove that Assumption 3-(i) holds. By [15, Theorem 10.8] and arguing as in the proof of Lemma 3 in [12] for getting the result in the sense of the convergence in probability, the pointwise convergence in Assumption 4-(i) implies the uniform convergence on the compact set B', that is,

$$\sup_{\boldsymbol{\phi} \in B'} |M_n(\boldsymbol{\phi}) - \Delta(\boldsymbol{\phi})| \xrightarrow{P} 0 \ . \tag{13}$$

Let Ω' be a probability 1 set on which M_n is convex and define

$$A_n = \left\{ \sup_{\boldsymbol{\phi} \in B'} |M_n(\boldsymbol{\phi}) - \Delta(\boldsymbol{\phi})| \leq \alpha/4 \right\} \cap \Omega' \ .$$

The set A_n is measurable since M_n and Δ are convex on $\mathbf{\Phi}$ and thus the sup can be replaced by a sup on a countable dense subset of B' without changing the definition of A_n. Let $\omega \in A_n$. For all $\boldsymbol{\phi} \in B' \setminus B$ and $\mathbf{t} \in [0, L]$, we have $M_n(\omega, \boldsymbol{\phi}) \geq \Delta(\boldsymbol{\phi}) - \alpha/4$, $\Delta(\boldsymbol{\phi}) \geq \Delta(\boldsymbol{\beta}) + \alpha$, and, since $\boldsymbol{\beta} \in B'$, $\Delta(\boldsymbol{\beta}) \geq M_n(\omega, \boldsymbol{\beta}) - \alpha/4$. Hence

$$\inf_{\boldsymbol{\phi} \in B' \setminus B} M_n(\omega, \boldsymbol{\phi}) \geq M_n(\omega, \boldsymbol{\beta}) + \alpha/2 \ .$$

By convexity of the function $M_n(\omega, \cdot)$ and of the set $\mathbf{\Phi}$, the last display implies that

$$\inf_{\boldsymbol{\phi} \in \mathbf{\Phi} \setminus B} M_n(\omega, \boldsymbol{\phi}) \geq M_n(\omega, \boldsymbol{\beta}) + \alpha/2 \ .$$

For all $\omega \in A_n$, using the definition of M in (11), we thus have, for all $\boldsymbol{\phi} \in \mathbf{\Phi} \setminus B$,

$$\{M(\boldsymbol{\phi}) - M_n(\omega, \boldsymbol{\phi})\}_+ = \{\Delta(\boldsymbol{\beta}) + \alpha/2 - M_n(\omega, \boldsymbol{\phi})\}_+ \leq |\Delta(\boldsymbol{\beta}) + M_n(\omega, \boldsymbol{\beta})| \ .$$

Using this with (13) and $P(A_n) \to 1$, we get Assumption 3-(i). We conclude that Assumption 3 holds and we obtain Assertion (a) as an application of Theorem 3.

Next we show Assertion (b) and thus assume that J_n is strictly convex. The proof of Assertion (c) is similar and thus omitted. We set

$$L_n = \frac{\alpha}{4 J_n(\boldsymbol{\beta})} \ ,$$

so that $L_n \to \infty$ by assumption on $J_n(\boldsymbol{\beta})$ and $\mathbf{t} J_n(\boldsymbol{\beta}) \leq \alpha/4$ for all $\mathbf{t} \leq L_n$. Let $\omega \in A_n$. Then, for all $\boldsymbol{\phi} \in B' \setminus B$ and $\mathbf{t} \in [0, L_n]$, using that $\Lambda_n(\omega, \boldsymbol{\phi}, \mathbf{t}) \geq M_n(\omega, \boldsymbol{\phi})$ and $M_n(\omega, \boldsymbol{\beta}) = \Lambda_n(\omega, \boldsymbol{\beta}, \mathbf{t}) - \mathbf{t} J_n(\boldsymbol{\beta}) \geq \Lambda_n(\omega, \boldsymbol{\beta}) - \alpha/4$, we obtain

$$\inf_{\mathbf{t} \in [0, L_n]} \inf_{\boldsymbol{\phi} \in B' \setminus B} \Lambda_n(\omega, \boldsymbol{\phi}, \mathbf{t}) \geq \Lambda_n(\omega, \boldsymbol{\beta}, \mathbf{t}) + \alpha/4 \ .$$

Since J_n is strictly convex, so is the function $\Lambda_n(\omega, \cdot, \mathbf{t})$ for $\mathbf{t} > 0$. By convexity of the set $\mathbf{\Phi}$, the previous display implies that for all $\mathbf{t} \in [0, L_n]$, the minimum of $\Lambda_n(\omega, \boldsymbol{\phi}, \mathbf{t})$ on

$\boldsymbol{\phi} \in \boldsymbol{\Phi}$ is attained within B. By strict convexity of J_n, this minimum is unique for $\mathbf{t} > 0$ and we let $\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t})$ be this unique minimum for $\mathbf{t} \in (0, L_n]$. For $\omega \in A_n^c$ (the complementary set of A_n in Ω) or $\mathbf{t} > L_n$, we define $\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t}) = \boldsymbol{\phi}_0$, where $\boldsymbol{\phi}_0$ is any fixed point of $\boldsymbol{\Phi}$. As for $\mathbf{t} = 0$ and $\omega \in A_n$, we define

$$\widehat{\boldsymbol{\beta}}_n(\omega, 0) = \liminf_{\mathbf{t} \downarrow 0} \widehat{\boldsymbol{\beta}}_n(\mathbf{t}) \in B \ ,$$

where the $\liminf$ is defined component-wise in a given coordinate system of the Euclidean space containing $\boldsymbol{\Phi}$. Since the minimum of $\Lambda_n(\omega, \boldsymbol{\phi}, \mathbf{t})$ on $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ is attained within the compact set B, by continuity of $J_n(\boldsymbol{\phi})$ and $M_n(\omega, \boldsymbol{\phi})$ in $\boldsymbol{\phi}$, $\widehat{\boldsymbol{\beta}}_n(\omega, 0)$ is a minimizer of $\Lambda_n(\omega, \boldsymbol{\phi}, 0)$ on $\boldsymbol{\phi} \in \boldsymbol{\Phi}$. Thus, we have defined a r.v. $\widehat{\boldsymbol{\beta}}_n(\cdot, \mathbf{t})$ for any $\mathbf{t} \geq 0$, for which Property (b1) holds.

To conclude the proof, we show that Property (b2) holds. The continuity on (L_n, ∞) for $\omega \in A_n$ and on $\mathbb{R}_+$ for $\omega \in A_n^c$ directly follows from the definition of $\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t})$. Let us now prove that $\widehat{\boldsymbol{\beta}}_n(\omega, \cdot)$ is continuous on $(0, L_n]$ for all $\omega \in A_n$. Since J_n is convex, it is bounded on B and since $\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t}) \in B$, we have $\sup_{\mathbf{t} \in (0, L_n]} J_n(\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t})) \leq \sup J_n(B) < \infty$. Let $\mathbf{t}$ and $\mathbf{t}_0$ be in $(0, L_n]$. We have

$$\begin{aligned}
\Lambda_n(\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t}), \mathbf{t}_0) &\leq \Lambda_n(\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t}), \mathbf{t}) + |\mathbf{t}_0 - \mathbf{t}| \ \sup J_n(B) \\
&\leq \Lambda_n(\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t}_0), \mathbf{t}) + |\mathbf{t}_0 - \mathbf{t}| \ \sup J_n(B) \\
&\leq \Lambda_n(\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t}_0), \mathbf{t}_0) + 2|\mathbf{t}_0 - \mathbf{t}| \ \sup J_n(B) \ .
\end{aligned}$$

Since $\Lambda_n(\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t}_0), \mathbf{t}_0) \leq \Lambda_n(\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t}), \mathbf{t}_0)$, we get that $\Lambda_n(\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t}), \mathbf{t}_0) \to \Lambda_n(\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t}_0), \mathbf{t}_0)$ as $\mathbf{t} \to \mathbf{t}_0$. Since, by strict convexity of Λ_n, $\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t}_0)$ is an isolated minimum of $\Lambda_n(\cdot, \mathbf{t}_0)$, this implies that $\widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t}) \to \widehat{\boldsymbol{\beta}}_n(\omega, \mathbf{t}_0)$ as $\mathbf{t} \to \mathbf{t}_0$. The continuity of $\widehat{\boldsymbol{\beta}}_n(\omega, \cdot)$ on $(0, L_n]$ follows and the proof is achieved. □

Remark 7. The proof of Assertion (c) is somewhat simpler than Assertion (b). However, in some cases, the first purpose of the penalization J_n is precisely to solve an ill-posed problem such as in the ridge regression (see [8]) where $M_n(\boldsymbol{\phi}) = \sum_k (y_k - \mathbf{x}_k^T \boldsymbol{\phi})^2$, $J_n(\boldsymbol{\phi}) \propto \|\boldsymbol{\phi}\|^2$ and the regression matrix $\mathbf{X}_n = [\mathbf{x}_1 \ \ldots \ \mathbf{x}_n]^T$ is not full rank. Thus J_n is strictly convex and M_n is not, in which case Assertion (b) can be useful.

4. An Argmin theorem for contrast processes depending on a parameter

To prove a CLT, we will rely on an Argmin theorem, which is of independent interest, and is adapted from [10] to fit the context of a contrast process depending on a parameter. We will in fact adapt a simpler proof provided by Van der Vaart and Wellner for their similar Theorem 3.2.2 in [18]. Let us recall some of the terminology and notation used in [18]. For a metric space $\mathcal{D}$, we say that a sequence of $\mathcal{D}$-valued maps (X_n) defined on Ω converges weakly to a $\mathcal{D}$-valued map X defined on $(\Omega, \mathcal{F})$, and denote $X_n \rightsquigarrow X$, if X is a Borel map and, for any real-valued bounded continuous function f defined on $\mathcal{D}$,

$$E^*[f(X_n)] \to E[f(X)] \ ,$$

where E denotes the expectation with respect to P and E^* denotes the outer expectation, defined for every real-valued map Z defined on Ω by $E^*[Z] = \inf\{E[U] \ : \ U \geq Z\}$. For any positive integer p and any set T we denote by $\ell^\infty(\mathrm{T}, p)$ the normed space of bounded

functions $f = (f_1, \dots, f_p)$ taking values in $\mathbb{R}^p$ and defined on T endowed with the sup norm on T, denoted by

$$\|f\|_{\mathrm{T}} = \sup_{t \in \mathrm{T}, i \in \{1,\dots,p\}} |f_i(t)| \, .$$

We will simply denote $\ell^\infty(\mathrm{T}, p)$ by $\ell^\infty(\mathrm{T})$ for $p = 1$.

Theorem 5. *Let $\mathbf{\Phi}$ be a metric space endowed with a metric d and T be a parameter set. We suppose that we are in one of the two following cases*

(C-1) T is a finite set. In this case, we set $\mathcal{D} = \mathbf{\Phi}^{\mathrm{T}}$ endowed with the product topology;
(C-2) $\mathbf{\Phi} = \mathbb{R}^p$ with $p \geq 1$, d being the Euclidean metric. In this case, we set $\mathcal{D} = \ell^\infty(\mathrm{T}, p)$.

Let $\{\mathbb{L}_n(\boldsymbol{\phi}, \mathbf{t}),\ \boldsymbol{\phi} \in \mathbf{\Phi}, \mathbf{t} \in \mathrm{T}\}$ be a sequence of real-valued processes, $\{\mathbb{L}(\boldsymbol{\phi}, \mathbf{t}),\ \boldsymbol{\phi} \in \mathbf{\Phi}, \mathbf{t} \in \mathrm{T}\}$ be a real-valued process, $\{\widehat{\boldsymbol{u}}(\mathbf{t}),\ \mathbf{t} \in \mathrm{T}\}$ be a $\mathbf{\Phi}$-valued process, and $\{\widehat{\boldsymbol{u}}_n(\mathbf{t}),\ \mathbf{t} \in \mathrm{T}\}$ be a sequence of $\mathbf{\Phi}$-valued processes. Assume that

(i) for any compact set $K \subset \mathbf{\Phi}$, $\mathbb{L}_n \rightsquigarrow \mathbb{L}$ in $\ell^\infty(K \times \mathrm{T})$ and $\mathbb{L}$ is a tight Borel map taking values in $\ell^\infty(K \times \mathrm{T})$;

(ii) for any $\eta > 0$, we have almost surely that

$$\inf_{\mathbf{t} \in \mathrm{T}} [\inf\{\mathbb{L}(\boldsymbol{\phi}, \mathbf{t}) \ :\ \boldsymbol{\phi} \in \mathbf{\Phi},\ d(\boldsymbol{\phi}, \widehat{\boldsymbol{u}}(\mathbf{t})) \geq \eta\} - \mathbb{L}(\widehat{\boldsymbol{u}}(\mathbf{t}), \mathbf{t})] > 0 \, ; \tag{14}$$

(iii) for any $\epsilon > 0$, there exists a compact $K \subset \mathbf{\Phi}$ such that

$$P\left(\widehat{\boldsymbol{u}}(\mathbf{t}) \in K^c \text{ for all } \mathbf{t} \in \mathrm{T}\right) \leq \epsilon \, ; \tag{15}$$

(iv) for any $\epsilon > 0$, there exists a compact $K \subset \mathbf{\Phi}$ such that

$$\limsup P^*\left(\widehat{\boldsymbol{u}}_n(\mathbf{t}) \in K^c \text{ for all } \mathbf{t} \in \mathrm{T}\right) \leq \epsilon \, ; \tag{16}$$

(v) $\widehat{\boldsymbol{u}}_n$ is approximately minimizing $\mathbb{L}_n$,

$$\sup_{\mathbf{t} \in \mathrm{T}} \left\{ \mathbb{L}_n(\widehat{\boldsymbol{u}}_n(\mathbf{t}), \mathbf{t}) - \inf_{\boldsymbol{\phi} \in \mathbf{\Phi}} \mathbb{L}_n(\boldsymbol{\phi}, \mathbf{t}) \right\}_+ = o_{P^*}(1) \, . \tag{17}$$

Then there is a version of $\widehat{\boldsymbol{u}}$ in $\mathcal{D}$ and $\widehat{\boldsymbol{u}}_n \rightsquigarrow \widehat{\boldsymbol{u}}$.

The case where T is finite is a natural extension of Theorem 3.2.2 in [18]. The second case relies on the first one for obtaining the convergence of finite-dimensional distributions and on a tightness condition, which is more involved to prove.

Proof of Theorem 5. We first consider the case (C-1), $\mathrm{T} = \{\mathbf{t}_1, \dots, \mathbf{t}_q\}$ for some $q \geq 1$. Let $F_1, \dots, F_q$ be some closed subsets of $\mathbf{\Phi}$. Let $\epsilon > 0$ be arbitrarily small. By Conditions (iii) and (iv) there is a compact set K such that (15) and (16) hold. Define the following sequences of subsets of Ω,

$$\begin{aligned} A_n &= \{\widehat{\boldsymbol{u}}_n(\mathbf{t}_i) \in F_i \text{ for all } i \in \{1, \dots, q\}\} \ , \\ B_n &= \{\widehat{\boldsymbol{u}}_n(\mathbf{t}) \in K^c \text{ for all } \mathbf{t} \in \mathrm{T}\} \ , \end{aligned} \tag{18}$$

$$\text{and} \quad C_n = \left\{ \sup_{\mathbf{t} \in \mathrm{T}} \left\{ \mathbb{L}_n(\widehat{\boldsymbol{u}}_n(\mathbf{t}), \mathbf{t}) - \inf_{\boldsymbol{\phi} \in \mathbf{\Phi}} \mathbb{L}_n(\boldsymbol{\phi}, \mathbf{t}) \right\}_+ > \epsilon' \right\} \ , \tag{19}$$

with $\epsilon' > 0$. In $A_n \cap B_n^c \cap C_n^c$, we have, for all $i \in \{1, \dots, q\}$,

$$\inf_{\boldsymbol{\phi} \in F_i \cap K} \mathbb{L}_n(\boldsymbol{\phi}, \mathbf{t}_i) \leq \mathbb{L}_n(\widehat{\boldsymbol{u}}_n(\mathbf{t}_i), \mathbf{t}_i) \leq \inf_{\boldsymbol{\phi} \in \mathbf{\Phi}} \mathbb{L}_n(\boldsymbol{\phi}, \mathbf{t}_i) + \epsilon' \leq \inf_{\boldsymbol{\phi} \in K} \mathbb{L}_n(\boldsymbol{\phi}, \mathbf{t}_i) + \epsilon' \, .$$

Since $\limsup P^*(B_n) \leq \epsilon$ and $\limsup P^*(C_n) = 0$ by Condition (iv) and Eq. (17), applying the continuous mapping Theorem (see [18, Theorem 1.3.6]) and Condition (i) in the previous display with $\epsilon' > 0$ arbitrarily small yields

$$\limsup P^*(A_n) \leq P(A) + \epsilon \ ,$$

where

$$A = \left\{ \inf_{\boldsymbol{\phi} \in F_i \cap K} \mathbb{L}(\boldsymbol{\phi}, \mathbf{t}_i) \leq \inf_{\boldsymbol{\phi} \in K} \mathbb{L}(\boldsymbol{\phi}, \mathbf{t}_i) + \epsilon \text{ for all } i \in \{1, \ldots, q\} \right\} \ .$$

Define

$$\begin{aligned} B &= \{\widehat{\boldsymbol{u}}(\mathbf{t}) \in K^c \text{ for all } \mathbf{t} \in \mathrm{T}\} \\ \text{and} \quad C &= \{\widehat{\boldsymbol{u}}(\mathbf{t}_i) \in F_i \text{ for all } i \in \{1, \ldots, q\}\} \ . \end{aligned} \tag{20}$$

In $A \cap B^c \cap C^c$, there exists $\eta > 0$ such that (14) does not hold. By Condition (ii), this event has probability 0, and, by the above definition of K, $P(B) \leq \epsilon$. Hence $P(A) \leq P(C) + \epsilon$. Letting ϵ tend to 0, we finally get

$$\limsup P^*(A_n) \leq P(C) \ .$$

This implies $\widehat{\boldsymbol{u}}_n \rightsquigarrow \widehat{\boldsymbol{u}}$ in $\boldsymbol{\Phi}^{\mathrm{T}}$ by a slight adaptation of the Portmanteau Theorem (see Theorem 1.3.4 in [18]).

We now consider the case (C-2). Observing that if Conditions (i)–(iv) and Eq. (17) hold for a given set T, then they also hold for any of its finite subsets, the previous case implies the weak convergence $(\widehat{\boldsymbol{u}}_n(\mathbf{t}_1), \ldots, \widehat{\boldsymbol{u}}_n(\mathbf{t}_q)) \rightsquigarrow (\widehat{\boldsymbol{u}}(\mathbf{t}_1), \ldots, \widehat{\boldsymbol{u}}(\mathbf{t}_q))$ for any positive integer q and any $(\mathbf{t}_1, \ldots, \mathbf{t}_q) \in \mathrm{T}^q$. By Theorem 1.5.4 in [18], we thus need to prove that $\widehat{\boldsymbol{u}}_n$ is asymptotically tight in $\ell^\infty(\mathrm{T}, p)$.

By successively applying Lemma 1.3.8 and Theorem 1.5.7 in [18], Condition (i) implies that, for any compact set $K \subset \mathbb{R}^p$, $\mathbb{L}_n$ is asymptotically tight in $\ell^\infty(K \times \mathrm{T})$ and there exists a semi-metric ρ on $K \times \mathrm{T}$ such that $(K \times \mathrm{T}, \rho)$ is totally bounded and $\mathbb{L}_n$ is asymptotically uniformly ρ-equicontinuous in probability. This means that, for any $\epsilon, \epsilon_0 > 0$, there exists $\delta > 0$ such that

$$\limsup P^* \left(\sup_{(\mathbf{u}, \mathbf{u}') \in \mathcal{S}_\delta(K)} |\mathbb{L}_n(\mathbf{u}) - \mathbb{L}_n(\mathbf{u}')| > \epsilon \right) \leq \epsilon_0 \ , \tag{21}$$

where

$$\mathcal{S}_\delta(K) = \left\{ ((\boldsymbol{\phi}, \mathbf{t}), (\boldsymbol{\phi}', \mathbf{t}')) \in (K \times \mathrm{T})^2 \ : \ \rho((\boldsymbol{\phi}, \mathbf{t}), (\boldsymbol{\phi}', \mathbf{t}')) < \delta \right\} \ .$$

Clearly, the semi-metric ρ can be assumed to be bounded and not to depend on the compact set K without loss of generality; in other words, a bounded semi-metric ρ can be defined on $\mathbb{R}^p \times \mathrm{T}$ so that $(\mathbb{R}^p \times \mathrm{T}, \rho)$ is totally bounded and $\mathbb{L}_n$ is asymptotically uniformly ρ-equicontinuous in probability on $K \times \mathrm{T}$ for any compact set K. We shall use this semi-metric in the following to show that $\widehat{\boldsymbol{u}}_n$ is asymptotically uniformly $\tilde{\rho}$-equicontinuous in probability, where $\tilde{\rho}$ is the semi-metric defined on T by

$$\tilde{\rho}(\mathbf{t}, \mathbf{t}') = \sup_{\boldsymbol{\phi} \in \mathbb{R}^p} \rho((\boldsymbol{\phi}, \mathbf{t}), (\boldsymbol{\phi}, \mathbf{t}')) \ .$$

By [18, Theorem 1.5.7 and Theorem 1.5.4], the asymptotic uniform $\tilde{\rho}$-equicontinuity in probability implies that $\widehat{\boldsymbol{u}}_n$ weakly converges to a tight limit in $\ell^\infty(\mathrm{T})$, which has the same finite–dimensional distributions as $\widehat{\boldsymbol{u}}$.

Let us now prove that $\widehat{\boldsymbol{u}}_n$ is asymptotically uniformly $\tilde{\rho}$-equicontinuous in probability. Let η and ϵ_0 be two arbitrarily small positive numbers. Let K be a compact subset of $\mathbb{R}^p$ such that Inequalities (15) and (16) hold. Using Condition (ii), we may find $\epsilon > 0$ arbitrarily small such that

$$P\left(\inf_{\mathbf{t}\in\mathrm{T}}\left[\inf_{\|\boldsymbol{\phi}-\widehat{\boldsymbol{u}}(\mathbf{t})\|\geq\eta/2}\mathbb{L}(\boldsymbol{\phi},\mathbf{t})-\mathbb{L}(\widehat{\boldsymbol{u}}(\mathbf{t}),\mathbf{t})\right]\leq 4\epsilon\right)\leq\epsilon_0\ . \tag{22}$$

We further choose $\delta > 0$ so that Inequality (21) holds. Define B_n as in (18) and C_n as in (19) with $\epsilon' = \epsilon$ and define

$$D_n = \left\{\sup_{\tilde{\rho}(\mathbf{t},\mathbf{t}')\leq\delta}\left\|\widehat{\boldsymbol{u}}_n(\mathbf{t})-\widehat{\boldsymbol{u}}_n(\mathbf{t}')\right\|>\eta\right\}\ ,$$

$$\text{and}\quad E_n = \left\{\sup_{(\mathbf{u},\mathbf{u}')\in\mathcal{S}_\delta(K)}|\mathbb{L}_n(\mathbf{u})-\mathbb{L}_n(\mathbf{u}')|>\epsilon\right\}\ .$$

Hence, with the previous definitions,

$$\limsup P^*(B_n)\leq\epsilon,\ \limsup P^*(C_n)=0\quad\text{and}\quad\limsup P^*(E_n)\leq\epsilon_0\ . \tag{23}$$

On $B_n^c\cap E_n^c$, we have, for any $\mathbf{t},\mathbf{t}'$,

$$\mathbb{L}_n(\widehat{\boldsymbol{u}}_n(\mathbf{t}'),\mathbf{t})\leq\mathbb{L}_n(\widehat{\boldsymbol{u}}_n(\mathbf{t}'),\mathbf{t}')+\epsilon\ . \tag{24}$$

On the set D_n, there exists $\mathbf{t},\mathbf{t}'$ with $\tilde{\rho}(\mathbf{t},\mathbf{t}')\leq\delta$, $\|\widehat{\boldsymbol{u}}_n(\mathbf{t})-\widehat{\boldsymbol{u}}_n(\mathbf{t}')\|>\eta$. On $D_n\cap C_n^c$, we have $\mathbb{L}_n(\widehat{\boldsymbol{u}}_n(\mathbf{t}'),\mathbf{t}')\leq\inf_{\boldsymbol{\phi}\in\boldsymbol{\Phi}}\mathbb{L}_n(\boldsymbol{\phi},\mathbf{t}')+\epsilon$. Intersecting with $B_n^c\cap E_n^c$ and applying (24), we further get

$$\mathbb{L}_n(\widehat{\boldsymbol{u}}_n(\mathbf{t}'),\mathbf{t})\leq\inf_{\boldsymbol{\phi}\in\boldsymbol{\Phi}}\mathbb{L}_n(\boldsymbol{\phi},\mathbf{t}')+2\epsilon\leq\mathbb{L}_n(\widehat{\boldsymbol{u}}_n(\mathbf{t}),\mathbf{t}')+2\epsilon\leq\mathbb{L}_n(\widehat{\boldsymbol{u}}_n(\mathbf{t}),\mathbf{t})+3\epsilon\ ,$$

where the last inequality is obtained by exchanging $\mathbf{t}$ with $\mathbf{t}'$ in (24). Applying again that we are on C_n^c, we have $\mathbb{L}_n(\widehat{\boldsymbol{u}}_n(\mathbf{t}),\mathbf{t})\leq\inf_{\boldsymbol{\phi}\in\boldsymbol{\Phi}}\mathbb{L}_n(\boldsymbol{\phi},\mathbf{t})+\epsilon$, and thus, with the last display, we get

$$\max\left(\mathbb{L}_n(\widehat{\boldsymbol{u}}_n(\mathbf{t}),\mathbf{t}),\mathbb{L}_n(\widehat{\boldsymbol{u}}_n(\mathbf{t}'),\mathbf{t})\right)\leq\inf_{\boldsymbol{\phi}\in\boldsymbol{\Phi}}\mathbb{L}_n(\boldsymbol{\phi},\mathbf{t})+4\epsilon\leq\inf_{\boldsymbol{\phi}\in K}\mathbb{L}_n(\boldsymbol{\phi},\mathbf{t})+4\epsilon\ .$$

Since $\|\widehat{\boldsymbol{u}}_n(\mathbf{t})-\widehat{\boldsymbol{u}}_n(\mathbf{t}')\|>\eta$ and $\widehat{\boldsymbol{u}}_n(\mathbf{t})$ and $\widehat{\boldsymbol{u}}_n(\mathbf{t}')$ belong to K on B_n^c, we just proved that $D_n\cap C_n^c\cap B_n^c\cap E_n^c$ is included in

$$F_n=\left\{\inf_{\mathbf{t}\in\mathrm{T}}\left[\inf_{(\boldsymbol{\phi},\boldsymbol{\phi}')\in\mathcal{B}_\eta(K)}\max\left(\mathbb{L}_n(\boldsymbol{\phi},\mathbf{t}),\mathbb{L}_n(\boldsymbol{\phi}',\mathbf{t})\right)-\inf_{\boldsymbol{\phi}\in K}\mathbb{L}_n(\boldsymbol{\phi},\mathbf{t})\right]\leq 4\epsilon\right\}\ ,$$

where

$$\mathcal{B}_\eta(K)=\left\{(\boldsymbol{\phi},\boldsymbol{\phi}')\in K^2\ :\ \|\boldsymbol{\phi}-\boldsymbol{\phi}'\|>\eta\right\}\ .$$

Using Condition (i) and the continuous mapping Theorem, we have $\limsup P^*(F_n)\leq P(F)$, where

$$F=\left\{\inf_{\mathbf{t}\in\mathrm{T}}\left[\inf_{(\boldsymbol{\phi},\boldsymbol{\phi}')\in\mathcal{B}_\eta(K)}\max\left(\mathbb{L}(\boldsymbol{\phi},\mathbf{t}),\mathbb{L}(\boldsymbol{\phi}',\mathbf{t})\right)-\inf_{\boldsymbol{\phi}\in K}\mathbb{L}(\boldsymbol{\phi},\mathbf{t})\right]\leq 4\epsilon\right\}\ .$$

Since $D_n\cap C_n^c\cap B_n^c\cap E_n^c\subset F_n$, using (23), we further obtain

$$\limsup P^*(D_n)\leq P(F)+\epsilon+\epsilon_0\ .$$

Define B as in (20). On $F \cap B^c$, we have, for any $\mathbf{t} \in \mathrm{T}$, $\inf_{\boldsymbol{\phi} \in K} \mathbb{L}(\boldsymbol{\phi}, \mathbf{t}) \leq \mathbb{L}(\widehat{\boldsymbol{u}}(\mathbf{t}), \mathbf{t})$, and, for all $(\boldsymbol{\phi}, \boldsymbol{\phi}') \in \mathcal{B}_\eta(K)$, since $\|\boldsymbol{\phi} - \widehat{\boldsymbol{u}}(\mathbf{t})\| > \eta/2$ or $\|\boldsymbol{\phi}' - \widehat{\boldsymbol{u}}(\mathbf{t})\| > \eta/2$, $\max\left(\mathbb{L}_n(\boldsymbol{\phi}, \mathbf{t}), \mathbb{L}_n(\boldsymbol{\phi}', \mathbf{t})\right) \geq \inf_{\|\boldsymbol{\phi}'' - \widehat{\boldsymbol{u}}(\mathbf{t})\| > \eta} \mathbb{L}(\boldsymbol{\phi}'', \mathbf{t})$. It follows that $F \cap B^c$ is included in

$$\left\{ \inf_{\mathbf{t} \in \mathrm{T}} \left[\inf_{\|\boldsymbol{\phi} - \widehat{\boldsymbol{u}}(\mathbf{t})\| > \eta/2} \mathbb{L}(\boldsymbol{\phi}, \mathbf{t}) - \mathbb{L}(\widehat{\boldsymbol{u}}(\mathbf{t}), \mathbf{t}) \right] \leq 4\epsilon \right\} ,$$

which, by (22), has probability at most ϵ_0 for our choice of ϵ. Since K has been chosen so that $P(B) \leq \epsilon$, we finally get

$$\limsup P^*(D_n) \leq 2\epsilon + 2\epsilon_0 .$$

Since ϵ is arbitrarily small, this implies that $\widehat{\boldsymbol{u}}_n$ is asymptotically uniformly $\tilde{\rho}$-equicontinuous in probability and the proof is achieved. □

5. Application to M-estimation depending on a parameter

Some general conditions for proving $\sqrt{n}$ asymptotic normality for M-estimators rely on the so called stochastic differentiability condition introduced in [14]. They exploit the idea introduced in [9] of using strong differentiability conditions on the limit contrast function rather than on the contrast process. Moreover it is explained in [14] how the empirical process theory can be used to prove the stochastic differentiability condition. Extensions of these ideas can be found in [18]. We now extend the setting of [14] to a contrast process depending on a parameter. First we obtain the $\sqrt{n}$-rate of convergence in probability; second we apply Theorem 5 to obtain a CLT for M-estimators depending on a parameter. This result will be applied in the context of penalized M-estimation in the next section.

Proposition 2. *Let $\boldsymbol{\Phi}$ be a subset of a metric space endowed with the metric d and T be any set. Let $\{\Lambda_n(\boldsymbol{\phi}, \mathbf{t}),\ \boldsymbol{\phi} \in \boldsymbol{\Phi}, \mathbf{t} \in \mathrm{T}\}$ be a sequence of real-valued processes, $\boldsymbol{\beta}$ be a $\mathrm{T} \to \boldsymbol{\Phi}$ map and $\{\widehat{\boldsymbol{\beta}}_n(\mathbf{t}),\ \mathbf{t} \in \mathrm{T}\}$ be a sequence of $\boldsymbol{\Phi}$-valued processes such that*

$$\sup_{\mathbf{t} \in \mathrm{T}} \left\{ \Lambda_n(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}), \mathbf{t}) - \Lambda_n(\boldsymbol{\beta}(\mathbf{t}), \mathbf{t}) \right\}_+ = O_{P^*}\left(n^{-1}\right) , \tag{25}$$

and the uniform P^*-consistency (9) holds. Assume that we have the following decomposition of the contrast process,

$$\Lambda_n(\boldsymbol{\phi}, \mathbf{t}) - \Lambda_n(\boldsymbol{\beta}(\mathbf{t}), \mathbf{t}) = G_n(\boldsymbol{\phi}, \mathbf{t}) + H(\boldsymbol{\phi}, \mathbf{t}) + d(\boldsymbol{\phi}, \boldsymbol{\beta}(\mathbf{t}))\ R_n(\boldsymbol{\phi}, \mathbf{t}) , \tag{26}$$

where G_n, H and R_n satisfy

(i) $\{G_n(\boldsymbol{\phi}, \mathbf{t}),\ \boldsymbol{\phi} \in \boldsymbol{\Phi}, \mathbf{t} \in \mathrm{T}\}$ is a sequence of real-valued processes such that

$$\sup_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} \sup_{\mathbf{t} \in \mathrm{T}} \frac{n\ |G_n(\boldsymbol{\phi}, \mathbf{t})|}{1 + \sqrt{n}\, d(\boldsymbol{\phi}, \boldsymbol{\beta}(\mathbf{t}))} = O_{P^*}(1) ; \tag{27}$$

(ii) H is a real-valued function defined on $\boldsymbol{\Phi} \times \mathrm{T}$ such that there exists $\epsilon > 0$ for which

$$\inf_{\mathbf{t} \in \mathrm{T}} \inf \left\{ \frac{H(\boldsymbol{\phi}, \mathbf{t})}{d^2(\boldsymbol{\phi}, \boldsymbol{\beta}(\mathbf{t}))} \ :\ \boldsymbol{\phi} \in \boldsymbol{\Phi},\ d(\boldsymbol{\phi}, \boldsymbol{\beta}(\mathbf{t})) \leq \epsilon \right\} > 0 ; \tag{28}$$

(iii) $\{R_n(\boldsymbol{\phi}, \mathbf{t}),\ \boldsymbol{\phi} \in \boldsymbol{\Phi}, \mathbf{t} \in \mathrm{T}\}$ is a sequence of real-valued processes such that, for any positive random sequence (r_n) converging to 0 in P^*–probability,

$$\sup_{\mathbf{t} \in \mathrm{T}} \sup \left\{ |R_n(\boldsymbol{\phi}, \mathbf{t})| \ ;\ \boldsymbol{\phi} \in \boldsymbol{\Phi},\ d(\boldsymbol{\phi}, \boldsymbol{\beta}(\mathbf{t})) \leq r_n \right\} = o_{P^*}(r_n) + O_{P^*}(n^{-1/2}) . \tag{29}$$

Then, $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ converges to $\boldsymbol{\beta}(\mathbf{t})$ uniformly in $\mathbf{t} \in \mathrm{T}$, in P^*–probability, with rate at least $\sqrt{n}$, that is,

$$\sqrt{n} \sup_{\mathbf{t}\in\mathrm{T}} d(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}), \boldsymbol{\beta}(\mathbf{t})) = O_{P^*}(1) \ . \tag{30}$$

Proof. Denote the left-hand side of (30) by U_n and the left-hand side of (27) by V_n. Let $\delta > 1$ and define $A_n = \{U_n > \delta\}$. Then for all $\omega \in A_n$, we have

$$\sup_{\mathbf{t}\in\mathrm{T}} \left| G_n(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}), \mathbf{t}) \right| \le 2n^{-1}\ \delta^{-1}\ U_n^2\ V_n \ . \tag{31}$$

By (iii), using the assumed uniform P^*-consistency (9), there exist non-negative random sequences w_n and W_n such that $w_n = o_{P^*}(1)$, $W_n = O_{P^*}(1)$ and

$$\sqrt{n} \sup_{\mathbf{t}\in\mathrm{T}} \left| R_n(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}), \mathbf{t}) \right| \le (U_n\ w_n + W_n) \ ,$$

hence, for all $\omega \in A_n$,

$$n \sup_{\mathbf{t}\in\mathrm{T}} \left\{ d(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}), \boldsymbol{\beta}(\mathbf{t})) \left| R_n(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}), \mathbf{t}) \right| \right\} \le U_n\ (U_n\ w_n + W_n) \le U_n^2\ (w_n + W_n/\delta) \ .$$

Denote the left-hand side of (25) by S_n. The last display, (31) and (26) imply that, for all $\omega \in A_n$ and all $\mathbf{t} \in \mathrm{T}$,

$$H(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}), \mathbf{t}) \le S_n + U_n^2\ n^{-1}\ \left\{ 2\delta^{-1}\ V_n + w_n + W_n/\delta \right\} \ .$$

Define $B_n = \{\sup_{\mathbf{t}\in\mathrm{T}} d(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}), \boldsymbol{\beta}(\mathbf{t})) > \epsilon\}$ where ϵ is the positive number in Condition (ii) and denote the left-hand side of (28) by α, which is positive. Then, for all $\omega \in B_n^c$, $\alpha\ U_n^2 \le n\ \sup_{\mathbf{t}\in\mathrm{T}} H(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}), \mathbf{t})$, and, using the previous display, if moreover $\omega \in A_n$,

$$\alpha\ U_n^2 \le n\ S_n + U_n^2\ \left\{ 2\delta^{-1}\ V_n + w_n + W_n/\delta \right\} \ .$$

Using that $P^*(B_n) \to 0$, $nS_n = O_{P^*}(1)$, $V_n = O_{P^*}(1)$, $w_n = o_{P^*}(1)$ and $W_n = O_{P^*}(1)$, we easily get that $\limsup P^*(A_n)$ can be made arbitrarily small by taking δ large enough. Hence (30) holds. □

Applying Proposition 2 and Theorem 5, we get the following result.

Theorem 6. *Let $\boldsymbol{\Phi} = \mathbb{R}^p$, $p \ge 1$, and T be any set. Let $\{\Lambda_n(\boldsymbol{\phi}, \mathbf{t}),\ \boldsymbol{\phi} \in \boldsymbol{\Phi}, \mathbf{t} \in \mathrm{T}\}$ be a sequence of real-valued processes, $\boldsymbol{\beta}$ be a $\mathrm{T} \to \boldsymbol{\Phi}$ map and $\{\widehat{\boldsymbol{\beta}}_n(\mathbf{t}),\ \mathbf{t} \in \mathrm{T}\}$ be a sequence of $\boldsymbol{\Phi}$-valued processes such that*

$$\sup_{\mathbf{t}\in\mathrm{T}} \left\{ \Lambda_n(\widehat{\boldsymbol{\beta}}_n(\mathbf{t}), \mathbf{t}) - \Lambda_n(\boldsymbol{\beta}(\mathbf{t}), \mathbf{t}) \right\}_+ = o_{P^*}\left(n^{-1}\right) \ , \tag{32}$$

and the uniform P^*-consistency (9) holds. Assume that the decomposition (26) of the contrast process holds where G_n, H and R_n satisfy:

(i) $\{G_n(\boldsymbol{\phi}, \mathbf{t}),\ \boldsymbol{\phi} \in \boldsymbol{\Phi}, \mathbf{t} \in \mathrm{T}\}$ is a sequence of real-valued processes satisfying (27);

(ii) H is real-valued function defined on $\boldsymbol{\Phi} \times \mathrm{T}$ and there exists a function Γ defined on T and taking values in the set of non-negative symmetric $p \times p$ matrices such that, denoting by $\lambda_{\min}(\Gamma(\mathbf{t}))$ and $\lambda_{\max}(\Gamma(\mathbf{t}))$ the smallest and largest eigenvalues of $\Gamma(\mathbf{t})$,

$$0 < \inf\{\lambda_{\min}(\Gamma(\mathbf{t})),\ \mathbf{t} \in \mathrm{T}\} < \sup\{\lambda_{\max}(\Gamma(\mathbf{t})),\ \mathbf{t} \in \mathrm{T}\} < \infty \ , \tag{33}$$

and, as $\boldsymbol{\phi} \to \boldsymbol{\beta}$ in $\ell^\infty(\mathrm{T}, p)$,

$$\left\| H(\boldsymbol{\phi}(\cdot), \cdot) - (\boldsymbol{\phi} - \boldsymbol{\beta})^T \Gamma (\boldsymbol{\phi} - \boldsymbol{\beta}) \right\|_{\mathrm{T}} = o\left(\|\boldsymbol{\phi} - \boldsymbol{\beta}\|_{\mathrm{T}}^2 \right) \ ; \tag{34}$$

(iii) $\{R_n(\boldsymbol{\phi}, \mathbf{t}),\ \boldsymbol{\phi} \in \boldsymbol{\Phi}, \mathbf{t} \in \mathrm{T}\}$ *is a sequence of real-valued processes such that, for any positive random sequence* (r_n) *converging to 0 in* P^**–probability,*

$$\sup_{\mathbf{t}\in\mathrm{T}} \sup \left\{ |R_n(\boldsymbol{\phi}, \mathbf{t})| \ ;\ \boldsymbol{\phi} \in \boldsymbol{\Phi},\ d(\boldsymbol{\phi}, \boldsymbol{\beta}(\mathbf{t})) \le r_n \right\} = o_{P^*}(r_n) + o_{P^*}(n^{-1/2}) \ . \tag{35}$$

Let us further define

$$\widehat{G}_n(\boldsymbol{\phi}, \mathbf{t}) = nG_n \left(\boldsymbol{\beta}(\mathbf{t}) + n^{-1/2}\boldsymbol{\phi}, \mathbf{t} \right) \ , \tag{36}$$

and assume that there exists a real-valued process $\{G(\boldsymbol{\phi}, \mathbf{t}),\ \boldsymbol{\phi} \in \boldsymbol{\Phi},\ \mathbf{t} \in \mathrm{T}\}$ *such that, for any compact* $K \subset \boldsymbol{\Phi}$*,* G *is tight in* $\ell^\infty(K \times \mathrm{T}, p)$ *and* $\widehat{G}_n \rightsquigarrow G$ *in* $\ell^\infty(K \times \mathrm{T}, p)$*. Define*

$$\mathbb{L}(\boldsymbol{\phi}, \mathbf{t}) = G\left(\boldsymbol{\phi}, \mathbf{t}\right) + \boldsymbol{\phi}^T \Gamma(\mathbf{t}) \boldsymbol{\phi} \ , \tag{37}$$

and assume that there exists a $\boldsymbol{\Phi}$*-valued process* $\{\widehat{\boldsymbol{u}}(\mathbf{t}),\ \mathbf{t} \in \mathrm{T}\}$ *such that Conditions (ii) and (iii) in Theorem 5 hold. Then there is a version of* $\widehat{\boldsymbol{u}}$ *in* $\ell^\infty(\mathrm{T}, p)$ *and*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \rightsquigarrow \widehat{\boldsymbol{u}} \ . \tag{38}$$

Remark 8. Observe that Eq. (32) is a strengthened version of (30) and that (33) and (34) imply (28). Hence Conditions (i)–(iii) in Theorem 6 imply Conditions (i)–(iii) in Proposition 2.

Proof. Let us define $\widehat{\boldsymbol{u}}_n = \sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ and

$$\mathbb{L}_n(\boldsymbol{\phi}, \mathbf{t}) = n \left\{ \Lambda_n(\boldsymbol{\beta}(\mathbf{t}) + n^{-1/2}\boldsymbol{\phi}, \mathbf{t}) - \Lambda_n(\boldsymbol{\beta}(\mathbf{t}), \mathbf{t}) \right\} \ . \tag{39}$$

We will apply Theorem 5 with these definitions (in the case (C-2)) and thus now proceed in checking the conditions of Theorem 5 successively. Let K be a compact subset of $\boldsymbol{\Phi}$. Using (26), (36) and (39), we get

$$\mathbb{L}_n(\boldsymbol{\phi}, \mathbf{t}) = \widehat{G}_n(\boldsymbol{\phi}, t) + nH \left(\boldsymbol{\beta}(\mathbf{t}) + n^{-1/2}\boldsymbol{\phi}, \mathbf{t} \right) + \sqrt{n}\|\boldsymbol{\phi}\| R_n \left(\boldsymbol{\beta}(\mathbf{t}) + n^{-1/2}\boldsymbol{\phi}, \mathbf{t} \right) \ .$$

Observe that by (33) and (34), as functions of $(\boldsymbol{\phi}, \mathbf{t})$,

$$nH \left(\boldsymbol{\beta}(\mathbf{t}) + n^{-1/2}\boldsymbol{\phi}, \mathbf{t} \right) \to \boldsymbol{\phi}^T \Gamma(\mathbf{t}) \boldsymbol{\phi} \quad \text{in} \quad \ell^\infty(K \times \mathrm{T}, p) \ .$$

Applying (35), we obtain

$$\sup_{(\boldsymbol{\phi},\mathbf{t})\in K\times\mathrm{T}} \sqrt{n}\|\boldsymbol{\phi}\| \left| R_n \left(\boldsymbol{\beta}(\mathbf{t}) + n^{-1/2}\boldsymbol{\phi}, \mathbf{t} \right) \right| = o_{P^*}(1).$$

Hence using that $\widehat{G}_n \rightsquigarrow G$ in $\ell^\infty(K \times \mathrm{T}, p)$, the three last displays yield $\mathbb{L}_n \rightsquigarrow \mathbb{L}$ in $\ell^\infty(K \times \mathrm{T}, p)$. Since G is tight in $\ell^\infty(K \times \mathrm{T}, p)$ by assumption, $\mathbb{L}$ also is and thus Condition (i) holds. Conditions (ii) and (iii) hold by assumption. Applying Proposition 2, we obtain (30) and thus Condition (iv) holds. Using (32) with the above definitions, we get that Condition (v) holds. □

6. APPLICATION TO PENALIZED M-ESTIMATION

We now apply Theorem 6 for extending Pollard's theorem in [14]. We will show that if the $\sqrt{n}$ asymptotic normality conditions in [14] are verified and if the penalty is reasonable then the penalized version of the M-estimator satisfies a CLT similar to the CLT in [11] for the mean square criterion. Moreover this CLT applies to the regularization path in a functional sense. In [14], Pollard proves the asymptotic normality of M-estimators based on a contrast process of the form

$$M_n(\boldsymbol{\phi}) = n^{-1} \sum_{k=1}^{n} g(\xi_k, \boldsymbol{\phi}) = P_n g(\cdot, \boldsymbol{\phi}) \ , \tag{40}$$

where (ξ_k) is a sequence of $\mathcal{X}$-valued random variables and g is a $\mathcal{X} \times \mathbb{R}^p$ function satisfying the following Taylor expansion around a given point $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$g(x, \boldsymbol{\phi}) = g(x, \boldsymbol{\beta}) + (\boldsymbol{\phi} - \boldsymbol{\beta})^T \Delta(x) + \|\boldsymbol{\phi} - \boldsymbol{\beta}\| \, r(x, \boldsymbol{\phi}) \ . \tag{41}$$

Let us recall Pollard's conditions that we will use on the contrast process M_n.

(P-1) (ξ_k) is a sequence of i.i.d. random variables with distribution P;
(P-2) the function $M(\boldsymbol{\phi}) = Pg(\cdot, \boldsymbol{\phi})$ has a nonsingular second derivative Γ at $\boldsymbol{\beta} \in \mathbb{R}^p$;
(P-3) $P\|\Delta\|^2 < \infty$ and $P\Delta = 0$;
(P-4) the stochastic differentiability condition holds on r, that is, for any sequence of positive r.v. (r_n) such that $r_n \xrightarrow{P} 0$,

$$\sup_{\|\boldsymbol{\phi} - \boldsymbol{\beta}\| \le r_n} \frac{|\nu_n \, r(\cdot, \boldsymbol{\phi})|}{1 + \sqrt{n}\|\boldsymbol{\phi} - \boldsymbol{\beta}\|} \xrightarrow{P} 0 \ . \tag{42}$$

Here we used the notations, standard in the empirical process literature, Pf, $P_n f$ and $\nu_n f$ for $\int f \mathrm{d}P$, $n^{-1} \sum_{k=1}^{n} f(\xi_k)$ and $\sqrt{n}(P_n f - Pf)$, respectively. Theorem 7 below provides a central limit theorem for the regularization path defined on the penalized contrast (1) when M_n satisfies Pollard's conditions (P-1)–(P-4) with some mild conditions on the penalty J_n.

Theorem 7. *Let $\boldsymbol{\Phi} = \mathbb{R}^p$, $p \ge 1$ and $\mathrm{T} = [0, L]$, with $L > 0$. Define Λ_n as in (1), where M_n is defined by (40) and satisfies Pollard's conditions (P-1)–(P-4) and J_n is a sequence of deterministic non-negative functions defined on $\mathbb{R}^p$. Further assume that there exists a positive constant C such that*

$$n \ |J_n(\boldsymbol{\phi}) - J_n(\boldsymbol{\beta})| \le C \, (1 + \sqrt{n} \, ||\boldsymbol{\phi} - \boldsymbol{\beta}||) \quad \textit{for} \quad \|\boldsymbol{\phi} - \boldsymbol{\beta}\| \le 1 \ , \tag{43}$$

and, for any compact $K \subset \mathbb{R}^p$,

$$\sup_{\boldsymbol{\phi} \in K} \left| n \ J_n(\boldsymbol{\beta} + n^{-1/2}\boldsymbol{\phi}) - n \ J_n(\boldsymbol{\beta}) - J_\infty(\boldsymbol{\phi}) \right| \to 0 \ . \tag{44}$$

Let $\{\widehat{\boldsymbol{\beta}}_n, \ \mathbf{t} \in \mathrm{T}\}$ be a sequence of $\boldsymbol{\Phi}$-valued processes such that (32) and the uniform P^*-consistency (9) hold. Let W be a centered Gaussian p-dimensional vector with covariance $P(\Delta\Delta^T)$ and define

$$\mathbb{L}(\boldsymbol{\phi}, \mathbf{t}) = W^T \phi + \boldsymbol{\phi}^T \Gamma \boldsymbol{\phi} + \mathbf{t} J_\infty(\boldsymbol{\phi}) \ . \tag{45}$$

Finally assume that there exists a $\boldsymbol{\Phi}$-valued process $\{\widehat{\boldsymbol{u}}(\mathbf{t}), \ \mathbf{t} \in \mathrm{T}\}$ such that Conditions (ii) and (iii) in Theorem 5 hold. Then there is a version of $\widehat{\boldsymbol{u}}$ in $\ell^\infty(\mathrm{T}, p)$ and

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \rightsquigarrow \widehat{\boldsymbol{u}} \ . \tag{46}$$

Proof. We shall apply Theorem 6 for Λ_n given by (1) and with $\boldsymbol{\beta}(\mathbf{t}) = \boldsymbol{\beta}$ for all $\mathbf{t} \in \mathrm{T}$. Let us check that the assumptions of this theorem hold in this context. Condition (32) and the uniform P^*-consistency (9) hold by assumption. The decomposition (26) holds with

$$G_n(\boldsymbol{\phi}, \mathbf{t}) = (\boldsymbol{\phi} - \boldsymbol{\beta})^T P_n \Delta + \mathbf{t} \left(J_n(\boldsymbol{\phi}) - J_n(\boldsymbol{\beta})\right) \mathbb{1}(\|\boldsymbol{\phi} - \boldsymbol{\beta}\| \leq 1) \ ,$$
$$H(\boldsymbol{\phi}, \mathbf{t}) = Pg(\cdot, \boldsymbol{\phi}) - Pg(\cdot, \boldsymbol{\beta}) - (\boldsymbol{\phi} - \boldsymbol{\beta})^T P\Delta \ ,$$
$$R_n(\boldsymbol{\phi}, \mathbf{t}) = n^{-1/2} \nu_n \, r(\cdot, \boldsymbol{\phi}) + \mathbf{t} \|\boldsymbol{\phi} - \boldsymbol{\beta}\|^{-1} \left(J_n(\boldsymbol{\phi}) - J_n(\boldsymbol{\beta})\right) \mathbb{1}(\|\boldsymbol{\phi} - \boldsymbol{\beta}\| > 1) \ .$$

Using (P-1) and (P-3), we have $\sum_{k=1}^n \Delta(\xi_k) = O_P(n^{1/2})$ and, using (43), we get that Condition (i) in Theorem 6 holds. Observe that $H(\boldsymbol{\phi}, \mathbf{t})$ does not depend on $\mathbf{t}$ and, by (P-3), we have

$$H(\boldsymbol{\phi}, \mathbf{t}) = M(\boldsymbol{\phi}) - M(\boldsymbol{\beta}) \ .$$

Integrating x with respect to P in (41) and using (P-4), we get that the first derivative of M at $\boldsymbol{\beta}$ is zero and, by (P-2),

$$H(\boldsymbol{\phi}, \mathbf{t}) = (\boldsymbol{\phi} - \boldsymbol{\beta})^T \Gamma (\boldsymbol{\phi} - \boldsymbol{\beta}) + o\left(\|\boldsymbol{\phi} - \boldsymbol{\beta}\|^2\right) \ .$$

Hence Condition (ii) in Theorem 6 holds.

We have, for any sequence of positive r.v. (r_n) such that $r_n \xrightarrow{P} 0$,

$$\sup_{\|\boldsymbol{\phi} - \boldsymbol{\beta}\| \leq r_n} \left\{ \left| n^{-1/2} \nu_n \, r(\cdot, \boldsymbol{\phi}) \right| \right\} \leq \frac{1 + \sqrt{n} r_n}{\sqrt{n}} \sup_{\|\boldsymbol{\phi} - \boldsymbol{\beta}\| \leq r_n} \left\{ \frac{|\nu_n \, r(\cdot, \boldsymbol{\phi})|}{1 + \sqrt{n} \|\boldsymbol{\phi} - \boldsymbol{\beta}\|} \right\}$$
$$= o_P(n^{-1/2}) + o_P(r_n) \ ,$$

where the last equality follows from (P-4). Observing that, for $\|\boldsymbol{\phi} - \boldsymbol{\beta}\| \leq r_n$ and $r_n \leq 1$ the second term defining R_n vanishes, we obtain Condition (35) in Theorem 6.

Defining $\widehat{G}_n$ as in (36) gives

$$\widehat{G}_n(\boldsymbol{\phi}, \mathbf{t}) = \boldsymbol{\phi}^T \left(\sqrt{n} P_n \Delta\right) + \mathbf{t} \left[n \, J_n(\boldsymbol{\beta} + n^{-1/2} \boldsymbol{\phi}) - n \, J_n(\boldsymbol{\beta}) \right] \ .$$

Using (P-1) and (P-3), we have that $\sqrt{n} P_n \Delta$ converge in distribution to W and, by (44), for any compact $K \subset \mathbb{R}^p$ $\widehat{G}_n \rightsquigarrow G$ in $\ell^\infty(K \times \mathrm{T}, p)$, where

$$G(\boldsymbol{\phi}, \mathbf{t}) = \boldsymbol{\phi}^T W + \mathbf{t} \, J_\infty(\boldsymbol{\phi}) \ .$$

This definition of G and (37) gives (45). Hence Theorem 6 yields (46). □

The following lemma shows that the penalties considered in [11] satisfy Conditions (43) and (44).

Lemma 1. *Let $\gamma > 0$ and define, for all $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p) \in \mathbb{R}^p$,*

$$J_n^{(\gamma)}(\boldsymbol{\phi}) = n^{(1 \wedge \gamma)/2 - 1} \sum_{k=1}^p |\phi_k|^\gamma \ . \tag{47}$$

Then for any $\boldsymbol{\beta} \in \mathbb{R}^p$, there exists $C > 0$ such that, for all $\boldsymbol{\phi} \in \mathbb{R}^p$,

$$n \left| J_n^{(\gamma)}(\boldsymbol{\phi}) - J_n^{(\gamma)}(\boldsymbol{\beta}) \right| \leq C \left(1 + \sqrt{n} \|\boldsymbol{\phi} - \boldsymbol{\beta}\| + \sqrt{n} \|\boldsymbol{\phi} - \boldsymbol{\beta}\|^{1 \vee \gamma}\right) \ , \tag{48}$$

and, for any compact $K \subset \mathbb{R}^p$,

$$\sup_{\boldsymbol{\phi} \in K} \left| n J_n^{(\gamma)}(\boldsymbol{\beta} + n^{-1/2} \boldsymbol{\phi}) - n J_n^{(\gamma)}(\boldsymbol{\beta}) - J_\infty^{(\gamma)}(\boldsymbol{\phi}) \right| \to 0 \ , \tag{49}$$

where

$$J_\infty^{(\gamma)}(\boldsymbol{\phi}) = \begin{cases} \sum_{j=1}^p |\phi_j|^\gamma \mathbb{1}_{\{\beta_j=0\}} & \text{if } \gamma < 1 \\ \sum_{j=1}^p \left\{ \phi_j \operatorname{sgn}(\beta_j)\, \mathbb{1}_{\{\beta_j\neq 0\}} + |\phi_j| \mathbb{1}_{\{\beta_j=0\}} \right\} & \text{if } \gamma = 1 \\ \gamma \sum_{j=1}^p \phi_j \operatorname{sgn}(\beta_j)\, |\beta_j|^{\gamma-1} \mathbb{1}_{\{\beta_j\neq 0\}} & \text{if } \gamma > 1. \end{cases} \tag{50}$$

Remark 9. The limit penalties in (50) correspond to those in Theorems 2 and 3 in [11], except for the multiplicative constant γ in the case $\gamma > 1$, which seems to have been forgotten in [11].

Proof. We have, for all $\boldsymbol{\phi} \in \mathbb{R}^p$,

$$\left| \sum_{k=1}^p |\phi_k|^\gamma - \sum_{k=1}^p |\beta_k|^\gamma \right| \le C \left(\|\boldsymbol{\phi} - \boldsymbol{\beta}\|^\gamma + \|\boldsymbol{\phi} - \boldsymbol{\beta}\| \right) ,$$

where C only depends on $\boldsymbol{\beta}$ and $\gamma > 0$. The bound (48) follows directly for $\gamma \ge 1$. For $\gamma < 1$, one obtains

$$n \left| J_n^{(\gamma)}(\boldsymbol{\phi}) - J_n^{(\gamma)}(\boldsymbol{\beta}) \right| \le C' \left((\sqrt{n}\|\boldsymbol{\phi} - \boldsymbol{\beta}\|)^\gamma + n^{\gamma/2} \|\boldsymbol{\phi} - \boldsymbol{\beta}\| \right) ,$$

and (48) follows by oberving that $a^\gamma \le 1 + a$ for $a \ge 0$, and $n^{\gamma/2} \le n^{1/2}$.

Relation (49) is easily obtained by using the Taylor expansion, valid for $x \neq 0$, $|x+y|^\gamma = |x|^\gamma + \gamma |x|^{\gamma-1} \operatorname{sgn}(x)\, y + O(y^2)$, which concludes the proof. □

7. Application to the lasso and hypothesis testing based on the regularization path

We are now in a position to prove Theorems 1 and 2. We next give a simple application for testing the null hypothesis $H_0 : \boldsymbol{\beta} = 0$ using a statistic based on the regularization path.

Proof of Theorem 1. As $\boldsymbol{\phi} \mapsto M_n(\boldsymbol{\phi}) = \frac{1}{n} \sum_{k=1}^n (y_k - \mathbf{x}_k^T \boldsymbol{\phi})^2$ is a convex function, we apply Theorem 4. In fact, by Assumption 1-(i), M_n is strictly convex for n large enough, and hence the more precise Assertion (c) applies. We now show that Assumption 4-(i) holds.

$$M_n(\boldsymbol{\phi}) - M_n(\boldsymbol{\beta}) = (\boldsymbol{\phi} - \boldsymbol{\beta})^T C_n (\boldsymbol{\phi} - \boldsymbol{\beta}) - \frac{2}{n} \varepsilon_n^T \mathbf{X}_n (\boldsymbol{\phi} - \boldsymbol{\beta}) \tag{51}$$

where $\varepsilon_n = Y_n - \mathbf{X}_n \boldsymbol{\beta}$. Since

$$\mathbb{E} \|\mathbf{X}_n^T \varepsilon_n\|^2 = \mathbb{E} \left[\operatorname{Tr}(\varepsilon_n^T \mathbf{X}_n \mathbf{X}_n^T \varepsilon_n) \right] = \operatorname{Tr} \left[\mathbf{X}_n \mathbf{X}_n^T \right] = O(n) ,$$

by Assumption 1-(i), it comes $-\frac{2}{n} \varepsilon_n^T \mathbf{X}_n (\boldsymbol{\phi} - \boldsymbol{\beta}) = O_P(n^{-1/2})$. And furthermore, by Assumption 1–(i) :

$$M_n(\boldsymbol{\phi}) - M_n(\boldsymbol{\beta}) \to_P (\boldsymbol{\phi} - \boldsymbol{\beta})^T C (\boldsymbol{\phi} - \boldsymbol{\beta}) = \Delta(\boldsymbol{\phi}) .$$

Since C is positive-definite, Δ is strictly convex and Assumption 4-(ii) holds. By definition of $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$, (7) holds. Finally, the condition $J_n(\boldsymbol{\beta}) \to 0$ holds, as the penalty is defined by $J_n(\boldsymbol{\beta}) = \lambda_n \|\boldsymbol{\beta}\|_1$, with $\|\cdot\|_1$ denoting the ℓ^1 norm. Uniform consistency follows as an application of Theorem 4. □

Proof of Theorem 2. We apply Theorem 6 with $\mathrm{T} = [0, L]$. By definition of $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$, condition (32) holds. We just obtained uniform consistency in Theorem 1. Using (51), we have the decomposition (26) of $\Lambda_n(\boldsymbol{\phi}, \mathbf{t})$, with

$$\begin{aligned} G_n(\boldsymbol{\phi}, \mathbf{t}) &= -2n^{-1/2} U_n^T(\boldsymbol{\phi} - \boldsymbol{\beta}) + \mathbf{t}\lambda_n \left(\|\boldsymbol{\phi}\|_1 - \|\boldsymbol{\beta}\|_1\right) \\ H(\boldsymbol{\phi}, \mathbf{t}) &= (\boldsymbol{\phi} - \boldsymbol{\beta})^T C(\boldsymbol{\phi} - \boldsymbol{\beta}) \\ R_n(\boldsymbol{\phi}, \mathbf{t}) &= \|\boldsymbol{\phi} - \boldsymbol{\beta}\|^{-1} (\boldsymbol{\phi} - \boldsymbol{\beta})^T (C_n - C)(\boldsymbol{\phi} - \boldsymbol{\beta}) \end{aligned}$$

where $U_n = n^{-1/2}\mathbf{X}_n^T \varepsilon_n$ and $\lambda_n = n^{-1/2}$, by Assumption 2-(iii).

The sequence $\{U_n\}$ converges in distribution to $U \sim \mathcal{N}(0, \sigma^2 C)$ by the Lindeberg-Feller theorem and Assumption 2. We have, for all $\boldsymbol{\phi} \in \mathbb{R}^p$ and $\mathbf{t} \in [0, L]$, $n|G_n(\boldsymbol{\phi}, \mathbf{t})| \leq \sqrt{n} U_n \|\boldsymbol{\phi} - \boldsymbol{\beta}\| + \mathbf{t}\sqrt{n} |\|\boldsymbol{\phi}\|_1 - \|\boldsymbol{\beta}\|_1| \leq \|\boldsymbol{\phi} - \boldsymbol{\beta}\| (O_P(\sqrt{n}) + cL\sqrt{n})$. Hence G_n satisfies (27).

Conditions (33) and (34) on H are immediately verified by taking $\Gamma(\mathbf{t}) = C$, for all $\mathbf{t} \in \mathrm{T}$ and using Assumption 2-(i).

Observe that $|R_n(\boldsymbol{\phi}, \mathbf{t})| \leq \rho(C_n - C)\, \|\boldsymbol{\phi} - \boldsymbol{\beta}\|$ where $\rho(C_n - C)$ is the spectral radius of $(C_n - C)$. Since $C_n \xrightarrow{P} C$, $\rho(C_n - C) = o_P(1)$ and $\sup \{R_n(\boldsymbol{\phi}, \mathbf{t}), \boldsymbol{\phi} \in \boldsymbol{\Phi}, \|\boldsymbol{\phi} - \boldsymbol{\beta}\| \leq r_n\} = o_P(r_n)$. Condition (35) on R_n follows.

As in (36), we define

$$\begin{aligned} \widehat{G}_n(\boldsymbol{\phi}, \mathbf{t}) &= nG_n\left(\boldsymbol{\beta} + n^{-1/2}\boldsymbol{\phi}, \mathbf{t}\right) \\ &= -2U_n^T \boldsymbol{\phi} + \mathbf{t} n^{1/2} \sum_{j=1}^{p} \left\{ \left| \beta_j + n^{-1/2}\phi_j \right| - |\beta_j| \right\} . \end{aligned}$$

For any compact $K \subseteq \mathbb{R}^p$, let f map $u \in \mathbb{R}^p$ to $f[u] \in \ell^\infty(K \times \mathrm{T})$, defined by $f[u](\boldsymbol{\phi}, \mathbf{t}) = u^T \boldsymbol{\phi}$. The map f is continuous and by the continuous mapping theorem, $f(U_n)$ converges to $f(U)$ in $\ell^\infty(K \times \mathrm{T})$. From this and (49) with $\gamma = 1$, it follows that $\widehat{G}_n$ converges to G in $\ell^\infty(K \times \mathrm{T})$, where

$$G(\boldsymbol{\phi}, \mathbf{t}) = -2U^T \boldsymbol{\phi} + \mathbf{t} \sum_{j=1}^{p} \left\{ \phi_j \operatorname{sgn}(\beta_j)\, \mathbb{1}_{\{\beta_j \neq 0\}} + |\phi_j| \mathbb{1}_{\{\beta_j = 0\}} \right\} .$$

By Assumption 1-(i) one has $\mathbb{L}(\boldsymbol{\phi}, \mathbf{t}) \geq c_1 \|\boldsymbol{\phi}\|^2 + c_2 \|\boldsymbol{\phi}\|$ for all $\boldsymbol{\phi} \in \mathbb{R}^p$ and $\mathbf{t} \in [0, L]$, with $c_1 > 0$ and c_2 a finite random variable. Since $\mathbb{L}(0, \mathbf{t}) = 0$, we get $0 \geq \mathbb{L}(\widehat{\boldsymbol{u}}(\mathbf{t}), \mathbf{t}) \geq c_1 \|\widehat{\boldsymbol{u}}(\mathbf{t})\|^2 + c_2 \|\widehat{\boldsymbol{u}}(\mathbf{t})\|$ thus $\widehat{\boldsymbol{u}}(\mathbf{t}) \leq -\frac{c_2}{c_1}$. Condition (ii) of Theorem 5 follows immediately and so does Condition (iii) of Theorem 5, observing that $\mathbb{L}(\boldsymbol{\phi}, \mathbf{t})$ is continuous in $(\boldsymbol{\phi}, \mathbf{t})$ and strictly convex in $\boldsymbol{\phi}$. The convergence (6) follows as an application of Theorem 6. □

As an illustration of Theorem 2, let us determine the asymptotic distribution of the following test statistic,

$$S_n = \inf_{\mathbf{t} \in [0, L]} \left\| \mathbf{X}_n \widehat{\boldsymbol{\beta}}(\mathbf{t}) \right\|^2 ,$$

under the null hypothesis $H_0 : \boldsymbol{\beta} = 0$. Using Theorem 2, Assumption 1-(i) and the continuous mapping theorem, this limit distribution is given by the convergence

$$S_n = \inf_{\mathbf{t} \in [0, L]} \sqrt{n} \widehat{\boldsymbol{\beta}}(\mathbf{t})^T \frac{\mathbf{X}_n^T \mathbf{X}_n}{n} \sqrt{n} \widehat{\boldsymbol{\beta}}(\mathbf{t}) \rightsquigarrow \inf_{\mathbf{t} \in [0, L]} \widehat{\boldsymbol{u}}(\mathbf{t})^T C \widehat{\boldsymbol{u}}(\mathbf{t}) = S_\infty ,$$

where $\widehat{\boldsymbol{u}}(\mathbf{t})$ is the minimizer of

$$\mathbb{L}(\boldsymbol{\phi}, \mathbf{t}) = -2U^T\boldsymbol{\phi} + \boldsymbol{\phi}^T C\boldsymbol{\phi} + \mathbf{t}\sum_{j=1}^{p} |\phi_j| \ , \tag{52}$$

which is (5) under H_0.

In practice, since the regularization path is continuous piecewise linear, the statistic S_n can easily be computed by using the Least Angle Regression (LAR) algorithm (see [4]). Simulations of S_∞ under H_0 are obtained in the same way : one simulates $U \sim \mathcal{N}(0, \sigma^2 C)$ and compute the corresponding S_∞ by using the LAR algorithm to obtain the solution path minimizing the limit contrast (52). This allows to compute approximate asymptotic p-values of the statistic S_n.

To assess the performance of the test statistic S_n defined with $L = 1$, we compute ROC curves obtained on simulated data sets. We take $n = 30$ and $p = 20$ and simulate the linear model (2) under $H_0 : \boldsymbol{\beta} = 0$ and under H_1, in which case the components of $\boldsymbol{\beta}$ are drawn independently uniformly in $[-1, 1]$. The regression vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are drawn independently according to the Gaussian distribution $\mathcal{N}(0, I)$. We consider two different marginal distributions for the additive noise (ε_k):

1) a Gaussian distribution $\mathcal{N}(0, 4)$,
2) a mixture of two Gaussian distributions $\mathcal{N}(0, 0.8)$ and $\mathcal{N}(0, 7.2)$ with weights 0.5.

The ROC curves of S_n are compared to those of the F-statistic

$$F_n = \frac{(n-p)\left\|\mathbf{X}_n\widehat{\boldsymbol{\beta}}(0)\right\|^2}{p\left\|\mathbf{Y}_n - \mathbf{X}_n\widehat{\boldsymbol{\beta}}(0)\right\|^2} \ ,$$

where $\mathbf{Y}_n = [y_1 \ \ldots \ y_n]^T$, computed on the same data sets. The results in Figure 1 indicate that the performance of S_n is superior to that of F_n .

8. Other examples of contrast processes

In [14], a wide variety of models and functions g are shown to satisfy Conditions (P-1)–(P-4). These conditions apply for the general linear model (GLM) as this model satisfies the pointwise assumptions of [14, Section 4] (provided some moment conditions). They also apply for the least absolute deviation (LAD) criterion, see Example 8 in [14, Section 6] (provided again some moment conditions on the model). We briefly write the corresponding results in these two cases as examples of applications of Theorem 7. Uniform consistencies for both examples are obtained as applications of Theorem 4, since in these cases M_n is convex. As for the penalty, we consider the same ones as in [11]. They fit the conditions of Theorem 7 as they satisfy (43) and (44) by Lemma 1. Observe however that the function J_∞ in Lemma 1 depends on the chosen penalty and thus so does the limit $\widehat{\boldsymbol{u}}$ in (46).

ℓ^1–penalized GLM Consider a canonical exponential family of density

$$p(y|\theta) = h(y)\exp\{y\theta - b(\theta)\} \ ,$$

with respect to a dominating measure μ. The function b, sometimes called the log-repartition function, is given by

$$b(\theta) = \log \int h(y)\exp\{y\theta\}\mu(\mathrm{d}y) \ ,$$

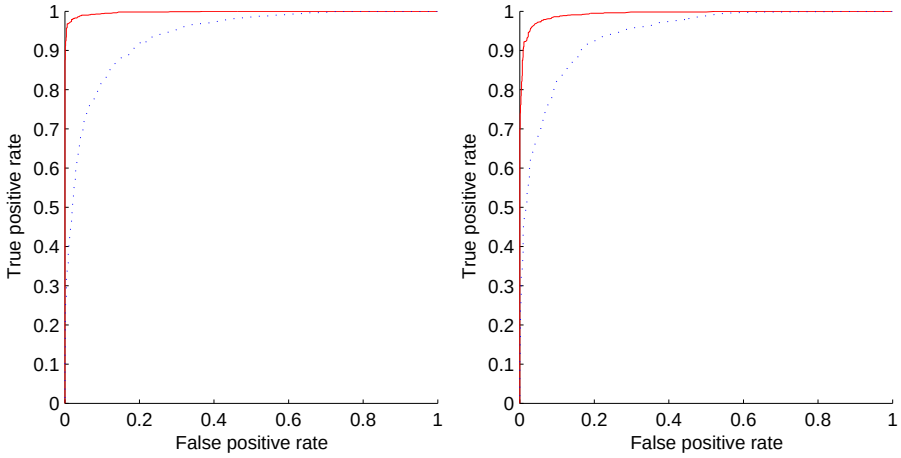

FIGURE 1. Roc curves of S_n (plain red line) and F_n (dotted blue line). Left: Gaussian noise. Right: Mixture noise. 1000 Monte-Carlo simulations have been used under H_0 and under H_1 to compute each ROC curve.

and thus is strictly convex and infinitely differentiable. In a GLM, one observes a sequence of i.i.d. $\mathbb{R} \times \mathbb{R}^p$-valued r.v.'s $(y_k, \mathbf{x}_k),\ k = 1, \ldots, n$, where y_k have conditional density $p(\cdot|\mathbf{x}_k^T\boldsymbol{\beta})$, given $\mathbf{x}_k$, with $\boldsymbol{\beta} \in \mathbb{R}^p$ denoting the parameter of interest. In this context, the non-penalized contrast process is given by the negated log-likelihood

$$M_n(\boldsymbol{\phi}) = n^{-1} \sum_{k=1}^{n} g((\mathbf{x}_k, y_k), \boldsymbol{\phi}) \ ,$$

where $g((\mathbf{x}, y), \boldsymbol{\phi}) = -y\mathbf{x}^T\boldsymbol{\phi} + b(\mathbf{x}^T\boldsymbol{\phi})$. Using that g is convex and smooth, and assuming some appropriate moment conditions on $\mathbf{x}_1$ for obtaining Pollard's conditions (P-1)–(P-4), we get the uniform consistency and a functional CLT on the regularization path $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ defined as the minimizer of (1) with $J_n(\boldsymbol{\phi}) = n^{-1/2} \sum_{i=1}^{p} |\phi_i|$ (this is the ℓ^1 penalty $J_n^{(1)}$ defined in (47)). In particular, for any $L > 0$,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \rightsquigarrow \widehat{\boldsymbol{u}} \ \text{ in } \ell^\infty([0, L], p) \ ,$$

where the limit $\widehat{\boldsymbol{u}}$ is defined as in the lasso case as the minimizer of (5) with $C = \mathbb{E}[b''(\mathbf{x}_1^T\boldsymbol{\beta})\mathbf{x}_1\mathbf{x}_1^T]$ (assumed positive-definite) and $U \sim \mathcal{N}(0, C)$. The numerical computation of $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ can be processed as proposed in [13].

ℓ^1 and ℓ^2–penalized LAD Given a sequence of $\mathbb{R}\times\mathbb{R}^p$-valued r.v.'s $(y_k, \mathbf{x}_k),\ k = 1, \ldots, n$, the LAD criterion is defined as

$$M_n(\boldsymbol{\phi}) = n^{-1} \sum_{k=1}^{n} |y_k - \mathbf{x}_k^T\boldsymbol{\phi}| \ .$$

It can be used to estimate the parameter $\boldsymbol{\beta} \in \mathbb{R}^p$ of a linear regression model $y_k = \mathbf{x}_k^T\boldsymbol{\beta} + \varepsilon_k$, with (ε_k) and $(\mathbf{x}_k)$ two independent sequence of i.i.d. r.v.'s. This contrast process is an alternative to the mean square criterion, resulting in an estimator less sensitive to the presence of outliers (for $\mathbf{x}_k = 1$, the minimizer of M_n is the sample median). In contrast

to the previous case, the contrast is not smooth, since the first derivative is discontinuous. However, as shown *e.g.* in [14], the minimizer of this contrast is asymptotically normal, provided some moment conditions and that

$$G(\boldsymbol{\phi}) = \mathbb{E}\left[\left|\varepsilon_1 + \mathbf{x}_1^T(\boldsymbol{\beta} - \boldsymbol{\phi})\right|\right]$$

has a non-singular second derivative at $\boldsymbol{\phi} = \boldsymbol{\beta}$. Observe that

$$G(\boldsymbol{\phi}) = \mathbb{E}\left[\mathbf{x}_1^T(\boldsymbol{\beta} - \boldsymbol{\phi}) + 2\int_0^{\mathbf{x}_1^T(\boldsymbol{\phi}-\boldsymbol{\beta})} F(s)\,\mathrm{d}s\right] ,$$

where F denotes the cumulative distribution function of ε_1. Thus, if ε_1 is distributed from a continuous density f, the second derivative of G at $\boldsymbol{\beta}$ is $\Gamma = 2f(0)\mathbb{E}\left[\mathbf{x}_1\mathbf{x}_1^T\right]$. Because the LAD criterion uses the ℓ^1 error function, the ℓ^2 penalty $J_n(\boldsymbol{\phi}) = n^{-1/2}\sum_{i=1}^p \phi_i^2$ could seem more reasonable. On the contrary Theorem 7 suggests that using an ℓ^1 error function contrast does not modify the asymptotic distribution of the regularization path, only the choice of the penalty does. In other words, the regularization path of the ℓ^1 and ℓ^2–penalized LAD has similar asymptotic distributions as the lasso and the ridge regression, respectively. Let us now precise the limit distribution of the regularization path $\widehat{\boldsymbol{\beta}}_n(\mathbf{t})$ defined as the minimizer of (1) with $J_n(\boldsymbol{\phi}) = n^{-1/2}\sum_{i=1}^p |\phi_i|$ and $J_n(\boldsymbol{\phi}) = n^{-1/2}\sum_{i=1}^p \phi_i^2$ respectively (these are the ℓ^1 and ℓ^2 penalty $J_n^{(1)}$ and $J_n^{(2)}$ defined in (47)). Under appropriate moment conditions on $(\varepsilon_1, \mathbf{x}_1)$ implying Pollard's conditions (P-1)–(P-4) (in particular $\mathbb{E}[\mathrm{sgn}(\varepsilon_1)] = 0$, $\mathbb{E}[\|\mathbf{x}_1\|^2] < \infty$ so that $\mathbb{E}[\Delta] = 0$, $\mathbb{E}[\|\Delta\|^2] < \infty$ and G is minimized at $\boldsymbol{\phi} = \boldsymbol{\beta}$), one has, for any $L > 0$,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \rightsquigarrow \widehat{\boldsymbol{u}} \text{ in } \ell^\infty([0, L], p) ,$$

where the limit $\widehat{\boldsymbol{u}}$ is defined as the minimizer of (45) where Γ is the (non-singular) second derivative of G at $\boldsymbol{\phi} = \boldsymbol{\beta}$, $W \sim \mathcal{N}(0, \mathbb{E}[\mathbf{x}_1\mathbf{x}_1^T])$ and J_∞ depends on the penalty. Namely, for the ℓ^1 penalty, one has $J_\infty = J_\infty^{(1)}$ and for the ℓ^2 penalty, one has $J_\infty = J_\infty^{(2)}$, where $J_\infty^{(\gamma)}$ is defined by (50).

REFERENCES

[1] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 2008. To appear.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[3] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194 (electronic), 2007. ISSN 1935-7524.

[4] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression. *Annals of Statistics*, 32:407–499, 2004.

[5] J.-F. Germain. A Two-steps Model Selection Procedure Based on the Regularization Path of a L_1-Penalized Logistic Likelihood. *Proceedings of SFdS*, June 2007.

[6] E. Greenshtein and Y. Ritov. Persistency in High Dimensional Linear Predictor-Selector and the Virtue of Over-Parametrization. *Bernoulli*, 10:971–988, 2004.

[7] S. J. Haberman. Concavity and estimation. *Ann. Statist.*, 17(4):1631–1661, 1989. ISSN 0090-5364.

[8] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(3):55–67, 1970.

[9] P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, pages 221–233. Univ. California Press, Berkeley, Calif., 1967.

[10] J. K. Kim and D. Pollard. Cube root asymptotics. *Ann. Statist.*, 18(1):191–219, 1990. ISSN 0090-5364.

[11] K. Knight and W. Fu. Asymptotics for LASSO-Type Estimators. *Annals of Statistics*, 28(5):1356–1378, 2000.

[12] W. Niemiro. Asymptotics for M-estimators defined by convex minimization. *Ann. Statist.*, 20(3):1514–1533, 1992. ISSN 0090-5364.

[13] M. Y. Park and T. Hastie. L_1-regularization path algorithm for generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(4):659–677, 2007. ISSN 1369-7412.

[14] D. Pollard. New Ways to Prove Central Limit Theorems. *Econometric Theory*, 1(3): 295–313, December 1985.

[15] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[16] R. Tibshirani. Regression Schrinkage and Selection via the LASSO. *J. Royal. Statist. Soc.*, B(58):229–243, 1996.

[17] A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[18] A. W. Van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3. With applications to statistics.

[19] P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7: 2541–2563, 2006. ISSN 1532-4435.

[20] H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *Ann. Statist.*, 35(5):2173–2192, 2007. ISSN 0090-5364.

RENAULT DREAM-DTAA, TECHNOCENTRE GUYANCOURT, 1, AVENUE DU GOLF, 78288 GUYANCOURT, FRANCE.

E-mail address: jean-francois.germain@renault.com

INSTITUT TELECOM, TELECOM PARISTECH, LTCI CNRS, 46, RUE BARRAULT, 75634 PARIS CEDEX 13, FRANCE

E-mail address: roueff@telecom-paristech.fr

Bibliographie

[1] N. Ansaldi. Contributions des méthodes statistiques à la quantification de l'agrément de conduite. 2002. Thèse de Doctorat.

[2] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous Analysis of Lasso and Dantzig Selector. 2008. Submitted to the Annals of Statistics.

[3] L. Birgé and P. Massart. Gaussian model selection. J. Eur. Math. Soc. (JEMS), 3(3) :203–268, 2001.

[4] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. Probab. Theory Related Fields, 138(1-2) :33–73, 2007.

[5] S. Boyd and L. Vandenberghe. Convex optimization. Cambridge University Press, Cambridge, 2004.

[6] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and regression trees. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984.

[7] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. Electron. J. Stat., 1 :169–194 (electronic), 2007.

[8] F. Cailliez and J.-P. Pagès. Introduction à l'analyse de données. SMASH, Paris, FR, 1976. 616 pp.

[9] E. J. Candès and T. Tao. Decoding by linear programming. IEEE Trans. Inform. Theory, 51(12) :4203–4215, 2005.

[10] E. J. Candès and T. Tao. The Dantzig selector : statistical estimation when p is much larger than n. Ann. Statist., 35(6) :2313–2351, 2007.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B, 39(1) :1–38, 1977. With discussion.

[12] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Inform. Theory, 52(1) :6–18, 2006.

[13] David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. Biometrika, 81(3) :425–455, 1994.

[14] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. Ann. Statist., 32(2) :407–499, 2004. With discussion, and a rejoinder by the authors.

[15] I. E. Frank and J. H. Friedman. A statiscal view of some chemometrics regression tools (with discussion). Technometrics, 35 :109–148, 1993.

[16] N. Freed and F. Glover. A Linear Programming Approach to Discriminant Problem. Decision Science, 12 :68–74, 1981.

[17] P. Geladi and B. Kowalski. Partial Least Square Regression : a Tutorial. Analytica Chemica Acta, 35 :1–17, 1986.

[18] J.-F. Germain. A Two-steps Model Selection Procedure Based on the Regularization Path of a L_1-Penalized Logistic Likelihood. Proceedings of SFdS, June 2007.

[19] C. J. Geyer. On the Asymptotics of Convex Stochastic Estimation. 1996. Unpublished manuscript.

[20] E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. Bernoulli, 10(6) :971–988, 2004.

[21] S. J. Haberman. Concavity and estimation. Ann. Statist., 17(4) :1631–1661, 1989.

[22] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning. Springer Series in Statistics. Springer-Verlag, New York, 2001. Data mining, inference, and prediction.

[23] N. L. Hjort and D. Pollard. Asymptotics for Minimisers of Convex Processes. 1993. Unpublished manuscript.

[24] A. E. Hoerl and R. W. Kennard. Ridge regression : Biased estimation for nonorthogonal problems. Technometrics, 12(3) :55–67, 1970.

[25] P. J. Huber. The behavior of maximum likelihood estimates under non-standard conditions. In Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I : Statistics, pages 221–233. Univ. California Press, Berkeley, Calif., 1967.

[26] E. A. Jaochimsthaler and A. Stam. Mathematical Programming Approaches for the Classification Problem in Two-Group Discriminant Analysis. Multivariate Behavioral Research, 25(4) :427–454, 1990.

[27] S. Keerthi and S. Shevade. A Fast Tracking Algorithm for Generalized Linear LARS/LASSO. IEEE transactions on Neural Networks, 2006. Submitted.

[28] J. K. Kim and D. Pollard. Cube root asymptotics. Ann. Statist., 18(1) :191–219, 1990.

[29] K. Knight and W. Fu. Asymptotics for lasso-type estimators. Ann. Statist., 28(5) :1356–1378, 2000.

[30] C. Leng, Y. Lin, and G. Wahba. A note on the lasso and related procedures in model selection. Statist. Sinica, 16(4) :1273–1284, 2006.

[31] R. D. Luce. Individual choice behavior : A theoretical analysis. John Wiley & Sons Inc., New York, 1959.

[32] C. L. Mallows. Some Comments on Cp. Technometrics, 15 :661–675, 1973.

[33] W. Niemiro. Asymptotics for M-estimators defined by convex minimization. Ann. Statist., 20(3) :1514–1533, 1992.

[34] M. R. Osborne, B. Presnell, and B. A. Turlach. Knot Selection for Regression Splines via the Lasso. Computing Science and Statistics, 30 :44–49, 1998.

[35] M. R. Osborne, B. Presnell, and B. A. Turlach. On the LASSO and its dual. J. Comput. Graph. Statist., 9(2) :319–337, 2000.

[36] M. Y. Park and T. Hastie. L_1-regularization path algorithm for generalized linear models. J. R. Stat. Soc. Ser. B Stat. Methodol., 69(4) :659–677, 2007.

[37] J.-M. Poggi and Tuleau C. Classification of objectivization data using CART and wavelets. Proceedings, June 2007.

[38] D. Pollard. New Ways to Prove Central Limit Theorems. Econometric Theory, 1(3) :295–313, December 1985.

[39] R. T. Rockafellar. Convex analysis. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[40] S. Rosset and J. Zhu. Piecewise linear regularized solution paths. Ann. Statist., 35(3) :1012–1030, 2007.

[41] M. Sauvé. Sélection de modèles en régression non gaussienne. Applications à la sélection de variables et aux tests de survie accélérés. 2006. Thèse de Doctorat.

[42] G. Schwarz. Estimating the dimension of a model. Ann. Statist., 6(2) :461–464, 1978.

[43] J. A. Swets and R. M. Pickett. Evaluation of Diagnostic Systems : Methods from Signal Detection Theory. Academic Press, New York, 1982. 253 pp.

[44] M. Tenenhaus. A PLS approach to multiple table analysis. In Classification, clustering, and data mining applications, Stud. Classification Data Anal. Knowledge Organ., pages 607–620. Springer, Berlin, 2004.

[45] M. Tenenhaus. La régression logistique PLS. In Modèles statistiques pour données qualitatives, pages 263–276. Technip, Paris, 2005.

[46] R. Tibshirani. Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B, 58(1) :267–288, 1996.

[47] K. E. Train. Discrete choice methods with simulation. Cambridge University Press, Cambridge, 2003.

[48] A. W. Van der Vaart. Asymptotic Statistics. Cambridge University Press, 1998.

[49] A. W. van der Vaart and J. A. Wellner. Weak convergence and empirical processes. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[50] V. Vapnik. Estimation of dependences based on empirical data. Springer Series in Statistics. Springer-Verlag, New York, 1982. Translated from the Russian by Samuel Kotz.

[51] H. Wold. Estimation of principal components and related models by iterative least squares. In Multivariate Analysis (Proc. Internat. Sympos., Dayton, Ohio, 1965), pages 391–420. Academic Press, New York, 1966.

[52] P. Zhao and B. Yu. On model selection consistency of Lasso. J. Mach. Learn. Res., 7 :2541–2563, 2006.

[53] H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. Ann. Statist., 35(5) :2173–2192, 2007.