Segmentation spatiale et sélection de modèle : théorie et applications statistiques

Antoine Chambaz

▶ **To cite this version:**

Antoine Chambaz. Segmentation spatiale et sélection de modèle : théorie et applications statistiques. Mathématiques [math]. Université Paris-Sud (1970-2019), 2003. Français. NNT : . tel-04144180

HAL Id: tel-04144180
https://theses.hal.science/tel-04144180

Submitted on 28 Jun 2023

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

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

63586

N° D'ORDRE : 7134

UNIVERSITÉ PARIS XI
UFR SCIENTIFIQUE D'ORSAY

THÈSE

présentée pour obtenir

le TITRE de DOCTEUR EN SCIENCES

DE L'UNIVERSITÉ PARIS XI – ORSAY

spécialité : Mathématiques

par

Antoine CHAMBAZ

Segmentation spatiale et sélection de modèle : théorie et applications statistiques.

Rapporteurs : M. Anestis ANTONIADIS
M. Eric MOULINES

Soutenue le 6 janvier 2003 devant la Commission d'examen composée de

M.	Raphaël CERF	(Président du jury)
Mme.	Elisabeth GASSIAT	(Directrice de thèse)
M.	Laurent GIRARD	(Examinateur)
M.	Marc LAVIELLE	(Co-Directeur de thèse)
M.	Christian LEONARD	(Examinateur)
M.	Eric MOULINES	(Rapporteur)

E quindi uscimmo a riveder le stelle. *Ce travail (ce voyage?) n'aurait pas abouti (la citation liminaire en clin d'œil est de Dante, plutôt que de Shakespeare ou Neruda) sans le soutien constant d'Elisabeth et Marc. Je vous remercie du fond du cœur pour votre enthousiasme, votre grande générosité, tant humaine que scientifique, et votre simplicité. Je vous dois une très belle aventure.*

Je souhaite exprimer ma grande gratitude à Laurent Girard de France Télécom R&D pour la confiance qu'il m'a témoignée. Je suis très heureux que tu aies accepté de faire partie de mon jury de thèse.

Je veux aussi remercier Anestis Antoniadis et Eric Moulines d'avoir bien voulu rapporter ce travail. Merci encore à Raphaël Cerf et Christian Léonard pour des conseils mathématiques éclairés et l'honneur que vous me faites, avec Eric, de participer au jury.

Je dois enfin à Pascal Massart et sa générosité jamais démentie l'éveil à la Statistique, de précieux enseignements et récentes suggestions. Et à Stéphane Boucheron, l'orientation vers le travail de Christian Léonard et Jamal Najim, ainsi que quelques « tournures shakespeariennes » du plus bel effet.

Je me suis beaucoup plu ces années au laboratoire de Mathématiques d'Orsay. Je ne saurais dresser une liste exhaustive des personnes que j'ai côtoyées ici avec bonheur. Je tiens cependant à citer tout particulièrement Dominique Hulin, Sophie Lemaire, Jean Coursol et les doctorants du bâtiment 430, dont Magalie Fromont-Renoir (ses gâteaux salvateurs), Servane Gey (son expertise CART), Estelle Kuhn-Sonnalier (sa bonne humeur), Khaled El-Dika (son intégrité), Alexandre Montaru (notre passion du jazz) et Vincent Sécherre (mon compagnon de rédaction tardive).

Je tiens par ailleurs à remercier vivement Anne Daviaud (qui a été de tous les déplacements) et Jean-Marc Kelif (de toutes les péripéties) de France Télécom R&D, ainsi que Zwi Altman, Nabil Benameur, Bénédicte Cherbonnel, Fabrice Clérot, Antoine Ferran, Arturo Ortega-Molina, Michel Ribeyron et Olivier Veyrunes ; sans oublier Marie-Hélène Busy et Arnaud Louis de Orange France, sans qui je n'aurais pu appliquer mon travail au problème original.

Je veux aussi redire toute mon amitié à ma « bande », Catherine Matias, Ivan Gentil et Vincent Lepez, pour tous les bons moments (et même les mauvais) passés et à venir ; à Vincent Beffara, pour ce mélange rare de gentillesse, de grande culture et de talent, et des conseils lumineux ; et au copain de la première heure, Gilles Blanchard, à qui je dois tant.

Merci enfin à Charles Lloyd (Lift every voice) et Wayne Shorter (Footprints !) pour leur accompagnement nocturne.

Finalement, pour l'essentiel, merci Julie, Lou & Fausto. Tout à coup je poussai un cri et courus sur le pont, C'est ça c'est ça, le bleu d'oultremer (Cendrars).

Je dédie ce travail à Martin et Clément, à notre enfance.

En route, le mieux c'est de se perdre. Lorsqu'on s'égare, les projets font place aux surprises et c'est alors, mais alors seulement, que le voyage commence.

Nicolas Bouvier, *in* Atlas.

Résumé :
Cette thèse trouve sa dynamique dans l'élaboration d'une méthode originale de raffinement de localisation du trafic de téléphonie mobile en zone urbaine pour France Télécom R&D, ainsi que dans l'étude de thèmes théoriques soulevés lors de notre exploration. Notre approche est de nature statistique. Il apparaît que les thèmes centraux de cette thèse sont la segmentation spatiale et la sélection de modèle.

Nous introduisons dans un premier temps les données sur lesquelles nous avons fondé notre approche du problème, que nous expliquons à leur lumière. Nous motivons le choix d'un modèle de régression hétéroscédastique.

Nous présentons ensuite une démarche non paramétrique par arbres de régression de type CART et ses extensions par ré-échantillonnage Bagging et Boosting dans un cadre de régression homoscédastique. Nous proposons une adaptation de ces techniques au cas hétéroscédastique. Une analyse originale de l'importance des variables y est associée. L'application commentée de notre méthode à divers jeux de données de trafic constitue notre réponse finale au problème initial.

Le travail appliqué évoqué plus tôt motive l'étude de la consistance d'une famille d'estimateurs de l'ordre et de la segmentation d'un modèle segmenté. Nous nous consacrons aussi, dans un cadre général de sélection de modèle dans un emboîtement, à l'estimation de l'ordre d'un modèle et aux propriétés de consistance, ainsi qu'aux vitesses de sur- ou sous-estimation. Une approche fonctionnelle, *i.e.* une approche pour laquelle les événements d'intérêt sont exprimés en termes d'événements sur la mesure empirique, permet d'unifier et de généraliser une large gamme de résultats antérieurs.

Les preuves font appel à une variété de techniques : arguments classiques de minimisation de contraste, concentration, inégalités maximales pour des variables dépendantes, lemme de Stein, pénalisation, Principes de Grandes et Moyennes Déviations pour la mesure empirique, *tour à la Huber*.

Abstract:
We tackle in this thesis the elaboration of an original method that provides refinement of the localization of the mobile telecommunication traffic in urban area for France Télécom R&D. This work involves both practical and theoretical developments. Our point of view is of statistical nature. The major themes are spatial segmentation and model selection.

We first introduce the various datasets from which our approach stems. They cast some light on the original problem. We motivate the choice of an heteroscedastic regression model.

We then present a practical nonparametric regression method based on CART regression trees and its Bagging and Boosting extensions by resampling. The latter classical methods are designed for homoscedastic models. We propose an adaptation to heteroscedastic ones, including an original analysis of variable importance. We apply the method to various traffic datasets. The final results are commented.

The above practical work motivates the theoretical study of the consistency of a family of estimators of the order of a segmented model and its associated segmentation. We also cope, in a general framework of model selection in a nested family of models, with the estimation of the order of a model. We are particularly concerned with consistency properties and rates of under- or overestimation. We tackle the problem at stake with a linear functional approach, *i.e.* an approach where the events of interest are described as events concerning the empirical measure. This allows to derive general results that gather and enhance earlier ones.

A large range of techniques are involved : classical arguments of M-estimation, concentration, maximal inequalities for dependent variables, Stein's lemma, penalization, Large and Moderate Deviations Principles for the empirical measure, *à la Huber trick*.

Keywords: abrupt changes, applications in engineering and industry, Bagging, Boosting, consistency, empirical measure, large deviations, linear functionals, M-estimation, maximal inequalities, mixing, model selection, nonparametric regression, order estimation, overestimation, penalization, mixture, rates of convergence, regression trees, resampling, segmentation, underestimation

MSC classification: 60E15, 60F10, 60G57, 62C99, 62F12, 62G08, 62G09, 62G20, 62M40, 62P30.

Table des matières

Ouverture **13**
Cheminement 15
Introduction à la sélection de modèle 18
Segmentation d'un champ aléatoire et sélection de modèle 25
Estimation de l'ordre d'un modèle 30
Pratique de la segmentation et agrégation 40
Eléments de raffinement de localisation de trafic 43

Prélude **45**
France Télécom R&D 47
Le laboratoire d'accueil 48
Un bref aperçu du réseau mobile Orange 49
Ce qu'il faut retenir 52
Introduction au problème initial 53
English summary 54

1. Les données France Télécom **57**
1.1. Introduction 59
1.2. Les données de trafic 59
1.3. Ebauche d'une étude statistique des durées d'appel 68
1.4. Les données de recouvrement cellulaire 70
1.5. Méthodes de localisation de trafic 77
1.6. Ce qu'il faut retenir 81
1.7. English summary 83

2. Interlude :
ILOTS15, Contoursllots et SIRENE **85**
2.1. Introduction 87
2.2. Les bases de données ILOTS15 et Contoursllots 87
2.3. Le répertoire SIRENE 89
2.4. Ce qu'il faut retenir 94
2.5. English summary 95
2.6. Annexe 96

3. Bagging and boosting CART regression trees: a user approach **101**
3.1. Introduction . . . 103
3.2. Prerequisites . . . 104
3.3. The CART regression algorithm . . . 112
3.4. Bagging and Boosting procedures . . . 121

4. Eléments de raffinement de localisation **129**
4.1. Introduction . . . 131
4.2. Tranche horaire matinale . . . 137
4.3. Tranche horaire de la mi-journée . . . 141
4.4. Tranche horaire de l'après-midi . . . 145
4.5. Tranche horaire de la soirée . . . 149
4.6. Journée complète . . . 153
4.7. Données HC2 . . . 155
4.8. Synthèse . . . 159
4.9. English summary . . . 163

5. Detecting abrupt changes in random fields **165**
5.1. Introduction . . . 167
5.2. The partitions and the associated parameters . . . 170
5.3. Modelization, observations, contrast . . . 173
5.4. Controlling random fluctuations *via* maximal inequalities . . . 176
5.5. The case of known cardinality of the true partition . . . 178
5.6. The case of unknown cardinality of the true partition . . . 183
5.7. Appendix . . . 184

6. Interlude :
A motivated introduction to Orlicz spaces and some LDP **193**
6.1. Orlicz spaces . . . 195
6.2. A Sanov's LDP with a view to statistical application . . . 197
6.3. Two MDP with a view to statistical application . . . 200
6.4. Appendix . . . 204

7. Estimating the order of a model **207**
7.1. Introduction . . . 209
7.2. Presentation of the model and three examples . . . 213
7.3. Two penalized maximum likelihood estimators . . . 218
7.4. Consistency . . . 219
7.5. Underestimation . . . 228
7.6. Overestimation . . . 243
7.7. Back to the three examples . . . 248
7.8. Appendix . . . 256

A. Glossaire **265**

B. Caccioppoli partitions **271**

C. Divers tableaux 277

Références 297

Ouverture

Résumé

Nous mettons en lumière l'unité de la thèse et la dynamique qui l'a portée. Nous éclairons ses contributions théoriques dans deux domaines connexes des thèmes de la *segmentation spatiale* et de la *sélection de modèle*. Finalement, nous évoquons les éléments de réponse qu'elle apporte au *problème appliqué original*. Une carte est offerte au lecteur, qui sera commentée tout au long de cette ouverture. On y représente le jeu des influences et des correspondances tissé entre les différentes partie de cette thèse.

Au menu

Cheminement **15**
Introduction à la sélection de modèle **18**
Deux cadres de travail 18
Estimation par minimum de contraste 19
Complexité d'un modèle et pénalisation 21
Deux approches non asymptotiques systématiques 23
Segmentation d'un champ aléatoire et sélection de modèle **25**
Motivations 25
Etat de l'art 26
Techniques mises en œuvre et résultats 27
Estimation de l'ordre d'un modèle **30**
Motivations 30
Etat de l'art 32
Techniques mises en œuvre et résultats 35
Pratique de la segmentation et agrégation **40**
Arbres de régression CART 40
Agrégation par Bagging ou Boosting 41
Eléments de raffinement de localisation de trafic **43**

Cette thèse trouve sa dynamique dans l'élaboration d'une méthode originale de tentative de résolution d'un problème auquel les ingénieurs de France Télécom R&D†* se sont trouvés confrontés et qu'ils nous ont soumis, ainsi que dans l'étude de thèmes théoriques soulevés lors de notre exploration. Notre travail de thèse nous apparaît, avec le recul, un peu comme un voyage. La carte présentée dans la Figure 1 donne un aperçu des routes que nous avons empruntées parmi tant d'autres et que nous exposons en détails dans le corps de cette thèse. Nous avons essayé de rendre compte de façon attrayante de notre cheminement, entre théorie et applications.

Cheminement

Notre approche est de nature statistique : nous avons analysé le problème original à la lumière des différents jeux d'observations que nous sommes parvenus à récolter et de nombreux échanges avec nos collègues de France Télécom R&D. Implicitement, ces observations sont considérées comme les réalisations de variables aléatoires sur un espace abstrait de probabilités. L'enjeu est d'extraire du jeu d'observations des informations relatives à la loi de ces variables aléatoires, c'est-à-dire sur la façon dont le hasard régit leur comportement.

Le problème original est résumé synthétiquement dans les lignes qui suivent. Nous devons tout d'abord introduire quelques rudiments de technologie de téléphonie mobile. Par téléphonie mobile, on entend technologie de téléphonie autorisant les déplacements en cours d'appel (c'est la mobilité†), un appel pouvant être initialisé de n'importe quel endroit (c'est l'itinérance†). Ces possibilités sont offertes par le système de téléphonie numérique et cellulaire que gèrent et

*Le lecteur pourra se reporter au Glossaire en Annexe A pour les mots signalés par un signe †.

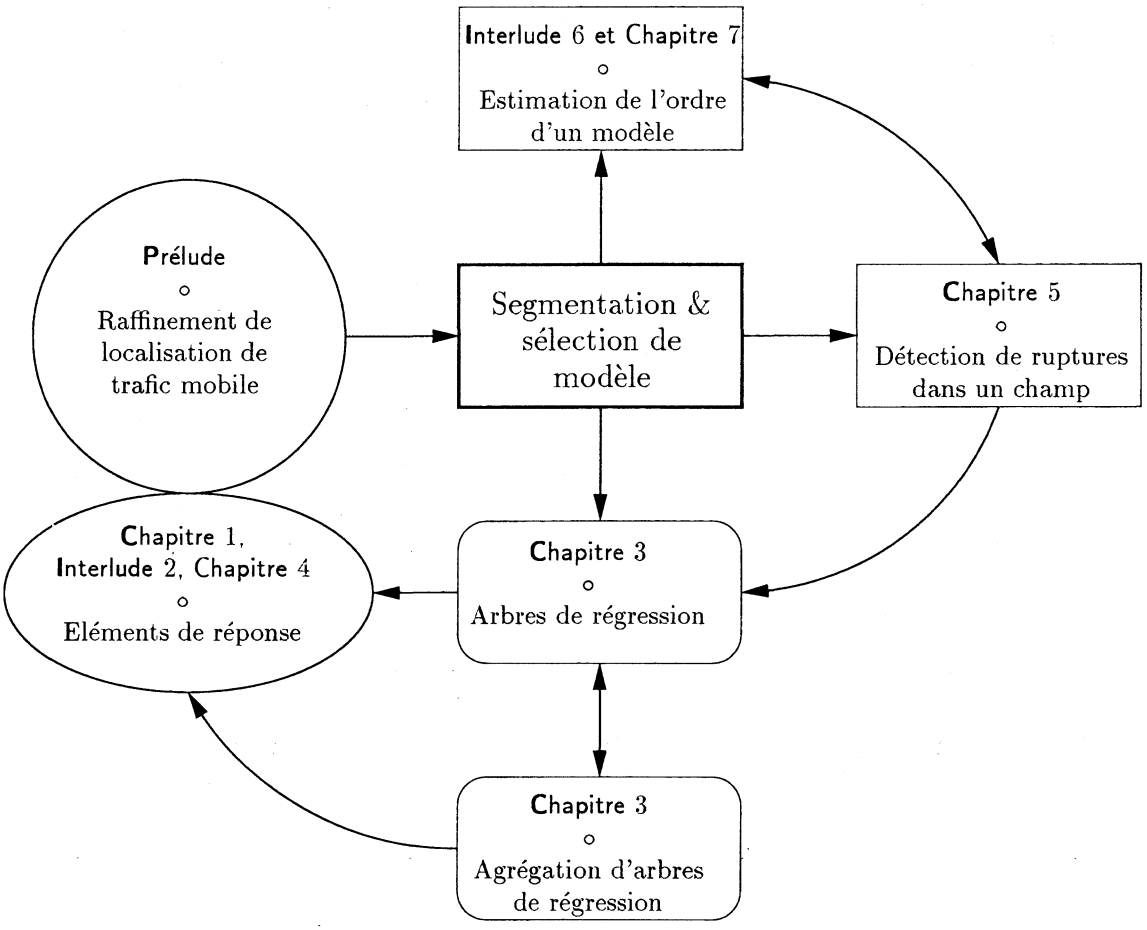


Figure 1 – Une carte à commenter. Le Prélude est consacré à la présentation de France Télécom R&D, l'entreprise qui est à l'origine du sujet de cette thèse, à quelques rudiments de technologie de téléphonie mobile et, finalement, à l'énoncé du problème original. Nous décrivons dans le Chapitre 1 et l'Interlude 2 les données sur lesquelles nous avons fondé notre approche du problème, que nous expliquons à leur lumière. Les Chapitres 3 et 4 sont dévolus à la mise en pratique de notre méthode. Celle-ci repose sur l'utilisation d'arbres de régression et sur leur agrégation. Il apparaît que les thèmes centraux de la thèse sont la segmentation spatiale et la sélection de modèle. Nous étudions soigneusement dans le Chapitre 5 (paru sous forme d'article dans ESAIM P&S) une méthode d'estimation dans un cadre commun de travail de segmentation spatiale et de sélection de modèle. L'Interlude 6 est dédié quant à lui à des préparatifs techniques pour le Chapitre 7, où nous étudions de façon approfondie certaines propriétés fines de consistance pour deux estimateurs de sélection de modèle. Les résultats généraux s'appliquent notamment dans un cadre de segmentation spatiale. Nous rassemblons enfin en Annexe un Glossaire, une présentation succincte des partitions de Caccioppoli et divers tableaux issus de l'exploitation de notre méthode.

développent notamment France Télécom R&D et Orange†. L'appellation cellulaire est motivée par l'architecture du réseau. Des émetteurs-récepteurs (les BTS†) desservent chacun une zone appelée cellule† : le principe schématique est qu'un appel en cours est géré par la BTS dont la cellule contient le point géographique physique d'appel depuis un terminal portable. Chaque BTS écoule ainsi un certain trafic que l'on peut mesurer. On peut en particulier en déduire une localisation du trafic à l'échelle des cellules du réseau, et par exemple identifier les cellules à trafic moyen faible, moyen, fort.

Le problème original de France Télélcom R&D sous sa formulation initiale consistait en la proposition d'une méthode de raffinement de localisation du trafic téléphonique mobile en zone urbaine.

L'expression de raffinement de localisation évoque une localisation du trafic qui ne soit pas limitée aux zones pour lesquelles on dispose naturellement d'observations, *i.e.* qui s'étende aux zones qui ne sont éventuellement pas des cellules. Nous pouvons avancer trois motivations qui éclairent l'intérêt d'un tel problème :

- l'accès à une connaissance qualitative et quantitative plus précise du trafic local qui ne soit pas fondée exclusivement sur les appréciations techniques et la connaissance intime du réseau local qu'en a un ingénieur-exploitant ;
- l'aide à la densification d'un réseau existant, *i.e.* au positionnement et paramétrage de nouveaux émetteurs-récepteurs ;
- l'aide à la planification d'un nouveau réseau.

Il est apparu à propos du premier des points évoqués ci-dessus que nos correspondants à France Télécom R&D et Orange souhaitaient que les résultats de la méthode de raffinement de localisation soient aisément lisibles. Typiquement, une *description discrète* du trafic les satisfaisait. Nous entendons par là un modèle statistique pour les quantités de trafic constitué d'un certain nombre de types de trafic (ce nombre est appelé ordre du modèle, il est à estimer), et d'une caractérisation de chacun de ces types *via* un jeu de paramètres (à estimer encore), par exemple la moyenne du trafic pour chacun des types.

Comme par ailleurs le trafic téléphonique mobile en zone urbaine dense (pensez à Paris, ville sur laquelle nous avons travaillé) dépend naturellement fortement de la nature socio-démographique et culturelle des lieux, nous nous sommes orientés vers une modélisation très générale du trafic $Y(\zeta)$ supporté par une zone ζ, vu comme variable à expliquer dans un modèle de régression dont la variable explicative $X(\zeta)$ est un vecteur dont les coordonnées sont de nature socio-démographique et culturelle :

$$Y(\zeta) = f_{\mathrm{m}}^{\star}\Big(X(\zeta)\Big) + f_{\mathrm{se}}^{\star}\Big(X(\zeta)\Big)\, e.$$

Ici, $f_{\mathrm{m}}^{\star}$ et $f_{\mathrm{se}}^{\star}$ sont deux fonctions de régression de moyenne et écart-type respectivement, et e est une variation aléatoire (un bruit) centrée et réduite conditionnellement à $X(\zeta)$. Le modèle est hétéroscédastique, c'est-à-dire que la variance du bruit n'est pas supposée constante. Un modèle discret au sens précisé plus haut est obtenu pour une fonction $(f_{\mathrm{m}}^{\star}, f_{\mathrm{se}}^{\star})$ constante par morceaux sur les parties d'une partition $\tau = (\tau_k)_{1 \leq k \leq K}$ de l'espace $\mathcal{X}$ auquel appartiennent les variables explicatives $X(\zeta)$. Dans ce cas, les choix du nombre de morceaux K, d'une bonne partition et des valeurs associées relèvent simultanément de la *segmentation spatiale* et de la *sélection de modèle* : ce sont les deux thèmes centraux de cette thèse.

Nous les avons explorés sous trois angles différents.

- Dans le Chapitre 3, nous nous intéressons à l'aspect pratique, *i.e.* à une façon de construire des modèles discrets à partir de nos observations. Nous y mettons en scène les arbres de régression CART (sous leur forme originale, avec aussi quelques modifications que nous leur apportons afin de les adapter selon nos intentions). Les arbres de régression CART offrent une solution efficace et largement appliquée dans des domaines variés. Ils sont d'ailleurs présents dans de nombreux logiciels statistiques.
  Ce chapitre est aussi consacré à une présentation de deux méthodes très en vogue d'agrégation de tels arbres, le Bagging et le Boosting. Elles sont dues à l'origine aux membres de la communauté de la théorie de l'intelligence artificielle (*machine learning community*). Elles permettent, au prix de la perte de lisibilité d'un arbre de régression, d'améliorer les performances statistiques d'estimation.
  Le Chapitre 4 est dédié aux résultats de l'application de notre méthode au problème original.
- Dans le Chapitre 5 (paru sous forme d'article en 2002 dans la revue ESAIM P&S), nous étudions soigneusement certaines propriétés statistiques asymptotiques d'un estimateur dans un cadre de segmentation spatiale et de sélection de modèle qui inclut notamment les modèles discrets évoqués au paragraphe précédent.
  Nous avons recours à diverses techniques, parmi lesquelles des arguments classiques de minimisation de contraste (*M-estimation* en anglais), de pénalisation, des inégalités de concentration et des inégalités maximales pour des variables aléatoires dépendantes.
  Les arbres de régression présentés dans le Chapitre 3 et exploités dans le Chapitre 4 apparaissent comme une mise en pratique de la théorie étudiée dans ce chapitre.
- Dans le Chapitre 7, l'accent est mis sur la sélection de modèle, puisqu'il s'agit d'estimer de quel modèle sont issues les observations parmi toute une famille de modèles potentiels. Les résultats, très généraux, s'appliquent entre autres à la segmentation spatiale. L'aspect technique de ce chapitre est en partie préparé dans l'Interlude 6 qui le précède.
  Les preuves reposent sur une approche fonctionnelle d'estimation pénalisée et l'application de Principes de Grandes et Moyennes Déviations pour la mesure empirique, sur une Loi du Logarithme Itéré ainsi que sur le lemme de Stein. Un exemple de « tour à la Huber » permet de raffiner l'ensemble des résultats en termes de contraintes sur la fonction de pénalité.

Introduction à la sélection de modèle

Deux cadres de travail

Modèles de densités

Supposons que l'on observe n variables aléatoires (v.a.) $Z_1, \ldots, Z_n$ à valeurs dans $\mathcal{Z}$, tirées sur un espace probabilisé $(\Omega, \mathcal{A}, \mathrm{P})$. Nous supposons qu'elles sont indépendantes et identiquement distribuées suivant une loi $P^\star$ de densité $f^\star$ relativement à une mesure μ.

L'objectif du statisticien est d'estimer la densité $f^\star$ sous la forme d'un estimateur $\widehat{f}$, *i.e.* d'une fonction mesurable des observations $\widehat{f}$. Nous désignons par $\mathcal{F}$ un espace d'estimateurs potentiels qui contient $f^\star$.

Nous abrégerons parfois par la suite l'expression *modèles de densités* en **DS**.

Modèles de régression

Supposons que l'on observe n v.a. $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$ à valeurs dans $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$, tirées sur un espace probabilisé $(\Omega, \mathcal{A}, \mathrm{P})$. Ajoutons que celles-ci sont identiquement distribuées suivant une loi jointe $P^\star$, puisqu'issues d'un même modèle hétéroscédastique de régression selon

$$Y_i = f^\star_{\mathrm{m}}(X_i) + f^\star_{\mathrm{se}}(X_i)\, e_i \quad (1 \leq i \leq n),$$

pour des bruits centrés réduits e_i conditionnellement à X_i. Nous notons P la loi marginale commune des X_i.

Remarque 1.

• Le choix d'un cadre de travail hétéroscédastique (*i.e.* où la variance du bruit n'est pas supposée constante, par opposition au cadre homoscédastique) est motivé par la nature de nos observations de trafic de téléphonie mobile. Dans la suite, nous abrégerons parfois les expressions *modèles de régression homoscédastique* et *modèles de régression hétéroscédastique* en **HO** et **HE**, respectivement.

• Il est à noter que l'on ne suppose pas pour l'instant que les v.a. sont indépendantes : c'est de nouveau parce que les données réelles avec lesquelles nous avons travaillé ne satisfont pas cette hypothèse. Le travail du Chapitre 5 permet d'ailleurs de traiter le cas des variables dépendantes, contrairement au Chapitre 7 où l'hypothèse d'indépendance est invoquée.

Notre tâche de statisticien est d'estimer la fonction de régression $f^\star = (f^\star_{\mathrm{m}}, f^\star_{\mathrm{se}})$ sous la forme d'un prédicteur $\widehat{f} = (\widehat{f}_{\mathrm{m}}, \widehat{f}_{\mathrm{se}})$, *i.e.* une fonction $\widehat{f}$ mesurable des observations. Nous désignons par $\mathcal{F}$ un espace de prédicteurs potentiels qui contient $f^\star$. Les éléments de $\mathcal{F}$ sont notés f et se décomposent suivant $f = (f_{\mathrm{m}}, f_{\mathrm{se}})$, hormis lorsque $f^\star_{\mathrm{se}} = \sigma$ est une constante : dans ce cas, nous notons $f = f_{\mathrm{m}}$ par souci de concision. De la même façon, $f^\star$ et $\widehat{f}$ coïncident avec $f^\star_{\mathrm{m}}$ et $\widehat{f}_{\mathrm{m}}$ quand $f^\star_{\mathrm{se}} = \sigma$ est constante. Alors le modèle de régression est homoscédastique et l'on a

$$Y_i = f^\star(X_i) + e'_i \quad (1 \leq i \leq n),$$

où e'_i est un bruit centré et de variance σ^2 conditionnellement à X_i.

Estimation par minimum de contraste

Une façon naturelle de contrôler la qualité de l'estimation de $f^\star$ par $\widehat{f}$ consiste à essayer de rendre *l'erreur moyenne* de $\widehat{f}$ aussi petite que possible. La vocation des *fonctions de perte* est de fournir de telles quantités de contrôle.

Une fonction de perte L est une fonction de $\mathcal{F}$ dans $\mathbb{R}$ qui atteint son minimum au seul point $f^\star$. Ainsi, on peut décider de juger de la qualité de l'estimation par $\widehat{f}$ à la vue de la quantité $L(\widehat{f})$. Un tel critère incite à construire $\widehat{f}$ par tentative de minimisation de L. Cependant, une fonction de perte dépend souvent de la loi inconnue $P^\star$, comme par exemple avec les *fonctions de perte par contraste*, qui sont de la forme

$$L_\gamma(f) = \mathrm{E}_{P^\star}\gamma(f, Z)$$

où γ est une fonction à valeur réelle appelée *contraste* (nous en donnons trois exemples ci-dessous) et $\mathrm{E}_{P^\star}$ représente l'espérance relative à la v.a. Z de loi $P^\star$.

A une fonction de perte par contraste, on associe naturellement la *perte relative*

$$RL_\gamma(f) = \mathrm{E}_{P^\star}\gamma(f, Z) - \mathrm{E}_{P^\star}\gamma(f^\star, Z)$$

qui, bien que coïncidant avec la perte originale à une constante additive (inconnue) près, jouit de propriétés statistiques distinctes.

Trois fonctions de contraste jouent un rôle prépondérant dans cette thèse : si l'on note z un élément générique de $\mathcal{Z}$ (avec la décomposition naturelle $z = (x, y)$ dans les cadres de régression), leurs définitions prennent les formes suivantes.

DS : on suppose que $\mathcal{F}$ est l'ensemble des densités sur $\mathcal{Z}$ relativement à μ. Le contraste

$$\gamma_0(f, z) = -\log f(z)$$

se voit associer pour perte relative la *divergence de Kullback-Leibler* (nous parlerons de *divergence*)

$$RL_{\gamma_0}(f) = RL_0(f) = H(P^\star \,|\, f\,\mu),$$

où la divergence $H(Q_1 \,|\, Q_2)$ de deux probabilités Q_1 et Q_2 est définie* en toute généralité par

$$H(Q_1 \,|\, Q_2) = Q_1 \log \frac{\mathrm{d}Q_1}{\mathrm{d}Q_2}$$

si $Q_1 << Q_2$, et $H(Q_1 \,|\, Q_2) = \infty$ dans le cas contraire. La divergence est notamment positive, nulle si et seulement si (ssi) les deux probabilités Q_1 et Q_2 coïncident.

HO : on suppose que $f^\star = f^\star_{\mathrm{m}} \in \mathcal{F} = L^2(P)$. Le contraste quadratique

$$\gamma_1(f, z) = \Big(y - f(x)\Big)^2$$

produit la classique *perte quadratique relative*

$$RL_{\gamma_1}(f) = RL_1(f) = P(f - f^\star)^2 \geq 0,$$

avec égalité ssi $f = f^\star$ P-presque sûrement (P-ps).

HE : on suppose que $f^\star = (f^\star_{\mathrm{m}}, f^\star_{\mathrm{se}}) \in \mathcal{F} = \mathcal{F}_{\mathrm{m}} \times \mathcal{F}_{\mathrm{se}}$ tels que $\mathcal{F}_{\mathrm{m}} \subset L^2(P)$ et que les éléments de $\mathcal{F}_{\mathrm{se}}$ sont uniformément éloignés de 0. Le contraste suivant (dont la définition est suggérée par la log-vraisemblance d'une v.a. gaussienne)

$$\gamma_2(f, z) = \frac{\Big(y - f_{\mathrm{m}}(x)\Big)^2}{{f_{\mathrm{se}}}^2(x)} + \log {f_{\mathrm{se}}}^2(x)$$

donne naissance à la perte relative moins commune

$$RL_{\gamma_2}(f) = RL_2(f) = P\left(\frac{f_{\mathrm{m}} - f^\star_{\mathrm{m}}}{f_{\mathrm{se}}}\right)^2 + P\left(\frac{{f^\star_{\mathrm{se}}}^2}{{f_{\mathrm{se}}}^2} + \log\frac{{f_{\mathrm{se}}}^2}{{f^\star_{\mathrm{se}}}^2} - 1\right) \geq 0,$$

où l'égalité est satisfaite ssi $f = f^\star$ P-ps, *i.e.* lorsque $f_{\mathrm{m}} = f^\star_{\mathrm{m}}$ et $f_{\mathrm{se}} = f^\star_{\mathrm{se}}$ P-ps.

*Par la suite, nous utiliserons les notations d'analyse fonctionnelle pour les espérances et les intégrales, *i.e.* nous écrirons $\mu\, f$ pour l'intégrale $\int f \mathrm{d}\mu$ d'une fonction f relativement à une mesure μ.

L'heuristique de l'estimation par minimum de contraste offre une réponse pratique à l'impossibilité (due à la dépendance des pertes relativement à $P^\star$) d'exploiter les fonctions de perte comme outils d'estimation. Ainsi, informellement, la minimisation du critère empirique de substitution

$$f \mapsto n^{-1} \sum_{i=1}^{n} \gamma(f, Z_i) \tag{1}$$

plutôt que de son espérance* définit une méthode pratique d'estimation. Celle-ci coïncide en particulier avec l'estimation par maximum de vraisemblance pour γ_0 dans le modèle **DS** et avec les moindres carrés quand la fonction de contraste est γ_1 dans le modèle **HO**.

Les deux exemples précédents sont parmi les plus célèbres des procédures de minimisation de contraste (*M-estimation procedures* en anglais). Les premiers travaux les concernant remontent aux années 20 et aux études de Fisher. Nous renvoyons à (van der Vaart 1998) pour une présentation lumineuse récente des propriétés asymptotiques des estimateurs de minimum de contraste.

Complexité d'un modèle et pénalisation

La décomposition biais/variance

L'un des enjeux majeurs lorsque l'on procède à une estimation par minimisation de contraste empirique est de se placer sur un bon espace de minimisation que nous appellerons *espace de référence*. En effet, l'ensemble $\mathcal{F}$ proposé dans chacun des exemples **DS**, **HO** et **HE** peut être qualifié *d'universel**, dans la mesure où il contient quasiment toutes les fonctions susceptibles sans aucun *a priori* d'être choisies. Or, il est connu que dans de telles circonstances, l'estimation peut s'avérer inconsistante, ou bien sous-optimale. Finalement, la difficulté consiste à choisir un modèle de référence dont la richesse assure un comportement satisfaisant à l'estimateur construit par minimisation sur lui, en veillant toutefois à ce que le modèle en question ne soit pas disproportionné.

On entrevoit donc l'intérêt de l'introduction d'une *famille de modèles* dont l'éventail en offre une variété de *complexités* variées, *i.e.* caractérisés par des degrés de sophistication divers. Schématiquement, le principe est d'en extraire, grâce à des arguments statistiques, un meilleur modèle de référence et l'estimée $\widehat{f}$ de $f^\star$ associée par minimisation de contraste.

Illustrons notre propos par un exemple simple. Nous considérons le modèle **HO** avec indépendance des observations $Z_1, \ldots, Z_n$. Soit une suite $\{\mathcal{F}_K\}_{K \geq 1}$ de sous-espaces affines emboîtés de $\mathcal{F} = L^2(\mathbb{P}_n)$ muni de la structure héritée de la mesure empirique

$$\mathbb{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$$

(nous notons $\|\cdot\|_2$ la norme correspondante). L'indice K correspond à la dimension du sous-espace $\mathcal{F}_K$. Soit $\widehat{f}_K$ l'estimateur de minimum de constraste calculé sur $\mathcal{F}_K$ et $\bar{f}_K$ la projection

* *i.e.* la minimisation de la fonction de perte originale $f \mapsto L_\gamma(f)$, ou encore de $f \mapsto RL_\gamma(f)$

* L'expression *exhaustif* conviendrait sans doute mieux si elle n'avait par ailleurs un sens précis en Statistique.

au sens de la norme $\|\cdot\|_2$ de $f^\star$ sur $\mathcal{F}_K$. Ces fonctions satisfont

$$\|f^\star - \widehat{f}_K\|_2^2 = \underbrace{\inf\left\{\|f^\star - f\|_2^2 : f \in \mathcal{F}_K\right\}}_{\|f^\star - \bar{f}_K\|_2^2} + \|\bar{f}_K - \widehat{f}_K\|_2^2$$

et le passage à l'espérance relative aux observations aboutit (après de menus calculs) à l'expression

$$\mathrm{E}\|f^\star - \widehat{f}_K\|_2^2 = \mathrm{E}\left[\inf\left\{\|f^\star - f\|_2^2 : f \in \mathcal{F}_K\right\}\right] + \sigma^2 \frac{K}{n}.$$

Le membre de gauche donne une distance moyenne entre la fonction cible $f^\star$ et l'estimée $\widehat{f}_K$ déterminée sur $\mathcal{F}_K$. Elle peut être interprétée comme une *erreur typique*, se décomposant ainsi naturellement en la somme de deux termes :

- la moyenne (relativement au tirage de $X_1, \ldots, X_n$) de la *distance* de $f^\star$ à l'ensemble $\mathcal{F}_K$ — terme que l'on qualifie souvent de *biais* ;
- un terme positif proportionnel à la variance σ^2 du bruit, à la dimension K du modèle et à l'inverse n^{-1} du nombre d'observations, qui apparaît comme un *défaut de performance* en tant que différence entre l'erreur typique et la distance moyenne — terme que l'on qualifie souvent de *variance*.

Il est important de noter que ces deux termes ont des comportements opposés lorsque K croît (*i.e.* lorsque le modèle se complexifie), puisqu'en effet alors :

- le terme de biais décroît ;
- le terme de variance croît.

Heuristiquement donc, le raffinement du modèle (*i.e.* sa complexification) au sein duquel on cherche un estimateur tend à assurer une réduction du terme de biais, au prix d'une augmentation du terme de variance. De façon imagée, plus le modèle est riche (*i.e.* complexe) plus il est facile de s'ajuster aux données d'une part, et plus il y a potentiellement d'erreur d'estimation et de risque de sur-ajustement de l'autre. Cette interprétation justifie elle aussi les expressions de biais et variance.

En vertu de l'exemple simple exposé ci-dessus, on qualifiera dans une procédure statistique de minimisation de contraste de *terme de biais* une quantité qui diminue quand la complexité du modèle de référence croît et de *terme de variance* une quantité qui au contraire croît avec la complexité de ce modèle. La somme de deux telles quantités constitue une mesure de qualité sensible aux effets antagonistes de l'augmentation de complexité du modèle de référence.

Le terme de pénalisation

En particulier, le critère empirique de contraste donné par (1) définit un terme de biais selon la terminologie introduite dans le paragraphe précédent. L'absence de terme de variance dans la quantité à minimiser lors de la procédure de minimisation de contraste empirique (*i.e.* la minimisation du seul terme de biais) favorise les modèles complexes au détriment des modèles simples et conduit certainement à un estimateur peu satisfaisant.

Cette remarque est à l'origine de l'introduction d'un terme supplémentaire de variance sous la forme d'une quantité positive appelée *pénalité*, dépendant du nombre n d'observations et de

l'ordre K du modèle (qui est une mesure de complexité), que nous notons $\text{pen}(n, K)$. On peut l'interpréter comme un *terme simulé de variance.*

On décide de substituer au seul terme de biais, la somme du terme de biais et de la pénalité, d'où le *contraste empirique pénalisé*

$$f \mapsto n^{-1} \sum_{i=1}^{n} \gamma(f, Z_i) + \text{pen}(n, K) \quad (\text{si } f \in \mathcal{F}_K) \tag{2}$$

dont la minimisation sur la collection $\{\mathcal{F}_K\}_{K \geq 1}$ de modèles indexés par leur ordre K définit une procédure d'estimation pénalisée qui nous intéressera au premier chef dans toute cette thèse.

Le principe fondamental est qu'un bon *réglage* du terme $\text{pen}(n, K)$ de pénalité peut équilibrer avantageusement les effets antagonistes de l'augmentation de l'ordre K du modèle de référence dans la procédure de minimisation de contraste pénalisé et produire un meilleur estimateur final *via* le choix d'un meilleur modèle de référence.

Historiquement, quatre articles fondateurs marquent l'entrée en scène de l'artifice de pénalisation sous des formes diverses : Mallows (1973), Akaike (1974), Rissanen (1978) et Schwarz (1978) définissent respectivement les critères C_p, AIC (Akaike Information Criterion), MDL (Minimum Description Length) et BIC (Bayesian Information Criterion). Leurs définitions sont motivées par des arguments asymptotiques et des approches bayésienne ou du domaine de la théorie de l'information. Ils s'appliquent à divers critères empiriques, dont les deux classiques du maximum de vraisemblance et des moindres carrés évoqués plus haut.

Deux approches non asymptotiques systématiques*

Un question théorique cruciale tient en la calibration du terme de pénalité en vue d'obtenir un contrôle aussi fin que possible des performances de la procédure d'estimation. Deux approches systématiques non asymptotiques de cette question ont été explorées par les pionniers Červonenkis et Vapnik à partir des années 70 et, dans un cadre très général, par Barron, Birgé et Massart au cours des années 90 (nous donnons des références précises ci-dessous).

Nous proposons ci-dessous une présentation vulgarisée de ces deux approches.

Minimisation structurelle du risque

La théorie de Vapnik et Červonenkis (voir par exemple Vapnik 1998) consiste à contrôler l'écart entre la perte empirique et la perte réelle des estimateurs $\widehat{f}_K$ obtenus par minimisation du critère empirique non pénalisé sur le modèle de référence $\mathcal{F}_K$. A cette fin, un contrôle *uniforme* de quantités du type

$$V_K = \sup_{f \in \mathcal{F}_K} \left| n^{-1} \sum_{i=1}^{n} \gamma(f, Z_i) - RL_\gamma(f) \right|$$

est requis. Nous observons que V_K s'exprime comme le supremum de valeurs absolues de différences entre des pertes empiriques et leurs espérances. Le contrôle de V_K suppose l'utilisation

*Cette section doit beaucoup au beau chapitre introductif de la thèse de Gilles Blanchard (2001).

d'inégalités de concentration.* Intuitivement, la complexité du modèle $\mathcal{F}_K$ est un élément fondamental du résultat de contrôle.

Ainsi, modèle par modèle, on peut obtenir une *perte maximale* garantie avec grande probabilité du type

$$RL_\gamma(f) \leq n^{-1} \sum_{i=1}^{n} \gamma(f, Z_i) + \text{pen}(n, K) \quad \text{(pour tout } f \in \mathcal{F}_K) \tag{3}$$

en fonction d'une expression qui se décompose sous la forme biais/variance comme la somme de la perte empirique et d'un terme de pénalité $\text{pen}(n, K)$ dépendant du modèle.

La méthode de *minimisation structurelle du risque* consiste à choisir comme estimateur $\widehat{f}$ le minimiseur du terme empirique sur le modèle de référence $\mathcal{F}_{\widehat{K}}$ tels que la somme

$$n^{-1} \sum_{i=1}^{n} \gamma(\widehat{f}, Z_i) + \text{pen}(n, \widehat{K})$$

minimise l'expression de droite de l'équation (3) en $K \leq K_{\max}$ et $f \in \mathcal{F}_K$.

Finalement, on obtient aisément une majoration du *risque* de $\widehat{f}$, *i.e.* de l'espérance $\text{E}[RL_\gamma(\widehat{f})]$ (relativement aux observations $Z_1, \ldots, Z_n$) qui est l'ultime quantification de la performance de la procédure d'estimation.

Sélection par estimation pénalisée

La sélection de modèle par estimation pénalisée a été récemment abordée plus finement qu'à la façon de Vapnik et Červonenkis, grâce notamment à de nouvelles inégalités de concentration dues en particulier à Talagrand et Ledoux. Pour une présentation exhaustive de cette procédure, le lecteur pourra se référer à (Barron et al. 1999) – (Massart 2000) est toutefois plus abordable. Cet article en offre une vision synthétique qui met en lumière le rôle des inégalités de concentration d'une part et une comparaison avec la méthode de minimisation empirique du risque de Vapnik et Červonenkis de l'autre.

Cette fois, ce sont les expressions du type

$$V_{K'}(f) = \sup_{f' \in \mathcal{F}_{K'}} \left| \frac{n^{-1} \sum_{i=1}^{n} \Big(\gamma(f', Z_i) - \gamma(f, Z_i)\Big) - \Big(L_\gamma(f') - L_\gamma(f)\Big)}{w_{K'}(f')} \right|$$

(pour tout $f \in \mathcal{F}_K$ et tout K, K') (4)

auxquelles on s'intéresse particulièrement. Notez que le numérateur de la fraction met en jeu des différences de pertes empiriques recentrées calculées sur deux modèles distincts. On trouve par ailleurs au dénominateur un facteur de pondération $w_{K'}(f')$ qui dépend de $f^\star$, K' et f'.

*Ce sont les outils mathématiques qui rendent rigoureux l'énoncé suivant de Talagrand (1996b) : « une variable aléatoire qui dépend (« régulièrement ») de l'influence de nombreuses variables indépendantes (sans toutefois dépendre particulièrement de l'une d'elles) est essentiellement constante. » L'expression « essentiellement constante » signifie que les v.a. sont proches de leur espérance avec grande probabilité.

Cette fois, le principe très schématique est de majorer l'espérance des $V_{K'}(f)$ (c'est ici qu'intervient la complexité du modèle $\mathcal{F}_{K'}$) puis de contrôler les écarts des $V_{K'}(f)$ à leur espérance par concentration. Si ces écarts sont sommables, on peut passer à un contrôle des déviations uniforme en $\{\mathcal{F}_{K'}\}_{K' \geq 1}$ pour tout $f \in \mathcal{F}_K$ et tout $K \geq 1$. Un argument final permet de basculer vers un contrôle du *risque* de notre estimateur $\widehat{f} \in \mathcal{F}_{\widehat{K}}$, qui est tel que l'expression

$$n^{-1} \sum_{i=1}^{n} \gamma(f, Z_i) + \mathrm{pen}(n, K)$$

soit minimisée pour le choix $K = \widehat{K}$ et $f = \widehat{f} \in \mathcal{F}_K$.

En guise de conclusion, le contrôle obtenu de cette façon est plus satisfaisant que le précédent, notamment parce qu'il permet un choix de fonction de pénalité à valeurs plus faibles et que ceci contribue à améliorer le contrôle du risque.

Nous renversons la symétrie de l'organisation de cette thèse en commençant par résumer les contenus théoriques des Chapitre 5, Interlude 6 et Chapitre 7.

Le résumé du Chapitre 3 suit, où l'on s'attache à la préparation de la mise en œuvre pratique des méthodes étudiées du point de vue théorique. Puis celui des Prélude, Chapitre 1, Interlude 2 et Chapitre 4, consacrés à part entière aux applications qui nous motivent.

Segmentation d'un champ aléatoire et sélection de modèle

Nous résumons dans la présente section le contenu du Chapitre 5 qui fait l'objet d'une publication dans le sixième volume de la revue ESAIM P&S daté de 2002.

Motivations

Nous observons ponctuellement (suivant un sens à préciser) un champ aléatoire $(Y_x)_{x \in \mathcal{X}}$ sur un espace probabilisé $(\Omega, \mathcal{A}, \mathrm{P})$ et à valeurs dans $\mathcal{Y} = \mathbb{R}^q$. L'ensemble $\mathcal{X}$ des indices est un espace probabilisé, muni de la probabilité P dont le support est $\mathcal{X}$ entier. Nous proposons d'adopter l'exemple où $\mathcal{X}$ est un sous-ensemble de $\mathbb{R}^d$.

Nous supposons que $\mathcal{X}$ est segmenté par une partition $\tau^\star = (\tau_j^\star)_{1 \leq j \leq K^\star} \in \mathcal{T}_{K^\star}$. De façon générale, $\mathcal{T}_K$ dénote l'ensemble des partitions τ de $\mathcal{X}$ de cardinalité K, *i.e.* les collections $\tau = (\tau_k)_{1 \leq k \leq K}$ telles que

$$P(\tau_h) > 0, P\left(\mathcal{X} \setminus \bigcup_{k=1}^{K} \tau_k\right) = 0 \quad \text{et} \quad P(\tau_h \cap \tau_{h'}) = 0$$

pour tout $1 \leq h < h' \leq K$. Les parties τ_k sont choisies dans une classe $\mathcal{S}$ d'ensembles mesurables fixée, stable par unions et intersections finies. La variété des ensembles contenus dans cette classe jouera un rôle prépondérant dans notre étude.

Nous supposons aussi que le champ $(Y_x)_{x \in \mathcal{X}}$ est segmenté par $\tau^\star$, dans la mesure où il existe un jeu de paramètres deux à deux distincts $\theta^\star = (\theta_j^\star)_{1 \leq j \leq K^\star} \in \Theta_{K^\star}$ tels que la loi de Y_x dépende de $\theta_j^\star$ si et seulement si $x \in \tau_j^\star$. De façon générale, Θ_K dénote l'ensemble des paramètres $\theta = (\theta_k)_{1 \leq k \leq K}$ à coordonnées dans un ouvert borné Θ de $\mathbb{R}^p$ fixé.

Les points d'observations sont des v.a. sur $(\Omega, \mathcal{A}, \mathrm{P})$ notées $X_1, \ldots, X_n$. Ils sont tirés indépendamment dans $\mathcal{X}$ suivant la loi P. Ils sont par ailleurs indépendants du champ $(Y_x)_{x\in\mathcal{X}}$. Les observations ponctuelles du champ en ces points sont notés $Y_i = Y_{X_i}$ $(i = 1, \ldots, n)$. En résumé, nous disposons de

$$Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$$

et notre tâche est d'estimer le couple $(\tau^\star, \theta^\star)$ dans la famille $\left(\mathcal{T}_K \times \Theta_K\right)_{1\leq K\leq K_{\max}}$.
$K_{\max}$ est une borne supérieure pour $K^\star$ que nous connaissons *a priori*.

Au titre de ce qui précède, ce problème rentre dans le cadre général de la segmentation spatiale et de la sélection de modèle.

Nous renvoyons aux trois monographies (Basseville and Nikiforov 1993; Brodsky and Darkhovsky 1993; Carlstein, Müller, and Siegmund 1994) pour une approche générale de ce thème sous le signe de la détection de ruptures, qui traite des processus caractérisés par une hétérogénéité à grande échelle et une homogénéité à petite échelle sur certaines régions.

La monographie (Korostelëv and Tsybakov 1993) aborde le thème par le biais de la reconstitution d'images par méthodes statistiques non paramétriques, dans la perspective asymptotique minimax. Le champ $(Y_x)_{x\in\mathcal{X}}$ y est supposé sans effet de dépendance.

Notre étude est de nature asymptotique. Nous concentrons nos efforts sur une propriété de consistance d'un estimateur de minimum de contraste pénalisé.

Etat de l'art

Comme on l'a déjà suggéré, la question de la segmentation peut être abordée sous le signe de la *détection de ruptures*, ou bien sous celui de la *reconstitution d'images*. Ces deux approches diffèrent de par la terminologie et les techniques associées, mais surtout parce que la première est généralement réservée aux modèles temporels (*i.e.* monodimensionnels), par opposition à la seconde, réservée aux modèles spatiaux (*i.e.* multidimensionnels).

Les gammes de techniques sont cependant variées au sein de chacune des approches globales. Nous pouvons citer* parmi les travaux récents relevant de la détection de ruptures des exemples de techniques non paramétriques ou bayésiennes (Antoniadis et al. 2000; Antoniadis and Gijbels 2002; Lavielle and Lebarbier 2001), ou bien des techniques de minimisation de contrastes comme dans (Lavielle 1999; Lavielle and Ludeña 2000; Lavielle and Moulines 2000) et dans le travail que nous commentons ci-dessous.

On trouve aussi des techniques de minimisation de contraste dans le cadre de la reconstitution d'images, comme par exemple dans (Mammen and Tsybakov 1995) (dans une perspective minimax de détermination de vitesses optimales) ou dans (Massart 2000) (dans une perspective de contrôle de risque à horizon fini).

Il nous semble que les deux articles (Lavielle 1999; Mammen and Tsybakov 1995) parmi ceux que nous citons ci-dessus s'imposent comme exemples de référence, au titre d'une certaine parenté avec notre travail.

*Il n'est pas question de dresser ici un portrait exhaustif de la littérature. Le lecteur intéressé pourra consulter les bibliographies incluses dans ces articles.

Lavielle (1999) considère le cas où $\mathcal{X} = [0,1]$ et les points d'observations $X_1, \ldots, X_n$ sont déterministes, uniformément éloignés les uns des autres. De plus, les partitions τ de son modèle sont constituées de segments disjoints, dont les extrémités sont les *instants de rupture* dans la loi marginale des Y_i. Le champ $(Y_x)_{x\in\mathcal{X}}$ peut en effet être sujet à des effets de dépendance. L'auteur prouve la consistance d'un estimateur des instants de ruptures et des paramètres associés à chaque segment lorsque le nombre de *vraies* ruptures est inconnu, borné *a priori*. La définition de l'estimateur d'intérêt repose sur la minimisation d'un contraste empirique pénalisé. Nous nous sommes inspirés de la preuve de Lavielle pour élaborer la nôtre. Il apparaît que le passage au cas spatial combiné à l'aléa des points d'observation complique sensiblement le problème et exige d'invoquer des propriétés fines de processus empiriques.

Mammen et Tsybakov (1995) considèrent quant à eux le cas où $\mathcal{X} = [0,1]^d$ et les points d'observations $X_1, \ldots, X_n$ sont indépendants, uniformément distribués. De plus, les partitions sont toutes de cardinalité 2 (l'image est en noir et blanc), à frontière régulière. Le champ $(Y_x)_{x\in\mathcal{X}}$ est sans effet de dépendance. Les auteurs déterminent les vitesses de convergence optimales pour leurs estimateurs de minimum de contraste dans une approche minimax. Leur preuves reposent fondamentalement sur des inégalités exponentielles qui découlent de l'indépendance des observations $Z_1, \ldots, Z_n$.

Techniques mises en œuvre et résultats

Notre preuve est valable pour une famille de contrastes dont nous précisons la forme ci-dessous. Par souci de cohérence, nous identifions un couple générique (τ, θ) et $(\tau^\star, \theta^\star)$ avec respectivement

$$f(x) = \sum_{k=1}^{K} \theta_k \mathbb{1}\{x \in \tau_k\} \quad \text{et} \quad f^\star(x) = \sum_{j=1}^{K^\star} \theta_j^\star \mathbb{1}\{x \in \tau_j^\star\} \quad (\text{tout } x \in \mathcal{X}).$$

Moyennant cela, les contrastes qui nous concernent ici peuvent s'écrire

$$\gamma(f, z) = \sum_{k=1}^{K} \Big\{ \varphi(\theta_k) + \psi(\theta_k)^T \xi(y) \Big\} \mathbb{1}\{x \in \tau_k\}$$

pour $z = (x, y)$, si f est associée à (τ, θ) de cardinalité K, *i.e.* où $\tau = (\tau_k)_{1\le k\le K}$ et $\theta = (\theta_k)_{1\le k\le K}$. Ici, φ et ψ sont deux fonctions de la fermeture $\overline{\Theta}$ dans $\mathbb{R}$ et $\mathbb{R}^r$, respectivement, deux fois continuement dérivables ; la fonction ξ de $\mathbb{R}^q$ dans $\mathbb{R}^r$ est telle que les v.a. $\xi(Y_x)$ (tout $x \in \mathcal{X}$) et $\xi(Y_X)$ soient intégrables relativement à P (X est de loi P) ; enfin, u^T représente la transposée du vecteur u.

Nous supposons que les contrastes sont appropriés, *i.e.* notant X_1^n le vecteur $(X_1, \ldots, X_n)$ et $\mathbb{P}_n$ la mesure empirique correspondante, P-presque sûrement pour tout j compris entre 1 et $K^\star$ et $\theta \in \overline{\Theta}$,

$$\mathrm{E}\left[n^{-1} \sum_{i=1}^{n} \Big\{ \varphi(\theta) + \psi(\theta)^T \xi(Y_i) \Big\} \mathbb{1}\{X_i \in \tau_j^\star\} \Big| X_1^n \right] = \mathbb{P}_n(\tau_j^\star)\, w(\theta_j^\star, \theta)$$

pour une fonction déterministe w de $\{\theta_1^\star, \ldots, \theta_{K^\star}^\star\} \times \overline{\Theta}$ dans $\mathbb{R}$ qui satisfait par ailleurs

$$v(\theta_j^\star, \theta) = w(\theta_j^\star, \theta) - w(\theta_j^\star, \theta_j^\star) \ge 0,$$

égalité si et seulement si $\theta = \theta_j^\star$.

Cette condition d'adéquation du contraste au modèle statistique est par exemple remplie pour les deux exemples de modèles de régression **HO** et **HE**, puisque :

HO : on obtient (à une constante près qui ne dépend que de z, et par conséquent ne joue pas dans la procédure de minimisation) $\gamma_1(f,z) = (y - f(x))^2$ pour $\varphi(\theta) = \theta^2$, $\psi(\theta) = -2\,\theta$ et $\xi(y) = y$. Dans ce cas, pour tout $1 \leq j \leq K^\star$ et $\theta \in \overline{\Theta}$,

$$v(\theta_j^\star, \theta) = (\theta - \theta_j^\star)^2.$$

HE : Ici, le paramètre θ est un couple de moyenne et écart-type (m,s) et f se décompose en $f = (f_{\mathrm{m}}, f_{\mathrm{se}})$ de la même façon qu'exposée plus haut.
On obtient $\gamma_2(f,z) = (y - f_{\mathrm{m}}(x))^2/f_{\mathrm{se}}(x)^2 + \log f_{\mathrm{se}}(x)^2$ pour $\varphi(\theta) = m^2/s^2 + \log s^2$, $\psi(\theta) = (-2\,m/s^2, 1/s^2)$ et $\xi(y) = (y, y^2)$. Dans ce cas, pour tout $1 \leq j \leq K^\star$ et $\theta \in \overline{\Theta}$,

$$v(\theta_j^\star, \theta) = H(\mathcal{N}_{\theta_j^\star} | \mathcal{N}_\theta),$$

où $\mathcal{N}_\theta$ désigne la loi gaussienne de moyenne et écart-type θ.

Les hypothèses majeures se répartissent naturellement en deux groupes, selon qu'elles concernent les points observations $X_1, \ldots, X_n$ à travers leur distribution commune P ou bien les réponses $Y_1, \ldots, Y_n$ du champ $(Y_x)_{x \in \mathcal{X}}$ en ces points.

Tout d'abord, nous requérons que la classe $\mathcal{S}$ dans laquelle sont choisies les parties τ_k des partitions τ soit P-Glivenko-Cantelli, *i.e.* que la v.a.* $\sup\{|\mathbb{P}_n(S) - P(S)| : S \in \mathcal{S}\}$ converge P-presque sûrement vers 0. De façon imagée, cela signifie qu'asymptotiquement, toutes les formes $S \in \mathcal{S}$ sont uniformément visitées par les points d'observation au prorata de leurs mesures $P(S)$. De plus, nous demandons l'existence d'une suite $\{r_n\} \downarrow 0$ telle que $\liminf_n n\,r_n > 0$ et que

$$\lim_{\eta \to \infty} \lim_{n \to \infty} \mathrm{P}\left(\sup \left\{ \frac{P(F) - \mathbb{P}_n(F)}{P(F)} \,:\, F \in \mathcal{F}, P(F) \geq \eta r_n \right\} \geq \frac{1}{2} \right) = 0.$$

Nous prouvons (Proposition 5.3.3) grâce à des arguments classiques de symétrisation (voir Massart 2000) et une inégalité de concentration (empruntée à van der Vaart and Wellner 1996) que la condition ci-dessus est satisfaite si la classe $\mathcal{S}$ est une classe de Vapnik-Červonenkis (abrégé VC) de dimension de VC finie* et si de plus la suite $\{(n\,r_n)^{-1} \log r_n\}$ est bornée.

L'hypothèse principale relative aux Y_i consiste à contrôler leurs fluctuations grâce à une *inégalité maximale*. Notons par souci de concision

$$\Sigma_{X_1^n}(S) = \sum_{i=1}^{n} \Big\{ \xi(Y_i) - \mathrm{E}\Big(\xi(Y_i) | X_i \Big) \Big\} \mathbb{1}\{X_i \in S\} \quad (\text{tout } S \in \mathcal{S})$$

et $\|\cdot\|_\infty$ la norme uniforme sur $\mathbb{R}^r$. Alors la condition peut s'énoncer comme suit :

*Pour une présentation récente de la propriété P-Glivenko-Cantelli, nous renvoyons à (van der Vaart 1998). Quant à la mesurabilité du supremum, nous supposons que toutes les expressions qui font intervenir des suprema coïncident P-presque sûrement avec des suprema sur un ensemble dénombrable. Ceci est valable dans toute cette thèse.

*Nous renvoyons de nouveau à (van der Vaart 1998) pour le détail. Il suffit d'avoir à l'idée que la dimension de VC quantifie la richesse de la classe.

Il existe $C_1 > 0$ *et* $h \in (1,2)$ *tels que, pour tout* $\delta > 0$, $G \in \mathcal{S}$, P-*presque sûrement*

$$\mathrm{P}\left(\sup\left\{\|\Sigma_{X_1^n}(S \cap G)\|_\infty \,:\, S \in \mathcal{S}\right\} \geq \delta \,\middle|\, X_1^n\right) \leq \frac{C_1}{\delta^2}\left(\sum_{i=1}^n \mathbb{1}\{X_i \in G\}\right)^h. \qquad (5)$$

Par définition, notre estimateur $\widehat{f}$ qui s'identifie au couple $(\hat{\tau}_n, \hat{\theta}_n)$ est un minimiseur en $K \leq K_{\max}$ et f identifié à (τ, θ) dans $\left(\mathcal{T}_K \times \Theta_K\right)_{1 \leq K \leq K_{\max}}$ de

$$n^{-1}\sum_{i=1}^n \gamma(f, Z_i) + \mathrm{pen}(n, K)$$

si f est de cardinalité K (*i.e.* la partition τ associée est de cardinalité K) et où notre pénalité s'écrit $\mathrm{pen}(n,K) = v_n K$.

Nous prouvons sa consistance (Théorème 5.5.2) et évaluons sa vitesse de convergence (Théorème 5.5.4) *lorsque la cardinalité* $K^\star$ *de* $\tau^\star$ *est* a priori *connue*. Dans ce cas, la pénalisation ne joue évidemment aucun rôle. En particulier, nous montrons que la suite $\{r_n^{-1}\, g(\hat{\tau}_n, \tau^\star)\}$ est uniformément bornée en probabilité. Ici, $g(\hat{\tau}_n, \tau^\star)$ quantifie un écart entre $\hat{\tau}_n$ et $\tau^\star$. Cet écart original n'est pas une distance entre partitions, mais jouit néanmoins de propriétés remarquables que nous exploitons largement. En particulier, $g(\tau, \tau') = 0$ si et seulement si τ est une partition plus fine que τ'. Si les cardinalités coïncident, τ et τ' sont identiques.

Enfin, nous prouvons la consistance de l'estimateur *sans connaissance* a priori *de* $K^\star$ (Théorème 5.6.1) dès lors que $\{v_n\}$ satisfait la condition de calibration

$$v_n = o(1) \quad \text{et} \quad n^{(h-2)/2}\, v_n^{-1} = o(1).$$

Par consistance, nous entendons ici que la cardinalité $\widehat{K}_n$ de $\hat{\tau}_n$ tend en probabilité vers $K^\star$ (*i.e.* l'ordre est bien estimé) et que par conséquent, les résultats de consistance ci-dessus à $K^\star$ connu *a priori* s'appliquent avec une probabilité tendant vers 1.

Nous attachons un soin tout particulier à l'exploration de l'hypothèse concernant l'inégalité maximale (5). Nous lui trouvons deux alternatives, fondées d'une part (Proposition 5.7.1) sur l'adaptation multidimensionnelle d'une inégalité de moments classique due à Móricz et al. (1982) – celle-ci est significative lorsque la classe $\mathcal{S}$ est composée de rectangles ; d'autre part (Proposition 5.7.3) sur une inégalité à la Marcinkiewicz-Zygmund (voir Rio 2000; Dedecker 2001). La preuve repose sur l'application d'une méthode de Pisier* permettant de basculer d'un contrôle fin de moments de tous ordres $p > 2$ à l'inégalité maximale d'intérêt par optimisation en p (voir Dedecker 2001). La puissance de la dite méthode assure que cette alternative est valable à condition que la classe $\mathcal{S}$ soit une classe de VC de dimension de VC finie.

Pour conclure, nous démontrons une ultime alternative commune aux deux alternatives évoquées ci-dessus dans le cas où $\mathcal{X} = \mathbb{Z}^d$ est un réseau et $(Y_x)_{x \in \mathcal{X}}$ est borné et stationnaire. Des arguments inspirés largement de (Dedecker 2001), et qui reposent sur des inégalités à la Burkholder (1973) et Serfling (1968), aboutissent à des conditions très concrètes de décroissance (en fonction de l'éloignement) de la dépendance dans le champ (exprimée par le coefficient de ϕ-mélange).

*Nous tenons à remercier chaleureusement Jérôme Dedecker pour l'introdution limpide aux techniques de mélange qu'il nous a prodiguée.

Nous étendons les résultats de (Lavielle 1999) au cas spatial, avec aléa des points d'observation. Les preuves demandent un investissement théorique et technique important. Nous nous consacrons par ailleurs avec beaucoup de soin à l'illustration des hypothèses formulées sur la nature de la dépendance dans le champ $(Y_x)_{x\in\mathcal{X}}$.

La comparaison avec (Mammen and Tsybakov 1995) est malaisée, tant l'hypothèse d'indépendance des observations est cruciale pour leur approche du problème, tandis que nous nous en affranchissons. Les auteurs l'exploitent cependant largement, puisqu'ils déterminent des vitesses de convergence optimales dans l'approche minimax, lorsque les nôtres ne le sont *a priori* pas. Par ailleurs, notre estimateur n'est pas fondé sur une connaissance *a priori* de la cardinalité $K^\star$, quand elle est fixée égale à 2 dans leur travail. Enfin, les conditions sur les parties τ_k qui composent les partitions sont de natures différentes : conditions de régularité de la frontière pour eux ; conditions combinatoires pour ce qui nous concerne (*via* la dimension de VC).

Estimation de l'ordre d'un modèle

Nous résumons dans la présente section le contenu de l'Interlude 6 et du Chapitre 7.

Motivations

Le problème statistique d'intérêt est le suivant : étant donné un jeu $Z_1, \ldots, Z_n$ d'observations indépendantes et identiquement distribuées, étant donnée une famille croissante $\{\Pi_K\}_{K\geq 1}$ de modèles emboîtés indexés par leur ordre K (qui, conformément à la terminologie introduite plus haut, quantifie la complexité du modèle), nous souhaitons estimer l'ordre du modèle dont est issue la loi commune des observations, le cas échéant.

Nous supposons par ailleurs que les modèles Π_K sont paramétriques, *i.e.* qu'il existe une famille croissante d'espaces de paramètres $\{\Theta_K\}_{K\geq 1}$ tels que $\Pi_K = \{P_\theta : \theta \in \Theta_K\}$, où P_θ dénote une probabilité. Enfin, si P_θ appartient à $\Pi_K \setminus \Pi_{K-1}$ (avec la convention $\Pi_0 = \emptyset$), nous disons que P_θ est d'ordre K. Ainsi, lorsque la loi commune $P^\star$ des observations appartient à $\Pi_{K^\star} \setminus \Pi_{K^\star-1}$, l'objectif est d'estimer $K^\star$.

A ce titre, ce problème rentre dans le cadre général de la sélection de modèle.

Notre étude est de nature asymptotique. Nous concentrons nos efforts sur les propriétés de consistance de deux estimateurs complémentaires de maximum de vraisemblance pénalisée. Nous nous consacrons aussi aux vitesses de sous- et sur-estimation, *i.e.* aux vitesses auxquelles décroissent les probabilités respectives de sous-estimer et de sur-estimer l'ordre du modèle qui régit la production au hasard des observations. Ces trois phénomènes sont de natures statistiques différentes et requièrent donc des études distinctes.

Voici trois exemples-phares qui sont l'objet de tous nos soins (les abréviations reprennent celles du Chapitre 7, qui est rédigé en anglais). Nous prenons le temps de les présenter en détail car ils illustreront la généralité des résultats que nous obtenons.

Modèle segmenté (AC) : Soit $(\mathcal{X}, \mathcal{B}, P)$ un ouvert de $\mathbb{R}^q$ muni de la trace de la tribu des Boréliens et d'une mesure de probabilité P dominée par la mesure de Lebesgue $\mu^{\otimes q}$, dont la densité relativement à cette dernière est notée p. Soit aussi $\mathcal{Y}$ une partie de $\mathbb{R}$ et

ζ une constante strictement positive. Nous notons $\mathcal{T}$ l'ensemble des partitions finies de Caccioppoli* de $\mathcal{X}$ à périmètres majorés uniformément par ζ.

Il existe une métrique d_P sur $\mathcal{T}$ pour laquelle $(\mathcal{T}, d_P)$ est un espace métrique compact et telle que l'ensemble $\mathcal{T}_{\leq K}$ des partitions de cardinal au plus K (nous reprenons la définition des partitions et la terminologie associée telles qu'exposées plus tôt) est fermé dans $\mathcal{T}$ pour la topologie due à d_P.

Nous choisissons maintenant une partie compacte $\mathcal{M}$ de $\mathbb{R}$ et définissons, pour tout ordre $K \geq 1$, pour toute partition $\tau \in \mathcal{T}_K$ (le sous-ensemble de $\mathcal{T}$ constitué des partitions de cardinalité K) et pour tout $\mathbf{m} = (m_1, \ldots, m_K) \in \mathcal{M}^K$,

$$\begin{aligned} &\theta = (\tau, \boldsymbol{m}), \\ &f_\theta(x) = \sum_{k=1}^{K} m_k \mathbb{1}\{x \in \tau_k\} \quad (\text{tout } x \in \mathcal{X}) \quad \text{et} \\ &p_\theta(x, y) = G(y\,; f_\theta(x))\, p(x) \quad (\text{tout } x \in \mathcal{X},\ y \in \mathcal{Y}) \end{aligned}$$

où $G(\cdot\,; m)$ est le noyau réel gaussien de moyenne m et de variance σ^2 fixée. Alors pour tout ordre $K \geq 1$,

$$\begin{aligned} \Pi_1 &= \left\{p_\theta \, \mathrm{d}\mu^{\otimes q+1} : \theta \in \Theta_1\right\} \quad \text{avec} \quad \Theta_1 = \{\mathcal{X}\} \times \mathcal{M} \\ \Pi_{K+1} &= \Pi_K \cup \left\{p_\theta \, \mathrm{d}\mu^{\otimes q+1} : \theta \in \Theta_{K+1}\right\} \quad \text{avec} \quad \Theta_{K+1} = \Theta_K \cup \left(\mathcal{T}_{K+1} \times \mathcal{M}^{K+1}\right). \end{aligned}$$

Ce cadre de travail correspond au modèle segmenté dans lequel la v.a. $Z = (X, Y)$ est observée, avec

$$Y = f^\star(X) + e$$

où $f^\star(x) = \sum_j m_j^\star \mathbb{1}\{x \in \tau_j^\star\}$ $(x \in \mathcal{X})$ pour une partition $\tau^\star$ appartenant à $\mathcal{T}_{K^\star}$, $K^\star$ inconnu ; où X suit la loi $p\,\mathrm{d}\mu^{\otimes q}$; et où e suit une loi gaussienne centrée et de variance σ^2, indépendante de X. Ainsi, $Z = (X, Y)$ a pour densité $p_{\theta^\star}$ si $\theta^\star = (\tau^\star, \mathbf{m}^\star)$ (ici, $\mathbf{m}^\star = (m_1^\star, \ldots, m_{K^\star}^\star)$).

Régression homoscédastique (HO) : Soit $\mathcal{X} = [0, 1]$ et $\mathcal{Y} = \mathbb{R}$ munis de la mesure μ de Lebesgue sur les Boréliens. Soit $\{t_K\}_{K \geq 1}$ un système libre de fonctions continues sur $\mathcal{X}$ et $\mathcal{U}$ un compact de $\mathbb{R}$ contenant 0. Nous supposons que la famille de réels $\{t_K(\mathcal{X})\}_{K \geq 1}$ est bornée. Définissons $\Theta_K = \mathcal{U}^K$ et pour tout $\theta \in \Theta_K$ et $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned} f_\theta(x) &= \sum_{k=1}^{K} \theta_k t_k(x) \quad \text{et} \\ p_\theta(z) &= G(y\,; f_\theta(x)), \end{aligned}$$

où $G(\cdot\,; m)$ est comme dans l'exemple précédent. Ici, K est le nombre maximal de fonctions de base t_k qui interviennent dans la définition de f_θ $(\theta \in \Theta_K)$.

*L'Annexe B est consacrée à la présentation de ces ensembles de partitions. Elle est fondée sur la prépublication (Leonardi and Tamanini 2002) qui nous a été conseillée gentiment et savamment par Raphaël Cerf, que nous remercions encore ici. Leur introduction est motivée par l'existence d'une structure d'espace métrique compact sous des hypothèses faibles.

Ce cadre de travail correspond au modèle de régression homoscédastique

$$Y = f^\star(X) + e$$

où X est distribuée uniformément sur $[0,1]$ indépendamment de e, de loi gaussienne centrée et de variance σ^2, et $f^\star = \sum_{k=1}^{K^\star} \theta_k^\star t_k$ pour un certain ordre $K^\star$ inconnu et pour la condition $\theta_{K^\star}^\star \neq 0$. Ainsi, $Z = (X, Y)$ admet pour densité $p_{\theta^\star}$ relativement à μ.

Mélange de gaussiennes en la moyenne (MGM) : Soit $\mathcal{Z} = \mathbb{R}$ et $\mathcal{D}$ l'ensemble des densités gaussiennes réelles G_m à moyennes m dans le compact $\mathcal{M}$ de $\mathbb{R}$ et de même variance σ^2 fixée.

Nous définissons $\Pi_1 = \mathcal{D}$ et pour tout $K \geq 2$, l'ensemble Π_K des mélanges finis d'ordre K

$$\Pi_K = \left\{ p_\theta \, \mathrm{d}\mu = \sum_{k=1}^{K-1} \pi_k G_{m_k} + \left(1 - \sum_{k=1}^{K-1} \pi_k\right) G_{m_K} \, \mathrm{d}\mu : \theta = (\boldsymbol{\pi}, \boldsymbol{m}) \in \Theta_K \right\},$$

où $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{K-1})$ est un vecteur de nombres positifs tels que $\sum_k \pi_k \leq 1$ et $\boldsymbol{m}$ est un vecteur de moyennes à coordonnées dans $\mathcal{M}$.

Ici, l'ordre K du modèle Π_K est le nombre maximal de densités élémentaires mélangées dans toute densité $p_\theta \in \Pi_K$.

Ce cadre de travail correspond au modèle de mélange de populations où la variable cachée (*i.e.* non observée) X prend ses valeurs dans un ensemble fini inconnu $\{m_1^\star, \ldots, m_{K^\star}^\star\}$ de $\mathcal{M}$ (nous ne connaissons pas $K^\star$ non plus), suivant la loi inconnue également (appelée *loi de mélange*) $\mathrm{P}(X = m_j^\star) = \pi_j^\star$ $(j = 1, \ldots, K^\star)$ et, e étant gaussienne centrée de variance σ^2 et indépendante de X,

$$Z = X + e.$$

Ainsi, Z admet pour densité $p_{\theta^\star}$ si $\theta^\star = (\boldsymbol{\pi}^\star, \mathbf{m}^\star)$, avec $\boldsymbol{\pi}^\star = (\pi_1^\star, \ldots, \pi_{K^\star-1}^\star)$ et $\mathbf{m}^\star = (m_1^\star, \ldots, m_{K^\star}^\star)$.

Remarque 2. Le modèle de mélange **MGM** est très délicat à étudier, notamment parce que les proportions de mélange pour un modèle donné Π_K sont dans la frontière de l'espace des paramètres $\Theta_{K'}$ de tout modèle $\Pi_{K'}$ le contenant (*i.e.* ici tel que $K \geq K'$). Dans le même ordre d'idées, une loi de mélange $P_\theta \in \Pi_K \setminus \Pi_{K-1}$ à K composantes peut coïncider simultanément avec deux loi P_{θ_1} et P_{θ_2} de Π_{K+1} bien que θ_1 et θ_2 diffèrent : cette perte d'identifiabilité est la source de sérieuses difficultés lorsque l'on cherche à adapter les schémas classiques de preuve.

Pour ces raisons, le modèle **MGM** jouit d'un statut particulier *d'exemple de référence* dans le cadre de l'étude de l'estimation de l'ordre d'un modèle.

Pour davantage de détails sur les modèles de mélange, nous renvoyons aux monographies (Everitt and Hand 1981; Titterington et al. 1985; McLachlan and Basford 1988) et à l'article (Lindsay and Lesperance 1995).

Etat de l'art

Il ne saurait être question de référencer toute la production statistique consacrée à l'estimation de l'ordre. Nous nous sommes efforcés de sélectionner des articles importants aux titres des résultats contenus, des avancées apportées, des méthodes employées et des bibliographies proposées. Nous encourageons le lecteur intéressé à consulter les articles eux-mêmes pour plus de détails.

Consistance

Pour la consistance, on trouve quatre types de résultats dans la littérature selon les méthodes d'estimation et les progrès de la recherche. Ainsi, la consistance est toujours au sens de consistance presque sûre dans les articles que nous citons. En revanche, la consistance peut s'appuyer ou non sur une hypothèse de connaissance d'une borne *a priori* pour l'ordre $K^\star$ du vrai modèle. Cette hypothèse est assez peu naturelle mais souvent indispensable. Par ailleurs, on se trouve *de facto* dans une situation où $K^\star$ est bornée lorsque l'on en programme une méthode d'estimation. On peut noter que l'hypothèse n'est bien entendu pas requise pour les résultats de non sous-estimation (*i.e.* estimation correcte ou sur-estimation) presque sûre. Enfin, certains auteurs sont parvenus à obtenir un contrôle non asymptotique de la probabilité de mauvaise estimation, dont ils déduisent la consistance.

Par ailleurs, la littérature sur l'estimation de l'ordre se partage naturellement en deux groupes (non disjoints) d'articles, selon que les méthodes mises en œuvre s'appliquent ou non aux modèles de mélange. Ceci est dû en particulier à la difficulté intrinsèque que pose la question de l'estimation de l'ordre pour ces modèles.

Nous commençons par les résultats qui concernent les modèles de mélange. Ainsi, Henna (1985) a prouvé la consistance *sans borne* a priori de son estimateur issu d'une procédure de moindres carrés.

L'année suivante, Antoniadis et Berruyer (1986) ont construit un estimateur consistant *sans borne* a priori de l'ordre d'un mélange de lois exponentielles à un paramètre (cela inclut l'exemple **MGM**). Cet estimateur exploite certaines propriétés de matrices particulières (dites de Hankel) associées aux moments de la loi inconnue de mélange.

Bien plus tard, Dacunha-Castelle et Gassiat (1997) ont de nouveau mis en scène un estimateur de l'ordre d'un mélange général fondé sur les matrices de Hankel et une approche différente de la précédente. Les auteurs ont prouvé que leur estimateur est consistant *sans borne* a priori et ont obtenu un contrôle non asymptotique de la probabilité de mauvaise estimation. L'hypothèse principale demande l'existence d'estimateurs consistants des moments de la loi inconnue de mélange et la méthode peut s'appliquer à d'autres cadres de travail que celui des mélanges, à condition que cette hypothèse soit vérifiée.

Leroux (1992) a mis en œuvre une méthode d'estimation par maximum de vraisemblance pénalisée,* dont il a prouvé qu'elle produit des estimateurs qui presque sûrement ne sous-estiment pas l'ordre du mélange. Il faut attendre 2000 et les travaux de Keribin pour une preuve de la presque sûre non sur-estimation, d'où finalement la consistance presque sûre. Ce résultat requiert toutefois une *borne* a priori *sur l'ordre* $K^\star$ du vrai modèle. L'avancée due à Keribin est permise par l'introduction de la *paramétrisation localement conique* à laquelle Dacunha-Castelle et Gassiat consacrent un article en 1999. Cette paramétrisation permet de surmonter les difficultés provoquées par la non identifiabilité évoquée plus tôt et d'adapter les méthodes classiques de preuve de consistance par développement du rapport de vraisemblances. Dans ce même article, les auteurs traitent largement la question du test de l'ordre du vrai modèle de mélange. Finalement, Gassiat (2002) a démontré récemment deux inégalités simples et néanmoins très puissantes sur les rapports de vraisemblances. Celles-ci permettent par exemple à leur auteur de prouver la consistance d'un estimateur du nombre de populations (*i.e.* de l'ordre) dans un

*C'est justement à une procédure de maximisation de vraisemblance pénalisée que nous nous sommes consacrés.

mélange à régime markovien par maximum de vraisemblance pénalisée. La connaissance d'une *borne* a priori est requise.

Des travaux que nous citons, le plus récent dédié à l'estimation de l'ordre d'un mélange est à l'actif de James et al. (2001), qui explorent une méthode semi-paramétrique d'estimation, consistante *sans borne* a priori.

De nouveau dans le domaine des procédures de maximisation de vraisemblance pénalisée, les travaux de (Haughton 1988) sur la sélection de modèle dans une famille exponentielle et leur généralisation à des familles régulières (Haughton and Keribin 2001) s'appliquent au problème d'estimation d'un ordre de modèle, mais les mélanges sont exclus. Leurs résultats de consistance reposent notamment sur une *borne* a priori et sur l'utilisation de la paramétrisation localement conique permettant de procéder par développement du rapport de vraisemblances.

Récemment toujours, Guyon et Yao (1999) ont prouvé la consistance d'un estimateur de l'ordre par minimisation de contraste pénalisé. Deux hypothèses majeures sont invoquées. La première assure une factorisation du contraste empirique comme produit d'une statistique exhaustive des observations et d'une fonction déterministe des paramètres. La seconde assure l'identifiabilité des modèles. Une *borne* a priori est requise et le résultat s'applique à une grande variété de modèles (dont **HO**) et pour des variables éventuellement dépendantes. Néanmoins, les hypothèses majeures excluent *de facto* le modèle de mélange.

Enfin, nous étudions dans le Chapitre 5 une procédure d'estimation de l'ordre dans un modèle segmenté du type de **AC** pour des observations éventuellement dépendantes. Une description précise en termes de techniques et résultats en a été faite dans la section précédente.

Sous- et sur-estimation

Ici encore, nous préciserons bien quand la connaissance d'une *borne* a priori *sur l'ordre $K^\star$ du vrai modèle* est invoquée pour obtenir des résultats de sur-estimation.

Dans un cadre d'étude de modèles de mélange, Dacunha-Castelle et Gassiat (1997) ont obtenu un contrôle à horizon fini de la probabilité de mal estimer l'ordre du vrai modèle par une expression décroissant exponentiellement vite avec le nombre d'observations. C'est à notre connaissance le seul résultat concernant les probabilités de sous- et sur-estimation applicable aux mélanges.

Dans les autres cadres de travail, et parmi les articles déjà cités ci-dessus, tous exigent une *borne* a priori lorsqu'il est question de la probabilité de sur-estimation.

Guyon et Yao (1999) prouvent une majoration non asymptotique des probabilités de sous- et sur-estimer, en se fondant sur une inégalité exponentielle d'écart à son espérance pour la statistique exhaustive de factorisation.

Finalement Haughton et Keribin (2001) parviennent à contrôler asymptotiquement les probabilités de sous- et sur-estimation. Le schéma de preuve s'articule encore autour d'un développement du rapport de vraisemblances et sur l'application de résultats de Moyennes et Grandes Déviations, sous des hypothèses lourdes de double ou triple différentiabilité des log-vraisemblances, d'inversibilité de la matrice d'information de Fisher et d'existence de moments exponentiels pour le gradient et pour le supremum de la Hessienne.

Nous expliquons ci-dessous comment nous avons essayé d'unifier les cadres de travail, proposant l'étude de deux estimateurs complémentaires de l'ordre d'un modèle sous des hypothèses

aussi légères que possible. Cette tâche a été rendue possible grâce à une approche originale, inspirée notamment de (Gassiat and Boucheron 2001), où les auteurs s'intéressent à l'estimation de l'ordre d'un Modèle de Markov Caché.

Techniques mises en œuvre et résultats

L'originalité de notre approche tient notamment

- d'une part dans l'identification de la mesure empirique $\mathbb{P}_n$ avec une forme linéaire sur une classe de fonctions qui contient les log-vraisemblances de nos modèles, *i.e.* les $\ell_\theta = \log p_\theta$ ($\theta \in \Theta_K$, $K \geq 1$) ;
- d'autre part dans le fait que nous n'exploitons pas de propriétés de l'estimateur du maximum de vraisemblance du paramètre $\theta^\star$ associé à la loi $P^\star$ du modèle qui génère nos observations, et ce bien que nos estimateurs de l'ordre reposent sur une procédure de maximisation de vraisemblance.

Il est d'ailleurs à ce titre naturel de comparer tout particulièrement nos résultats à ceux de Leroux, Haughton et Keribin, qui nous ont précédés dans l'étude d'estimateurs de l'ordre fondés sur des méthodes de maximisation de vraisemblance.

Le Chapitre 7 est donc dédié à l'étude de propriétés asymptotiques des deux estimateurs de l'ordre

$$\widehat{K}_n^{\mathrm{L}} = \inf\left\{ K \geq 1 : \sup_{\theta\in\Theta_K} \mathbb{P}_n \ell_\theta - n^{-1}\mathrm{pen}(n,K) \geq \sup_{\theta\in\Theta_{K+1}} \mathbb{P}_n \ell_\theta - n^{-1}\mathrm{pen}(n,K+1) \right\},$$

$$\widehat{K}_n^{\mathrm{G}} = \arg\sup_{K\geq 1} \left\{ \sup_{\theta\in\Theta_K} \mathbb{P}_n \ell_\theta - n^{-1}\mathrm{pen}(n,K) \right\},$$

où pen est une pénalité positive. Nous soulignons que $n\mathbb{P}_n\ell_\theta$ est la log-vraisemblance des observations, soit

$$n\mathbb{P}_n\ell_\theta = -\sum_{i=1}^{n} \gamma_0(p_\theta, Z_i)$$

si l'on reprend les notations de l'exemple **DS** dans le cadre duquel ceci s'inscrit.

$\widehat{K}_n^{\mathrm{L}}$ est le premier maximiseur local du critère (d'où la lettre L) et $\widehat{K}_n^{\mathrm{G}}$ est un maximiseur global (d'où la lettre G). L'étude en parallèle de ces deux estimateurs est motivée par l'éclairage qu'apportent leurs qualités et défauts respectifs sur la compréhension que nous avons d'eux.

Nous pouvons souligner d'ores et déjà que $\widehat{K}_n^{\mathrm{L}}$ est algorithmiquement plus intéressant que $\widehat{K}_n^{\mathrm{G}}$ puisqu'il suffit de calculer des contrastes sur des modèles jusqu'à la première inversion du sens de variation.

Avant de poursuivre, précisons ici qu'une des hypothèses fondamentales dans notre étude est la *compacité* de chaque ensemble Π_K pour la topologie faible sur l'ensemble $M_1(\mathcal{Z})$ des probabilités sur $\mathcal{Z}$. Celle-ci est vérifiée pour les exemples **HO** et **MGM**. Elle l'est aussi pour l'exemple **AC**, grâce à la théorie des partitions de Caccioppoli.

Consistance

Nous commençons par régler le cas où la famille de modèles n'a pas été bien choisie, au sens où (exceptionnellement) $P^\star \notin \cup_K \Pi_K$. Dans ce cas, les estimateurs $\widehat{K}_n^{\mathrm{L}}$ et $\widehat{K}_n^{\mathrm{G}}$ tendent vers l'infini presque sûrement (Théorème 7.4.1).

Nous noterons par la suite $H(P\,|\Pi)$ pour l'infimum $\inf\{H(P\,|Q) : Q \in \Pi\}$ et $H(\Pi\,|P)$ pour l'infimum $\inf\{H(Q\,|P) : Q \in \Pi\}$.

Nous sommes amenés à invoquer une hypothèse de renforcement de l'inclusion du modèle générique Π_K dans son successeur Π_{K+1} en exigeant que, si $P^\star$ n'appartient pas à Π_K, alors la divergence $H(P^\star\,|\Pi_{K+1})$ est nécessairement strictement plus petite que $H(P^\star\,|\Pi_K)$. Le théorème s'applique aux deux exemples **AC** et **MGM**, mais pas à **HO**.

Cette hypothèse de renforcement est invoquée dans tous les résultats de consistance relatifs à $\widehat{K}_n^{\mathrm{L}}$. Elle ne l'est en revanche pas pour ceux relatifs à $\widehat{K}_n^{\mathrm{G}}$. Au contraire, tous les résultats de consistance relatifs à $\widehat{K}_n^{\mathrm{G}}$ requièrent la connaissance d'une borne $K_{\max}$ *a priori* sur $K^\star$, hypothèse inutile lorsque l'on considère ceux relatifs à $\widehat{K}_n^{\mathrm{L}}$.

Désormais, la famille de modèles aura été bien choisie, de sorte que $P^\star \in \Pi_{K^\star} \setminus \Pi_{K^\star-1}$, avec $P^\star = P_{\theta^\star}$ pour un paramètre $\theta^\star \in \Theta_{K^\star}$.

Nous prouvons un premier résultat de consistance presque sûre pour $\widehat{K}_n^{\mathrm{L}}$ et $\widehat{K}_n^{\mathrm{G}}$ moyennant, entre autres, les hypothèses distinctes ébauchées plus haut (Théorème 7.4.3). Celui-ci est valable à condition que la pénalité satisfasse, pour tout $K \geq 1$,

$$\liminf_{n\to\infty} \frac{\mathrm{pen}(n, K+1)}{\mathrm{pen}(n, K)} > 1 \quad \text{et} \quad \limsup_{n\to\infty} \frac{(n \log\log n)^{1/2}}{\mathrm{pen}(n, K)} = 0.$$

La preuve en est en deux volets, l'un pour la sous-estimation, l'autre pour la sur-estimation. Dans ce dernier, la clef de voûte est une Loi du Logarithme Itéré borné fonctionnelle due à Dudley et Philipp (1983). Nous pouvons l'utiliser par exemple si une certaine classe de rapports de vraisemblances du type $g_\theta = (\ell_\theta - \ell^\star)$ (tout $\theta \in \Theta_K$) est $P^\star$-Donsker* à enveloppe dans $L^2(P^\star)$. Le résultat s'applique bien aux exemples **HO** et **MGM**. Pour l'exemple **AC** en revanche, la vérification de l'hypothèse Donsker est difficile. Nous sommes toutefois assurés que le volet de presque sûre non sous-estimation s'applique.

Un procédé permet de raffiner considérablement le précédent résultat de consistance vis-à-vis des contraintes sur la pénalité, qui doit satisfaire dans cette seconde version, pour tout $K \geq 1$,

$$\liminf_{n\to\infty} \frac{\mathrm{pen}(n, K+1)}{\mathrm{pen}(n, K)} > 1 \quad \text{et} \quad \limsup_{n\to\infty} \frac{\log\log n}{\mathrm{pen}(n, K)} = 0.$$

La preuve de ce résultat (Théorème 7.4.5) suit les mêmes lignes que la précédente après qu'on a appliqué un « tour à la Huber* ». Il s'agit en fait d'une illustration du principe général selon lequel un recentrage et une renormalisation peuvent conduire à des gains en performance sensibles pour les méthodes impliquant des processus empiriques. L'idée originale revient à Huber (1967) ; on en trouve de nombreuses illustrations dans la littérature et notamment dans

*Pour la définition de la propriété de Donsker, nous renvoyons à (van der Vaart 1998). Quant à la condition sur l'enveloppe, elle peut être relachée légèrement.

*Nous remercions Pascal Massart qui a attiré notre attention sur ce « tour ».

la littérature de la sélection de modèle, voyez par exemple l'équation (4) rencontrée lors de la description schématique de la sélection de modèle selon Birgé et Massart (1999, 2000).

Pour ce qui nous concerne, la forme que prend le « tour à la Huber » tient dans la comparaison ci-dessous. Nous notons $H(\theta) = H(P^\star \,|P_\theta)$ (tout $\theta \in \cup_K \Theta_K$).

Soit $K_2 > K_1 \geq K^\star$. Alors les deux inégalités suivantes sont satisfaites (Proposition 7.4.10), la seconde offrant une illustration de « tour à la Huber » :

$$\sup_{\theta\in\Theta_{K_2}} (\mathbb{P}_n - P^\star)(\,\ell_\theta - \ell^\star) \geq \sup_{\theta\in\Theta_{K_2}} \mathbb{P}_n \ell_\theta - \sup_{\theta\in\Theta_{K_1}} \mathbb{P}_n \ell_\theta \quad \text{et}$$

$$\sup_{\theta\in\Theta_{K_2}} \left((\mathbb{P}_n - P^\star) \frac{\ell_\theta - \ell^\star}{H(\theta)^{1/2}} \right)^2 \geq \sup_{\theta\in\Theta_{K_2}} \mathbb{P}_n \ell_\theta - \sup_{\theta\in\Theta_{K_1}} \mathbb{P}_n \ell_\theta .$$

En retour, l'hypothèse $P^\star$-Donsker initiale est à vérifier pour une classe de rapports de vraisemblances renormalisés de la forme $g_\theta = H(\theta)^{-1/2}\,(\,\ell_\theta - \ell^\star)$. Cela constitue une tâche plus difficile que la première. Nous avons pu mener à bien la vérification (fastidieuse) sur l'exemple **MGM**.

En résumé, nous étendons tous les résultats asymptotiques de consistance, qui tiennent sous des hypothèses faibles au regard de leur adéquation à des modèles variés, et au regard de celles requises par les études antérieures des procédures fondées sur le maximum de vraisemblance. En particulier, la consistance peut être assurée pour $\widehat{K}_n^{\mathrm{L}}$ sans connaissance *a priori* d'une borne sur $K^\star$. Le « tour à la Huber » permet un progrès spectaculaire en termes de condition associée portant sur la pénalité, mais au prix de la formulation d'une hypothèse techniquement difficile à vérifier.

Sous-estimation

Bahadur et al. (1980) proposent la version concise suivante du célèbre *lemme de Stein* :
Pour deux probabilités P et Q sur $(\Omega, \mathcal{A})$ et une suite $\{A_n\}$ d'ensembles mesurables,

$$\liminf_{n\to\infty} Q(A_n) > 0 \Longrightarrow \liminf_{n\to\infty} n^{-1} \log P(A_n) \geq -H(P\,|Q).$$

Ce lemme nous conduit à identifier la meilleure vitesse de sous-estimation, valable pour tout estimateur $\widetilde{K}_n$ de l'ordre qui ne le sur-estime pas presque sûrement. En effet, des choix adéquats de probabilités P, Q et de suites $\{A_n\}$ aboutissent à la minoration (Proposition 7.5.1)

$$\liminf_{n\to\infty} n^{-1} \log \mathrm{P}(\widetilde{K}_n < K^\star) \geq - \inf_{K<K^\star} H(\Pi_K \,|P^\star).$$

Dès lors, l'objectif est de majorer l'expression de gauche ci-dessus pour $\widetilde{K}_n$ égal à $\widehat{K}_n^{\mathrm{L}}$ ou $\widehat{K}_n^{\mathrm{G}}$. Plus le majorant sera proche de l'expression de droite ci-dessus, meilleure sera considérée la majoration.

Nous exprimons l'événement de sous-estimation en termes d'un événement concernant la mesure empirique $\mathbb{P}_n$. Nous souhaitons alors appliquer la majoration d'un théorème de Grandes Déviations de type Sanov pour la mesure empirique. Le problème revient alors à l'évaluation de l'infimum de la fonction de taux correspondante sur la fermeture d'un certain ensemble.

Le choix du théorème de Sanov classique sur $M_1(\mathcal{Z})$ pour la topologie rendant continues les $Q \mapsto Q\,f$ de $M_1(\mathcal{Z})$ dans $\mathbb{R}$ pour toute fonction mesurable bornée (la τ-topologie) impose des conditions drastiques sur les log-vraisemblances ℓ_θ ($\theta \in \Theta_K$, $K < K^\star$), qui doivent être bornées.

Il s'avère en fait (Schied 1998) que notre schéma de preuve ne peut aboutir pour un théorème de Sanov sur $M_1(\mathcal{Z})$ si les log-vraisemblances ℓ_θ n'admettent pas tous leur moments exponentiels (*i.e.* ne satisfont pas $P^\star \exp(a|\,\ell_\theta|) < \infty$ pour tout $a > 0$). C'est une condition moins contraignante que l'existence d'une borne, qui exclut toutefois notre modèle **MGM**.

Le cadre naturel dans lequel notre schéma de preuve peut être mené à bien implique l'ensemble $\mathcal{L}_\tau^*$ des formes linéaires sur la classe $\mathcal{L}_\tau$ des fonctions f qui admettent un moment exponentiel (*i.e.* pour lesquelles il existe $a > 0$ tel que $P^\star \exp(a|f|) < \infty$). On munit $\mathcal{L}_\tau^*$ de la topologie $\sigma(\mathcal{L}_\tau^*, \mathcal{L}_\tau)$ qui rend continues les $Q \mapsto Q\,f$ pour tout élément de $\mathcal{L}_\tau$. Or, Léonard et Najim (2000) ont prouvé récemment un tel théorème de Sanov (Théorème 6.2.4).*

Nous parvenons finalement à démontrer un premier résultat sur la sous-estimation (Théorème 7.5.3) selon lequel

$$\limsup_{n\to\infty} n^{-1} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} \vee \widehat{K}_n^{\mathrm{G}} < K^\star) < 0.$$

Brièvement, les hypothèses principales demandent *primo* que les log-vraisemblances ℓ_θ ($\theta \in \Theta_K$, $K < K^\star$) admettent un moment exponentiel (nous nous plaçons dans un cadre où le théorème de Sanov s'applique bien) ; *secundo* qu'elles soient uniformément (en $\theta \in \Theta_K$, $K < K^\star$) minorées et majorées par deux fonctions l et u dont la différence $(u - l)$ admet tous ses moments exponentiels ; *tertio* que pour tout compact C de $\mathcal{Z}$ et tout $K \leq K^\star$, l'ensemble $\{\,\ell_\theta \mathbb{1}\{C\} : \theta \in \Theta_K\}$ soit précompact dans la classe $\mathrm{C}^0(C, \|\cdot\|_\infty)$ des fonctions continues munie de la norme uniforme.

La nouveauté et l'intérêt de ce résultat résident dans la nature de l'approche et des hypothèses sur lesquelles il repose. Celles-ci sont simples au regard des hypothèses que requiert une preuve fondée sur un développement de rapport de vraisemblances (triple différentiabilité des log-vraisemblances, inversibilité de l'information de Fisher, existence de moments exponentiels pour le gradient et le supremum de la Hessienne pour la preuve de Haughton et Keribin de 2001). Elles sont simples aussi du seul fait que les exemples **HO** et **MGM** les satisfont simultanément.

En guise de conclusion, un dernier argument vient étayer notre illustration de la nouveauté et de l'intérêt de notre approche. Ainsi, une hypothèse de différentiabilité des fonctions $\theta \mapsto \ell_\theta$ ($\theta \in \Theta_K$, $K < K^\star$) de sorte que les gradients $\dot{\ell}_\theta$ admettent tous leurs moments exponentiels permet de prouver que, dans un modèle de régression exponentiel (du type **HO**), *la meilleure vitesse de sous-estimation est atteinte par* $\widehat{K}_n^{\mathrm{G}}$.

Ceci est un résultat inédit. Sa preuve met en jeu des projections au sens de la divergence de Kullback-Leibler et des arguments d'analyse convexe.

Sur-estimation

Cette fois, le lemme de Stein nous renseigne sur une borne supérieure pour la vitesse de sur-estimation valable pour tout estimateur $\widetilde{K}_n$ qui ne le sous-estime pas presque sûrement. Ainsi

*Nous remercions Stéphane Boucheron qui nous a appris l'existence de ce théorème de Sanov, ainsi que les auteurs Christian Léonard et Jamal Najim pour d'instructives discussions.

(Proposition 7.6.1)

$$\liminf_{n\to\infty} n^{-1} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} < K^\star) = \liminf_{n\to\infty} n^{-1} \log \mathrm{P}(\widehat{K}_n^{\mathrm{G}} < K^\star) = 0.$$

Là encore, nous exprimons l'événement de sur-estimation en termes d'un événement concernant la mesure empirique $\mathbb{P}_n$. Nous souhaitons appliquer, non plus un résultat de Grandes Déviations (en vertu de ce qui précède), mais un résultat de Moyennes Déviations pour la mesure empirique. Le problème glisse ainsi vers l'évaluation de l'infimum de la fonction de taux correspondante sur la fermeture d'un certain ensemble, visant à prouver sa stricte positivité.

Une discussion similaire à celle que nous avons tenue pour le cas de la sous-estimation écarte les résultats classiques de Moyennes Déviations, dont celui pour la τ-topologie (voir par exemple de Acosta 1994). Nous avons démontré un nouveau théorème de Moyennes Déviations dans l'esprit du théorème de Sanov de Léonard et Najim (2000) sur l'ensemble $M(\mathcal{Z})$ des mesures sur $\mathcal{Z}$ et pour la topologie $\sigma(M(\mathcal{Z}), \mathcal{L}_\tau)$ qui rend continues les $Q \mapsto Q f$ (tout $f \in \mathcal{L}_\tau$). Malheureusement, nous n'avons pas su démontrer la stricte positivité de l'infimum d'intérêt dans ce cas.

Nous concluons tout de même grâce à un résultat antérieur de Moyennes Déviations dû à Wu (1994). Les hypothèses suffisantes sont sans doute plus contraignantes que celles auxquelles nous aurions recours si nous parvenions à prouver la positivité du majorant produit par l'application de notre résultat de Moyennes Déviations. Nous interprétons la mesure empirique recentrée $(\mathbb{P}_n - P^\star)$ comme un élément de la classe $\ell^\infty(\mathcal{G})$ des fonctions réelles uniformément bornées sur $\mathcal{G}$, munie de la norme uniforme sur $\mathcal{G}$. Nous supposons que la classe $\mathcal{G}$ est $P^\star$-Donsker et qu'elle a une enveloppe admettant un moment exponentiel. Nous choisissons des pénalités de la forme $\mathrm{pen}(n, K) = v_n\, D(K)$, où $D \in \mathbb{R}^{\mathbb{N}}$ croît et $v_n = o(n)$, avec une condition technique supplémentaire qui empêche $\{v_n\}$ de trop se rapprocher de $\{n\}$. Enfin, la connaissance d'une borne *a priori* sur $K^\star$ est requise pour le cas de $\widehat{K}_n^{\mathrm{G}}$.

Dans un premier temps, $\mathcal{G}$ est l'ensemble des $g_\theta = (\ell_\theta - \ell^\star)$ $(\theta \in \Theta_K)$ pour $K = K^\star + 1$ ou $K = K_{\mathbf{max}}$ selon que l'on étudie $\widehat{K}_n^{\mathrm{L}}$ ou $\widehat{K}_n^{\mathrm{G}}$. Alors, si $n^{1/2} v_n^{-1} = o(1)$,

$$\limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}(\widetilde{K}_n > K^\star) < 0,$$

où $\widetilde{K}_n$ coïncide avec $\widehat{K}_n^{\mathrm{L}}$ ou $\widehat{K}_n^{\mathrm{G}}$ (Théorème 7.6.3). Ce résultat s'applique aux exemples **HO** et **MGM**.

Dans un second temps, $\mathcal{G}$ est l'ensemble des $g_\theta = H(\theta)^{-1/2}\,(\ell_\theta - \ell^\star)$ (tout $\theta \in \Theta_K$ tel que $H(\theta) > 0$) pour $K = K^\star + 1$ ou $K = K_{\mathbf{max}}$ selon que l'on étudie $\widehat{K}_n^{\mathrm{L}}$ ou $\widehat{K}_n^{\mathrm{G}}$. Alors, si $(\log n)\, v_n^{-1} = o(1)$,

$$\limsup_{n\to\infty} v_n^{-1} \log \mathrm{P}(\widetilde{K}_n > K^\star) < 0,$$

où $\widetilde{K}_n$ coïncide avec $\widehat{K}_n^{\mathrm{L}}$ ou $\widehat{K}_n^{\mathrm{G}}$ (Théorème 7.6.5). Ici, le « tour à la Huber » est de nouveau à l'origine du raffinement du résultat en termes de latitude sur le choix des pénalités. Le renforcement des hypothèses est conséquent. Les exemples **HO** et **MGM** ne rentrent pas dans le cadre imposé.

Pour conclure, notre approche du problème produit des résultats dont les hypothèses sont simples au regard de celles que requiert une preuve par développement de rapport de vraisemblances. Cette simplicité est aussi illustrée par la vérification simultanée des hypothèses par **HO** et **MGM**.

Pratique de la segmentation et agrégation

Nous résumons dans la présente section le contenu du Chapitre 3.

Arbres de régression CART

Une méthode prisée

Nous avons surtout discuté des thèmes croisés de la segmentation spatiale et de la sélection de modèle du point de vue théorique depuis que nous en avons motivé l'intérêt du point de vue applicatif pour le problème original de France Télécom R&D qui nous occupe.

Nous avons consacré une part importante de notre travail à l'aspect pratique. Nous l'avons abordé notamment *via* les arbres de régression CART.

CART (un acronyme de Classification And Regression Trees) est une méthode d'estimation non linéaire qui s'applique à la fois dans un cadre de classification et de régression. Son développement a débuté à la fin des années 70. L'ouvrage de référence (Breiman et al. 1984) en a fixé la forme. La popularité de CART est allée grandissante depuis. Elle connaît d'ailleurs un nouveau sursaut, du fait particulièrement de la diffusion et l'utilisation récentes des idées et des outils de *l'analyse exploratoire des données* (*Data Mining* en anglais – le lecteur pourra par exemple se référer à Hastie et al. 2001).

Le succès de CART est entre autres dû à sa simplicité algorithmique et à une interprétabilité aisée des estimateurs qu'elle produit. Ce dernier point est en effet souvent jugé très important.* Il l'est ainsi notamment par nos collègues de France Télécom R&D, nous l'avons évoqué dans la première section. C'est l'une des raisons qui nous a fait nous intéresser à cette méthode.

Soulignons enfin que CART est implémenté dans de nombreux logiciels statistiques commerciaux et qu'elle a été appliquée par exemple à la biochimie, la médecine, la météorologie, la reconnaissance de caractères (voir Breiman et al. 1984 pour les premiers exemples), ou encore plus récemment à l'imagerie (Chou et al. 1989) et à la prévision de pollution (Ghattas 1999).

Description succincte

Nous nous replaçons dans le cadre de travail de régression homoscédastique et hétéroscédastique des exemples **HO** et **HE**. Nous reprenons aussi les notions de segmentation introduites lors du commentaire sur le Chapitre 5.

CART peut être appréhendée comme une procédure combinée de minimisation de contraste (sous-optimale mais au demeurant simple et rapide) et de sélection de modèle dans ces cadres.

La sous-optimalité est due au fait que l'on minimise itérativement le contraste empirique *non pénalisé* en considérant des partitions emboîtées (*i.e.* de plus en plus fine à partir de la partition

*Certains auteurs (dont Breiman) s'élèvent cependant contre la pertinence d'un tel critère de jugement. Nous en discuterons brièvement ci-dessous.

naïve à un seul élément). On ne passe alors bien entendu pas toutes les partitions en revue. La collection des partitions emboîtées est naturellement décrite simplement par un *arbre* maximal binaire complet (tout nœud admet soit deux descendants, soit aucun – dans ce cas, le nœud est appelé *feuille*), ce qui justifie le vocabulaire forestier : *feuille, arbres, élagage, forêt.*

L'*élagage* correspond à l'inclusion dans CART d'une méthode de sélection de modèle par pénalisation du contraste empirique à minimiser selon un principe surprenant. Choisissons des fonctions de pénalité de la forme $\text{pen}(n, K) = v\,K$, où v est un paramètre positif à régler et K dénote la cardinalité d'une partition (*i.e.* le nombre de feuilles de l'arbre associé). On peut en effet prouver (Théorème 3.3.6, dû à Breiman et al.) l'existence d'une suite de partitions emboîtées (une *forêt*) extraites par élagage de la partition maximale et telle que, quel que soit le choix de v, le contraste pénalisé

$$n^{-1}\sum_{i=1}^{n}\gamma(f, Z_i) + v\,K$$

soit uniquement minimalisé (lorsque l'on considère *tous* les arbres élagués de l'arbre maximal) par un des éléments de la suite.

On peut noter que cette expression de pénalité recoupe celle que nous avons étudiée dans le Chapitre 5 et son étude théorique de la segmentation spatiale et de la sélection de modèle. Le paramètre v est substitué au facteur v_n dépendant du nombre d'observations. Il doit être calibré pour un jeu d'observations donné.

L'estimateur final est un arbre choisi par validation au sein de la forêt construite par élagage à partir de l'arbre maximal.

Dans le Chapitre 3, nous proposons :

- Une présentation fouillée de la procédure originale. Celle-ci correspond au cadre homoscédastique et au contraste de moindres carrés de l'exemple **HO**.
- Une adaptation de la procédure originale afin de la rendre pertinente dans le cadre hétéroscédastique de l'exemple **HE**, avec son contraste associé.
  Ceci est une contribution inédite. Elle offre un élément de réponse à une remarque des auteurs de (Breiman et al. 1984) sensibilisant les utilisateurs de CART dans un modèle hétéroscédastique aux effets indésirables susceptibles d'apparaître.

Il convient finalement de préciser que la procédure CART est *instable*, *i.e.* heurisitiquement qu'une faible perturbation des données d'entrée peut avoir une répercussion importante sur la sortie de l'algorithme. C'est un constat que peut faire tout utilisateur et que Breiman (1996b) a exploré méthodiquement. Pour une approche théorique de l'instabilité, le lecteur pourra consulter (Bousquet and Elisseeff 2002).

Cet aspect au premier abord fâcheux est exploité dans la prochaine sous-section.

Agrégation par Bagging ou Boosting : modélisation par boîte noire*

Heuristique de l'agrégation

Le titre de cette section est en forme de clin d'œil au récent article stimulant de Breiman (2001). Dans ce dernier, l'auteur confronte ce qui lui apparaît comme deux cultures statistiques

*Nous souhaitons remercier Gilles Blanchard et Servane Gey pour de nombreuses discussions enrichissantes sur ces deux procédures.

antagonistes, à savoir celle de la modélisation de données *versus* celle de la modélisation algorithmique.

Ainsi, étant donné un problème de régression et un jeu de données lui correspondant, le processus original (« naturel ») peut être envisagé comme une *boîte noire* qui associe à une donnée en entrée, une réponse en sortie. L'objectif statistique est double. Il s'agit *primo* d'informer, *i.e.* d'extraire du jeu des observations des éléments explicatifs de la façon dont la boîte noire agit ; *secundo* de prédire, *i.e.* de reproduire artificiellement le fonctionnement de la boîte noire. Selon l'auteur, on rencontre deux principales approches :

- la modélisation de donnée, où un modèle statistique est proposé pour la boîte noire, ajusté au vue des observations puis testé ;
- la modélisation algorithmique, où une boîte artificielle complexe est construite et testée au vue des observations.

Dans le premier cas ci-dessus, une bonne interprétabilité est un critère important de satisfaction. Ce n'est plus vrai pour le second, où prime l'intérêt apporté à la qualité de l'information fournie.

La procédure CART participe naturellement à la seconde approche. Elle y joue en vertu de son instabilité le rôle de prédicteur faible (*weak learner*), *i.e.* d'élément basique de construction de la boîte artificielle, *i.e.* du prédicteur final. Le principe schématique consiste à rééchantillonner à plusieurs reprises l'ensemble d'apprentissage (une partie du jeu d'observation réservé à la fabrication de la boîte artificielle, le reste étant réservé à la validation), à construire systématiquement un prédicteur faible sur les jeux produits, pour finalement agréger les prédicteurs faibles en un prédicteur fort final.

Heuristiquement, le prédicteur fort final est bien plus performant (sur des critères statistiques) que chacun des prédicteurs faibles qui participent à son élaboration.

Bagging et Boosting

Nous pouvons distinguer deux familles de procédures selon la nature du rééchantillonnage impliqué :

uniforme : les procédures P&C (*Pertub and Combine*) correspondent aux cas où le rééchantillonnage de l'ensemble d'apprentissage est *uniforme*. Le Bagging (*Bootstrap aggregating*) en est un exemple. Il a été proposé par Breiman (1996a) dans un cadre de classification, puis prolongé au cas de la régression. Nous renvoyons à (Bühlmann and Yu 2002a) pour une bibliographie complète et des résultats théoriques concernant le Bagging.

Soulignons que pour le Bagging, l'agrégation consiste heuristiquement à moyenner les prédicteurs faibles.

adaptatif : les procédures Arcing (*Adaptively resampling and combining*), initiées par Breiman (1998) dans un cadre de classification, correspondent aux cas où le rééchantillonnage de l'ensemble d'apprentissage est *adaptatif*. Nous entendons par là que la loi du tirage à l'itération $(\ell + 1)$ dépend des performances du prédicteur agrégé à l'itération ℓ. La mise à jour de la loi de tirage tend à favoriser les exemples mal estimés par le prédicteur agrégé à l'étape ℓ.

Le Boosting est un exemple de procédure Arcing dû à (Freund and Schapire 1996) dans un cadre de classification (avec une comparaison au Bagging). Le Boosting d'arbres de régression peut suivre une approche de gradient (Bühlmann and Yu 2002b) ou respecter

fidèlement l'algorithme original suivant (Drucker 1997). C'est à cette approche que nous nous sommes attachés particulièrement.

Soulignons que pour le Boosting, l'agrégation consiste heuristiquement à prendre une médiane pondérée des prédicteurs faibles.

Le Bagging joue traditionnellement le rôle d'élément de comparaison avec d'autres méthodes d'agrégation, dont le Boosting.

Remarque 3. Ces procédures d'agrégation sont issues des travaux produits par les membres de la communauté de l'intelligence artificielle. La communauté statisticienne s'intéresse à celles-ci depuis un dizaine d'années. Des outils tels que les inégalités de concentration permettent de comprendre pourquoi ces méthodes jouissent des qualités *surprenantes* qu'elles exhibent. *Il faut en effet bien prendre la mesure de combien une méthode fondée sur une accumulation d'estimations successives et leur mise en commun va à l'encontre du classique principe statistique fondamental de l'économie d'estimation.*

Ces procédures d'agrégation ne sont pas étrangères à celles de sélection de modèle, qui pourtant peuvent être interprétées comme des techniques économiques. En effet, il faut veiller à ce que les prédicteurs faibles (soit les arbres de régression CART pour ce qui nous concerne) ne soient ni trop naïfs, ni trop sophistiqués, pour échapper aux écueils du sous-ajustement ou du sur-ajustement. Les méthodes de validation fournissent notamment des éléments d'appréciation relatifs à ces deux écueils.

Dans le Chapitre 3, nous proposons :

- Une présentation fouillée des procédures originales de Bagging et de Boosting. Celles-ci correspondent au cadre de la régression homoscédastique et au contaste des moindres carrés de l'exemple **HO**.
- Une adaptation des procédures originales afin de les rendre pertinentes dans le cadre hétéroscédastique de l'exemple **HE**, avec son contraste associé.
  Ceci est une contribution inédite. Nous la mettons en œuvre dans la partie applicative de la thèse consacrée au problème initial posé par France Tél'ecom R&D.
- Une analyse originale de l'importance des variables explicatives de régressions qui peut conduire à la réduction de dimension pour ce type de problèmes.

Eléments de raffinement de localisation de trafic

Nous résumons dans cette dernière section le contenu des Prélude, Chapitre 1, Interlude 2 et Chapitre 4. Ceux-ci sont dédiés directement au problème original proposé par France Télécom R&D.

Le Prélude introduit des rudiments de technologie de téléphonie mobile et le problème original proposé par France Télécom R&D.

Nous présentons dans le Chapitre 1 les données issues de France Télécom R&D et Orange que nous sommes parvenus à obtenir. Elles ont conditionné toute notre approche du problème.

L'Interlude 2 est consacré à la description de la base de données explicatives de nature socio-démographique et culturelle que nous avons construite *ad hoc* à partir de trois bases de données classiques entretenues par l'INSEE†, auprès de qui nous les avons acquises.

Enfin, nous présentons dans le Chapitre 4 l'application de la méthode élaborée soigneusement aux diverses données disponibles. Notre intention est d'une part d'en illustrer la flexibilité ; de l'autre d'offrir une vision transversale au lecteur des qualités et performances de la dite méthode. Pour cela, la présentation des résultats se fait de façon systématique d'un exemple à l'autre. Le tout est par ailleurs commenté largement.

Nous construisons ainsi des prédicteurs (forts, au sens d'agrégés de prédicteurs faibles CART) du trafic qui permettent de le localiser. Nous extrayons aussi un certain nombre de variables explicatives les plus pertinentes, conformément au souhait de nos collègues de France Télécom R&D. Nous renvoyons le lecteur aux chapitres eux-mêmes pour de plus amples détails.

La mise en œuvre de notre méthode a supposé un important travail d'implémentation sous `Perl` (préparation des données, très volumineuses) et `Matlab` (programmation *in extenso* de la procédure, dont la version inédite de CART adaptée au modèle **HE**).

Prélude

Résumé

Le lecteur découvrira dans ce prélude une présentation de l'entreprise qui a été à l'origine de ce travail de thèse, et plus particulièrement du laboratoire qui m'y a accueilli. Il trouvera aussi une introduction succincte aux techniques de communication mobile telles que développées, quotidiennement entretenues et innovées en son sein. Le problème initial sera introduit. Le prélude se concluera par un résumé de son contenu en version anglaise.

Abstract

We shall briefly introduce to the reader in the prélude hereafter the firm from which stems this thesis work. A more precise presentation of the laboratory that welcome us will lead to a concise yet hopefully not inaccurate summary of the basis of roaming and mobile telecommunication. Our work indeed copes with questions that arose from that technological field as engineers and researchers both developed and improved the current techniques, even anticipated the forthcoming ones. The original problem will be stated. The prélude will close with an english summary.

Au menu

France Télécom R&D 47
Le laboratoire d'accueil 48
Un bref aperçu du réseau mobile Orange 49
Ce qu'il faut retenir 52
Introduction au problème initial 53
English summary 54

France Télécom R&D*

France Télécom R&D est le centre de recherche du grand groupe français de télécommunications France Télécom. Il est le fruit de l'évolution du Centre National d'Etudes des Télécommunications (CNET), créé en 1944 avec le statut de centre de recherche rattaché au Ministère des PTT. En 1991, France Télécom est devenu Etablissement Autonome de droit public et les missions du CNET se sont recentrées sur les métiers de France Télécom. Avec le changement de statut de France Télécom de 1997, qui devient une Société Anonyme, le CNET a été rattaché à la Branche Développement de l'opérateur. Le 1er mars 2000, le centre quinquagénaire change de nom et embrasse celui de France Télécom R&D. Ce changement illustre la confirmation de rôle de centre de recherche et développement de l'ensemble des branches opérationnelles de France Télécom, avec des objectifs de recherche à long terme, mais aussi à plus court terme selon les besoins des unités opérationnelles. Elle est aussi le fruit de l'ouverture du marché français des télécommunications.

L'action de France Télécom R&D s'articule autour de trois grands thèmes :
- amélioration quotidienne des produits et services existants,
- imagination et réalisation d'innovations,
- exploration créatrice libérée de toutes contraintes

et se répartit en six grands domaines d'exploration :
- nouveaux usages : anticipation des usages personnels et professionels des clients ;
- mobilité : course technologique à la mobilité et à la variété et la qualité des services offerts (voix, image, vidéo) ;
- architectures de réseau : intervention sur les infrastructures de tous les réseaux (fixe, mobile, internet) pour optimiser les flux, enrichir les services et réduire les coûts ;
- internet : fournisseur d'accès, d'hébergement, de contenus ; commerce électronique, paiement sécurisé, fusion de l'audiovisuel et d'internet ;
- convergence fixe/mobile/internet : recherche sur des réseaux plus fluides, plus sûrs et plus compatibles ;
- réseaux de transport et d'accès : gain en débit, travaux sur différents supports de transmission d'information.

France Télécom R&D compte neuf unités de recherche. Le siège se trouve en région parisienne, à Issy-les-Moulineaux, où travaillent la plupart des personnes avec qui j'ai collaboré. On trouve aussi entres autres un centre à Lannion et un à Belfort, dont sont issus d'autres ingénieurs-chercheurs que j'ai rencontrés au cours de ma thèse. Je les remercie ici encore pour leur contribution.

*Cette section s'inspire de la brochure [France Télécom, loin devant et proche de vous].

Le laboratoire d'accueil

J'ai collaboré avec un laboratoire de la Direction des services Mobiles et systèmes Radio (*DMR* dans la suite). Cette direction (parmi les sept directions de recherche et développement que compte France Télécom R&D) se consacre au développement et à l'innovation. Elle honore de nombreux contrats, notamment avec les branches Entreprises et Réseaux de France Télécom et bien sûr avec la branche Mobiles *Orange*, du nom de l'opérateur britannique racheté par France Télécom en 2000. L'opérateur Orange domine les marchés français et britanniques et, fort de ses presque 40 millions de clients début 2002, occupe la place de numéro deux européen des mobiles.

Les prestations de la DMR concernent (la liste n'est pas exhaustive) les domaines des :
- technologie de radiodiffusion :
  - développement du savoir-faire technique de haut niveau sur l'électromagnétisme, les antennes et la propagation ;
  - développement et expérimentation de télé-radiodiffusion sonore numérique et télévision numérique ;
  - étude des effets biologiques des ondes électromagnétiques ;
  - étude technique de l'accès radio haut débit à internet, des risques et opportunités inhérents ;
- traitement du signal :
  - évaluation des techniques de transmission radio (réseau mobile, réseau de proximité) ;
  - recherche sur les techniques avancées de transmission radio ;
- réseaux, systèmes radio et outils associés :
  - coopération des réseaux physiques avec les réseaux sans fils ;
  - élaboration et développement d'outils de conception et de dimensionnement de réseaux ;
  - méthode d'ingénierie d'optimisation de la capacité des systèmes radio et de la qualité des services offerts ;
  - méthodes et outils d'observation des divers trafics dans les réseaux ;
  - modélisations et algorithmes d'optimisation de paramétrage des ressources radio ;
- services et réseaux à satellites :
  - diversification de l'offre de services sur réseaux à satellites (par exemple l'accès IP/satellite et le commerce électronique) ;
  - études techniques destinées à évaluer les projets à satellites multimedia mobiles ;
- services et réseaux pour les mobiles :
  - développement de la troisième génération de services et réseaux mobiles (dont l'Universal Mobile Telecommunications System, plus connu sous l'acronyme *UMTS*) et des technologies multimedia ;
  - développement du réseau et offre de services pour le groupe Orange ;
  - étude de l'offre de services de mobilité dans le cadre de la convergence fixe/mobile/internet (comme par exemple l'instauration d'un numéro unique d'appel pour ces trois types de communication).

C'est au laboratoire Interface radio et Ingénierie pour réseaux Mobiles (appelé *IIM* désormais) que j'ai été accueilli. Ses missions portent sur les réseaux mobiles et sur les réseaux d'accès mobiles aux réseaux fixes.

Un bref aperçu du réseau mobile Orange

Le réseau mobile Orange fonctionne suivant le système Global System for Mobile communications (soit *GSM*) enrichi du service General Packet Radio Service (soit *GPRS*). Nous allons tâcher ici de proposer une vue d'ensemble concise et fonctionnelle du système de communication mobile qui est à l'origine du sujet de cette thèse. Celle-ci se fera naturellement au prix de quelques simplifications au sujet desquelles nous prions les spécialistes d'être cléments. Le lecteur intéressé par les détails pourra se référer à l'ouvrage (Lagrange et al. 1999) dont cette présentation est très largement inspirée.

GSM : un système numérique cellulaire

Le GSM est la première norme de téléphonie radiomobile de seconde génération, *i.e.* pleinement numérique par opposition à la première génération analogique. C'est néanmoins dans le cadre des systèmes de première génération que le concept cellulaire a vu le jour pour finalement s'imposer comme procédé d'exploitation efficace de la ressource radio.

La finalité d'un système de téléphonie mobile est de permettre l'accès au réseau téléphonique sur un territoire étendu depuis un terminal portatif que nous appelerons désormais *mobile*. Il met en jeu une liaison radioélectrique entre le mobile et le réseau dans plusieurs bandes de fréquences, dont nous ne retiendrons que les 900 MHz et 1800 MHz.

L'accès au réseau depuis un mobile est conditionnel à une qualité suffisante de la liaison radio et par conséquent à la puissance et au nombre des émetteurs et de leurs localisations. L'opérateur dispose donc sur le territoire un ensemble de stations de base de sorte que la couverture totale par les zones de desserte des stations de base soit la plus complète possible dans un souci de limitation du nombre de stations.

L'architecture du réseau

Le réseau GSM doit permettre des communications entre mobiles et téléphones fixes abonnés du Réseau Téléphonique Commuté (soit *RTC*). De tels échanges nécessitent l'existence d'interfaces entre les deux réseaux, appelés *commutateurs*. De fait, on peut diviser le réseau GSM en trois sous-ensembles :

- le sous-système radio (soit *BSS* pour Base Station Sub-system) ;
- le sous-système d'acheminement, ou réseau fixe (soit *NSS* pour Network Sub-System) ;
- le sous-système d'exploitation et de maintenance (soit *OSS* pour Operation Sub-System).

L'architecture de réseau BSS/NSS est représentée dans la Figure 2.

Le BSS assure les transmissions radioélectriques et gère la ressource radio. C'est lui qui offre la posibilité de se déplacer au cours d'une communication, *i.e.* c'est lui qui permet la mobilité. Cette notion de mobilité implique la fonction de continuité de service alors que l'utilisateur se meut. Il peut être nécessaire de changer la station de base avec laquelle le mobile est en relations tout en maintenant la communication : c'est le transfert inter-cellulaire, ou *handover*.

Le NSS comprend l'ensemble des fonctions nécessaires à l'établissement des appels et à l'itinérance (*i.e.* à la possibilité de téléphoner de n'importe où, pas de pouvoir se déplacer en cours de communication).

L'OSS permet à l'opérateur d'administrer son réseau.

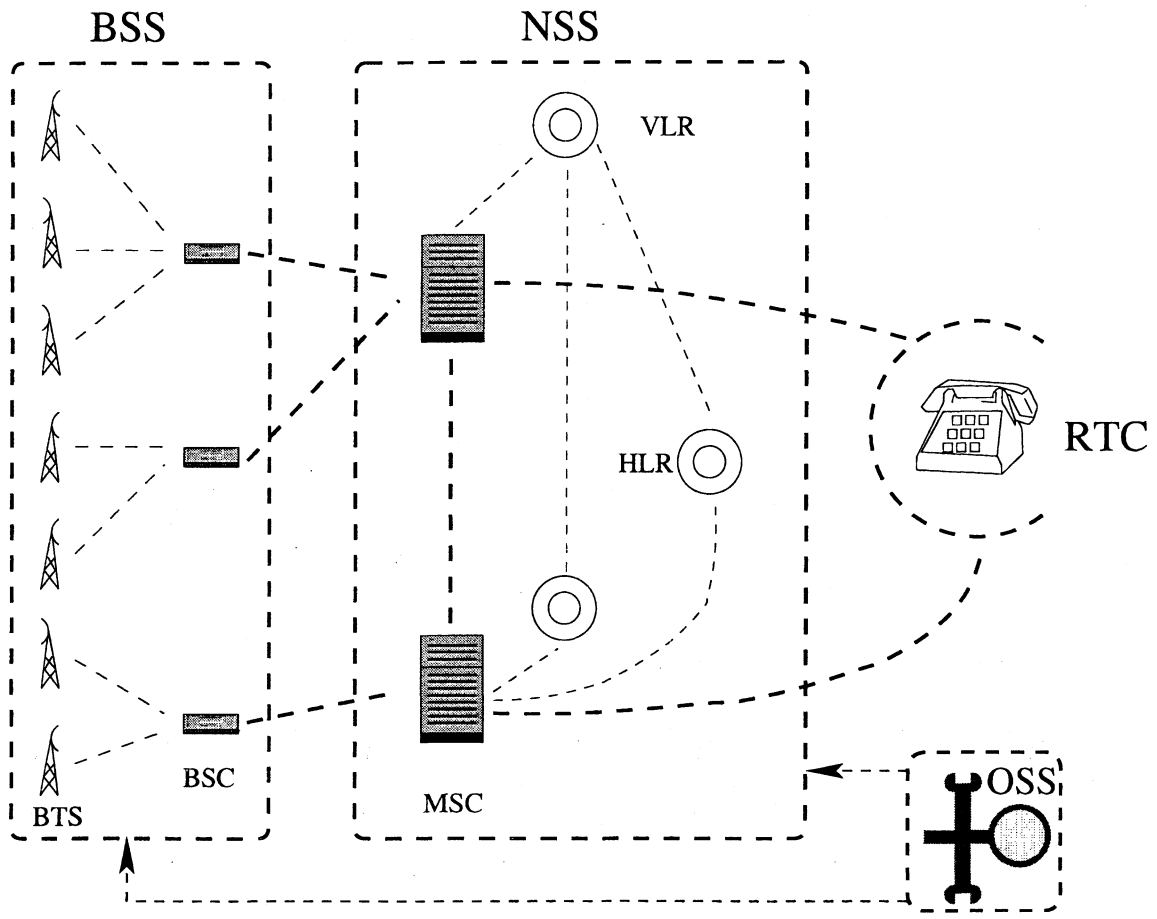


Figure 2 – Architecture de réseau.

Description du BSS

A un niveau de description moins grossier qu'au paragraphe précédent, le BSS comprend :
- les Base Transceiver Stations (désormais *BTS*) qui sont des émetteurs-récepteurs ;
- les Base Station Controllers (soit *BSC*) qui contrôlent un ensemble de BTS et permettent une première concentration des circuits.

Description du BSS : les BTS

Paris compte en 2002 environ deux milliers de BTS (pour un peu plus d'une centaine fin 1994). Chaque BTS est composée de plusieurs émetteurs-récepteurs appelés *TRX*, généralement de un à quatre en zone urbaine. Un TRX peut écouler entre 14 et 28 communications simultanées. La BTS a la charge de la transmission radio et procède aussi à des mesures qui, une fois transmises au BSC avec laquelle la BTS est en relation, permettent de s'assurer qu'une communication en cours se déroule correctement.

Un *site* est un endroit où sont regroupées plusieurs BTS, dont le nombre est généralement entre deux et quatre.

Cellules, microcellules et macrocellules

La zone de desserte d'une BTS est appelée *cellule.*

Il existe deux types de cellules, indépendants de la bande de fréquences 900 MHz ou 1800 MHz : les *microcellules* (à superficie réduite*) et les *macrocellules* (à plus grande superficie). Les microcellules servent à couvrir très finement une certaine zone – généralement parce que son trafic moyen est élevé (pensez à une gare par exemple). De telles cellules sont obtenues en installant des stations de base au-dessous du niveau des toits, voyez à titre illustratif la Figure 3. Il serait bien sûr trop coûteux de couvrir une ville uniquement à partir de microcellules. Aussi, l'opérateur conserve une couverture avec des cellules classiques (*i.e.*, par opposition, dont les stations de base se trouvent au-dessus des toits). Le réseau est alors constitué de deux couches, l'une microcellulaire et l'autre macrocellulaire ; on parle de réseau hiérarchique. Quant au fonctionnement concret, il faut savoir que l'opérateur peut paramétrer son système de sorte que les mobiles à déplacement rapide ne se connectent pas aux cellules de la couche micro mais plutôt à celles de la couche macro afin, d'une part de réduire le nombre de transferts intercellulaires et d'autre part, de ne pas charger inutilement les microcellules, dont la vocation est d'écouler un trafic local peu mobile.

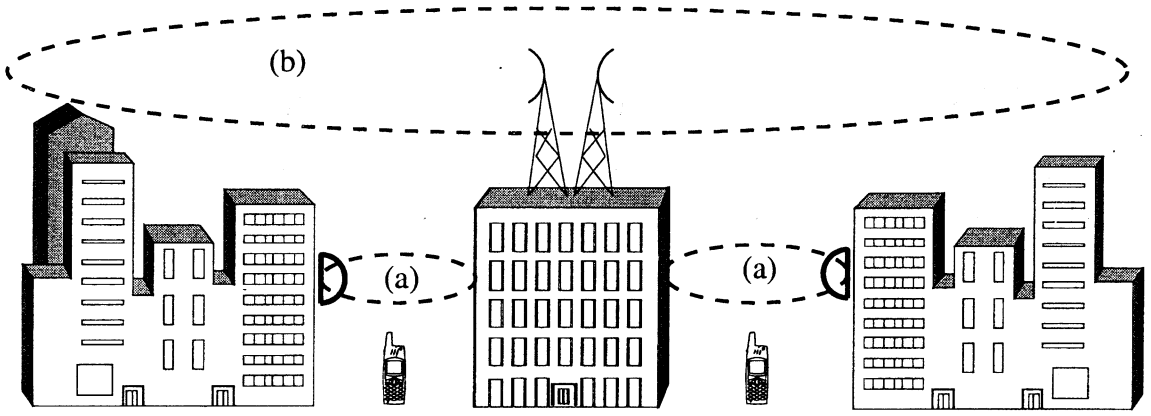


Figure 3 – Microcellules (a) et macrocellule (b).

Description du BSS : les BSC

Les BSC, contrôleurs des stations de base, constituent « l'organe intelligent » du BSS. Un BSC a pour fonction principale de gérer la ressource radio : il commande les allocations des canaux, utilise les mesures effectuées par les BTS évoquées plus haut pour contrôler les puissances d'émission du mobile et/ou de la BTS, il prend la décision de l'exécution des handover. Il est par ailleurs relié au NSS, comme expliqué plus bas.

On compte aujourd'hui une cinquantaine de BSC sur Paris.

Description du NSS

Rentrons maintenant davantage dans les détails du NSS. On observe qu'il est constitué de commutateurs et de bases de données, avec :

*Une description statistique des superficies sera traitée dans le Chapitre 1.

- les Mobile-services Switching Centers (*MSC* pour la suite), qui sont des commutateurs mobiles associés en général aux bases de données Visitor Location Register (*VLR*) ;
- le Home Location Register (*HLR*), base de données de localisation et de caractérisation des abonnés.

Une demi-douzaine de MSC prennent en charge la téléphonie mobile sur Paris. Un MSC endosse principalement deux rôles : il gère d'une part l'établissement des communications entre un mobile et un autre MSC et assure d'autre part le lien entre le réseau mobile et le réseau fixe RTC. Il peut par ailleurs dialoguer avec les VLR et le HLR pour prendre en compte les informations qu'ils détiennent sur les usagers (données d'abonnement, dernier MSC où le mobile s'est manifesté, par exemple).

Un ensemble colossal de données d'exploitation instantanée transite par les MSC. Le système OSS permet d'y avoir accès ponctuellement pour en extraire des données brutes : une telle extraction est gérée par l'interface Cigale et produit des relevés Cigale que nous présentons de façon approfondie dans le Chapitre 1. *Le système OSS fournit aussi des rapports prétraités aux contenus variés, parmi lesquels les Compte-Rendus d'Appels et les relevés HC2 que nous présentons aussi dans le Chapitre* 1. *Ces trois jeux de données constituent l'ensemble de nos observations du trafic.*

GPRS : un pas vers l'UMTS

Le GPRS représente une évolution majeure du GSM. Il consiste grossièrement à permettre des transmissions de données par paquets sur la voie radio, libérant ainsi de la ressource radio : un mobile peut être susceptible de recevoir et d'émettre des données à tout moment sans que le réseau lui alloue un canal dédié. Cette modification ouvre la voie aux applications mobiles multimédia, amorçant ainsi la transition vers le système de troisième génération.

Cette nouvelle génération prendra sous peu forme *via* l'UMTS. Son aspect technique excitant et ardu qui est associé à l'expression « troisième génération » s'est effacé totalement au profit de la seule notion de *services*, véritable mot-clef, suite à une campagne de publicité ayant changé totalement d'approche. L'UMTS assurera une télécommunication mobile à haut débit grâce aux réseaux fixe, mobile et satellite, qui permettra d'écouler des contenus variés, de la traditionnelle voix aux transfert de données (par exemple données audiovisuelles, messagerie électronique, internet) et commerce électronique.

Ce qu'il faut retenir*

France Télécom R&D† est la branche Recherche et Développement du groupe français France Télécom. Cette branche honore de nombreux contrats auprès d'Orange†, branche téléphonie mobile du groupe France Télécom et premier opérateur britannique et français, second européen.

L'une des vocations de France Télécom R&D est d'innover en matière de mobilité† (course technologique à la mobilité et à la variété et la qualité des services offerts : voix, image, vidéo) et de l'architecture de réseau (intervention sur les infrastructures des réseaux fixe, mobile, internet pour optimiser les flux, enrichir les services et réduire les coûts). Le laboratoire IIM† de la direction DMR† est surtout dédié aux réseaux mobiles et à l'accès mobile au réseau fixe RTC†.

*Le lecteur pourra consulter le Glossaire en Annexe pour trouver une définition des acronymes et mots-clefs indiqués par le signe †.

C'est au titre de la course technologique à la mobilité et de la réduction des coûts sur les infrastructures que le laboratoire IIM m'a accueilli pour cette thèse.

Il est nécessaire, pour bien prendre la mesure du problème initial qui m'a été posé et auquel j'ai consacré ces trois ans, de maîtriser un b-a-ba de la technologie de téléphonie mobile telle que déployée par Orange. Il s'agit de la technologie GSM†-GPRS† de téléphonie mobile de seconde génération, *i.e.* numérique et cellulaire, assurant de surcroît la transmission de données par paquets sur la voie radio et à ce titre précurseur de la technologie de troisième génération (dont l'UMTS† sera la norme en Europe). Ce réseau permet de téléphoner de façon itinérante† et mobile†.

Le réseau peut être divisé en trois sous-ensembles (voir la Figure 2) :
- le sous-système radio BSS†, composé de BTS† et de BSC† ;
- le sous-système d'acheminement NSS†, composé de MSC†, VLR† et HLR† ;
- le sous-système d'exploitation et de maintenance OSS†.

Une BTS est composée d'élémetteurs-récepteurs élémentaires appelés TRX†. La zone de desserte d'une BTS est une cellule†. Il en existe de deux types : les macrocellules† et les microcellules†. Le MSC permet d'extraire des informations très diverses sur l'état du système à chaque instant et à volonté (théoriquement du moins, car une telle extraction est assez contraignante à réaliser).

Introduction au problème initial

Le problème initial était formulé comme suit :

> Proposition d'une méthode de raffinement de la localisation du trafic téléphonie mobile en zone urbaine.

Il faut entendre ici *localisation fine* par opposition à la *localisation cellulaire* à laquelle l'opérateur accède grâce au système d'exploitation OSS. Il faut aussi souligner que la localisation du *trafic* diffère de la localisation *d'un mobile* en cours de communication (sujet par ailleurs largement étudié car crucial en termes d'optimisation pour la norme UMTS).

On peut avancer trois motivations qui apportent un éclairage sur l'intérêt de la question soulevée :
- Accéder à une *connaissance qualitative et quantitative plus précise du trafic local* qui ne soit pas fondée exclusivement sur les appréciations techniques et la connaissance intime du réseau local qu'en a un ingénieur-exploitant.
- Aider à la *densification* d'un réseau existant, *i.e.* au positionnement et au paramétrage de nouvelles BTS.
- Aider à la *planification* d'un nouveau réseau.

A quoi on peut ajouter, étant donné les éléments de réponse que nous avons donnés au cours de notre étude, l'Aide à la *prospective* pour les ingénieurs de la branche prospective commerciale d'Orange.

Nous abordons cette question par une approche *statistique*. Cela supose que l'on ait accès à des *données* et que l'on élabore un *modèle*.

English summary

Light prerequisites

France Télécom R&D, french telecommunications laboratory, brings its worldwide known savoir faire to the french France Télécom Group. In particular, France Télécom R&D regularly completes contracts for Orange, the wireless telecommunications company of the same group.

Two of the six major fields of research of France Télécom R&D are concerned with *mobility* (technological race to enhance mobility ant to develop voice, data, images, video services) and with *network architectures* (optimization of flows, improvement of services and cut of costs by upgraduating the network infrastructure of fixed, wireless, internet telecommunications). The IIM (Interface radio et Ingénierie pour réseaux mobiles) laboratory of the DMR (Direction des services Mobiles et systèmes Radio) direction focuses on mobile networks and on the access of mobile phones to the fixed network RTC (Réseau Téléphonique Commuté). I have been working partially in the IIM laboratory for the last three years.

In order to better understand the original problem I was proposed, one should have the following light prerequisites on the wireless telecommunications technology developped by Orange, namely the GSM–GPRS norm (Global System for Mobile communications–General Packet Radio Service). This norm is a second generation technology, *i.e.* both numerical and cellular. It is actually a second and half technology, since the GPRS system allows packet transmission of data *via* radio channel, a first step in the direction of the third generation technology (as defined *e.g.* in the UMTS norm that will equip Europe in a near future – UMTS is an acronym for Universal Mobile Telecommunications System).

The GSM-GPRS network makes roaming (roughly: on can call from everywhere) and mobile (roughly: one can move during a call) telecommunications possible. This network is divided into three subsystems (see Figure 2):

- the radio subsystem BSS (Base Station Sub-system), composed of BTS (Base Transceiver Station) and BSC (Base Station Controller) – a BTS should be seen as a collection of antennas (about two thousands BTS for Paris and one to four TRX for each BTS) ;
- the routing subsystem NSS (Network Sub-System), composed of MSC (Mobile-services Switching Center), VLR and HLR (respectively Visitor/Home Location Register);
- the exploitation and maintenance subsystem OSS (Operation Sub-System).

A BTS covers a zone called "cell". There exist two kinds of them: macrocell and microcell (according to their area – which depends on the heigth of the site where the corresponding BTS is installed, see Figure 3). The MSC provides various informations on the instantaneous state of the whole subsystem BSS through days.

The original problem

The original problem presented in its original form is:

> Elaboration of a method providing refinement of the localization of the mobile telecommunication traffic in urban area.

To localize the traffic actually means "to localize the traffic on a scale thinner than the natural cellular scale for which the MSC provide observations". One has to emphasize that localizing

traffic differs from localizing *a particular mobile in communication* (this is another important subject related to problems of optimization for the UMTS norm).

There are at least three main interests in connection with the original question we are asked:

- Access to a more comprehensive, both qualitative and quantitative, knowledge of the local traffic that is not exclusively based on technical considerations and intimate knowledge of the local network of an engineer-operator.
- Provide help for the densification of an already existing network, *i.e.* for the installation and parametrization of new BTS.
- Provide help for the design of a brand new network.

Furthermore, the method we propose to try to solve the original problem could possibly be of some help for prospective projection of the future traffic.

Our point of view is of statistical nature. Thus, it requires *data* and *model* (even slight).

1

Les données France Télécom

Résumé

Notre approche du problème est de nature statistique. Nous présentons dans ce chapitre les données issues de France Télécom qui jouent un rôle dans notre étude. Elles peuvent être regroupées en deux classes bien distinctes : les données de trafic et les données de recouvrement cellulaire. Nous détaillons les contenus et les procédures d'extraction pour chacune des bases de données. Cela débouche sur une meilleure compréhension des enjeux du problème initial et des réponses que l'on peut y apporter. Les points les plus importants du chapitre sont finalement résumés en anglais.

Abstract

We tackle the problem introduced in the Prélude with a statistical approach. This chapter is dedicated to the presentation of the datasets provided by France Télécom. Two classes arise: the class of traffic data and the class of cellular covering data. The description yields better understanding of the problem at stake. It also points out what kind of solutions might be elaborated. The chapter is finally summarized in english.

Au menu

1.1. Introduction **59**
1.2. Les données de trafic **59**
1.2.1. Les Comptes-Rendus d'Appels 60
1.2.2. Les relevés Cigale 60
1.2.3. Les relevés HC2 64
1.2.4. Quelles données pour quelle localisation? 65
1.3. Ebauche d'une étude statistique des durées d'appel **68**
1.4. Les données de recouvrement cellulaire **70**
1.4.1. Emission, affaiblissement, réception 70
1.4.2. Méthodologie pour l'évaluation des cellules 71
1.4.3. Brève description quantitative des recouvrements cellulaires 76
1.5. Méthodes de localisation de trafic **77**
1.5.1. Une méthode antérieure de localisation par triangulation 77
1.5.2. Des données de trafic agrégées 78
1.5.3. Heuristique pour une méthode originale 79
1.5.4. Mise en garde 80
1.5.5. Esquisse de la méthode originale 80
1.6. Ce qu'il faut retenir **81**
1.7. English summary **83**

1.1. Introduction

Nous avons introduit dans le Prélude le problème initial que nous a posé France Télécom. Il s'agit de raffiner la localisation du trafic téléphonique mobile en zone urbaine. Nous nous penchons particulièrement sur Paris. Nous consacrons ce chapitre à la présentation des données issues de France Télécom à partir desquelles nous allons tâcher d'élaborer une réponse au problème initial.

On distingue naturellement deux classes de jeux de données :

- les jeux de données de trafic, qui participent chacun à la description empirique du trafic téléphonique mobile ;
- les jeux de données de recouvrement cellulaire, qui tentent de rendre compte du système complexe de desserte de Paris par les cellules du réseau.

La première d'entre ces classes est présentée dans la Section 1.2, la seconde dans la Section 1.4. La section intermédiaire est dédiée à une étude statistique des durées d'appel.

Toutes les manipulations de données (prétraitement et exploitation) ont été effectuées en `Perl` et `Matlab`.

1.2. Les données de trafic

Les *données de trafic* sont des données issues de l'exploitation quotidienne du réseau de téléphonie mobile qui contiennent sous une forme ou une autre des informations relatives à l'ensemble des appels téléphoniques mobiles qui ont transité par le réseau. Une des difficultés de

notre travail a consisté à nous procurer un jeu de telles données qui soit pertinent vis-à-vis du problème initial de localisation de trafic, et exploitable.

La variété des données que nous aurons finalement récoltées témoigne du lent processus de recherche et de la difficulté à identifier une bonne source de données. Nous avons eu accès à trois types de jeux de données au cours de ces trois années de recherche, qui offraient des informations de qualités très diverses. Je tiens à remercier tout particulièrement Michel Ribeyron de FTR&D/DAC/ISS (pour les CRA) ; Jean-Marc Kelif de FTR&D/DMR/IIM et Arnaud Louis de l'UR Ile-de-France (pour les relevés Cigale) ; Marie-Hélène Busy de OF/DOD/EXT et Anne Daviaud de FTR&D/DMR/ISS (pour les HC2).

1.2.1. Les Comptes-Rendus d'Appels

Le premier jeu de données est un jeu de Comptes-Rendus d'Appels (abrégé *CRA*). Il est en particulier utilisé à des fins de facturation des usagers. Le fichier que nous avons obtenu comptait environ 56.10^4 lignes pour un volume après compression (`Gzip`) de 6.7 Mo. Nous en donnons un court échantillon dans le Tableau 1.1 à titre d'illustration.

```
00 331460 33148742882 208010310B320F FFFFFFFFFFFFFF 02 12 35 52 00003
00 331460 33610300423 20801031085E8F FFFFFFFFFFFFFF 02 12 35 50 00027
00 331460 33164681525 208010300EBCDF FFFFFFFFFFFFFF 02 12 26 23 00065
00 331460 33146603294 2080103007F5BF FFFFFFFFFFFFFF 02 12 27 34 00049
00 331460 33616951068 208010300A701F FFFFFFFFFFFFFF 02 12 27 41 00047
```

Tableau 1.1 – Echantillon de CRA. Les zones séparées par des espaces correspondent respectivement aux : appel entrant ou sortant (ici `00` pour sortant à l'échelle de tout le fichier) ; code du MSC où l'appel a été initialisé (ici toujours `331460`) ; numéro appelé ; code de la BTS qui a pris en charge l'appel à son initialisation ; complément d'information (ici brouillé) ; jour, heures, minutes auxquels l'appel a été initialisé ; durée de l'appel en secondes

Ces données sont insatisfaisantes à plus d'un titre. En particulier parce que les informations de localisation qu'elles contiennent se limitent à la cellule où l'appel a été initialisé. Ou encore parce que seuls les appels sortants sont répertoriés. On ne peut ainsi pas calculer le trafic instantané supporté par une BTS†.

1.2.2. Les relevés Cigale

Présentation

Les relevés Cigale (acronyme pour Contrôle de l'Interface Généralisée A partir des Lectures d'Enregistrements) sont effectués automatiquement au niveau des MSC† par le biais du sous-système OSS†. Nous en avons obtenu trois exemplaires datés du 5 février 2002 pour trois MSC parisiens.* Le nombre total de lignes est de l'ordre de 20.10^6 pour un volume après compression (`Bzip2`) d'environ 650 Mo.

*Nous avons travaillé avec les relevés Cigale sous leur forme brute, soit le format `.x13`. Il existe bien des formats prétraités, prévus notamment à des fins de description statistique, mais nous avons jugé que la perte d'informations était bien trop importante au regard du problème initial et des réponses que nous comptions essayer d'y apporter.

Nous pouvons affirmer que ces données sont exhaustives dans la mesure où elles gardent la trace de tous les échanges (*via* le NSS†) entre tous les mobiles et les trois MSC observés ce jour-là.

Une description précise du format original n'a pas sa place ici. Il est important en revanche de résumer la somme des informations d'intérêt contenues dans les relevés et que nous avons exploitées. La réduction des données originales par conservation uniquement des informations d'intérêt fait passer à environ 230 Mo leur volume compressé (`Bzip2`), soit environ 35% du volume initial.

Fin connexion	9:05:03:598	9:50:44:403	10:05:04:260
Début connexion	9:04:56:735	9:50:11:553	10:04:28:677
Durée connexion	6863	32850	35583
CI	TUCUMAN	SURABAYA	FLORES
Etat	OC_HO_COMM	TC_COMM	OC_COMM
IMEI	4491945669775	5449194566977	7544919456697
Numéro connexion	587060	886989	872548
Nombre handovers	>	.	2

Tableau 1.2 – Echantillon filtré d'un relevé Cigale. `Fin connexion` : heures, minutes, secondes, millisecondes de la fin de l'échange que cette colonne résume. `Début connexion` : comme avant, pour le début de l'échange. `Durée connexion` : durée de l'échange en millisecondes. `CI` : code de la cellule gérant la connexion au cours de l'échange (ici brouillé). `Etat` : toute une nomenclature de codes décrivant la nature de l'échange. Les préfixes `OC_` et `TC_` sont associés respectivement à des **appels sortants** (Outcoming Call) ou des **appels entrants** (Terminating Call). `IMEI` : numéro **unique** d'identification de l'appareil mobile attribué à sa fabrication (ici brouillé). `Numéro de connexion` : numéro de la connexion auquel cet échange correspond. `Nombre handovers` : `>` pour un début de connexion, `.` pour une continuation de connexion non terminale et sinon, nombre total de handovers.

Une *connexion* consiste en une série d'*échanges* (qui en sont les parties élémentaires) entre un mobile, le BSS† et le NSS. Une connexion peut correspondre à un appel, mais aussi plus généralement à des échanges protocolaires, dont des mises à jour automatiques de localisation pour les bases HLR† et VLR†.

Toute série d'échanges entre un mobile et un MSC est identifiée par un unique `Numéro de connexion` (nombre utilisé une fois et une seule par jour). Chaque échange émet un rapport lorsqu'il s'achève (sous la forme d'une ligne dans notre fichier). Les rapports sont toujours repérés par le `Numéro de connexion` de la connexion correspondante. Une nomenclature de codes permet de qualifier la nature de l'échange grâce à la variable `Etat`. Le mobile qui est à l'origine de la connexion est repéré par son `IMEI`, *i.e.* par le numéro unique de fabrication attribué au téléphone. De plus, le rapport d'échange précise les heures de début et de fin de l'échange ainsi que sa durée *via* les variables `Début connexion`, `Fin connexion` et `Durée connexion`. La variable `CI` indique la BTS† qui gérait la connexion au moment de l'émision du rapport. Enfin, la variable `Nombre handovers` informe de la nature initiale (`>`), intermédiaire (`.`) ou finale (nombre total de handovers) de l'échange correspondant relativement à la série d'échanges qui constitue

l'intégralité de la connexion.

En résumé, on peut répartir ces variables en quatre catégories :

- variables de suivi : `Numéro connexion`, `IMEI`, `Nombre handovers`;
- variable de nature de l'échange : `Etat`;
- variable de localisation : `CI`;
- variables temporelles : `Début connexion`, `Fin connexion`, `Durée connexion`.

Le Tableau 1.2 offre un aperçu de ces informations.

Extraction d'informations

En vertu de la brève description proposée en légende du Tableau 1.2, nous pouvons filtrer grâce à la variable `Etat` toutes les lignes de nos relevés qui concernent des échanges relatifs à un appel, avec la distinction entre les appels sortants et entrants. Nous pouvons ainsi laisser de côté les lignes dues à des opérations automatiques (comme la mise à jour de localisation pour les bases VLR et HLR, certaines mesures de champ *etc*, voir le Prélude). Nous donnons dans le Tableau 1.3 une description statistique simple des échanges en fonction de leur nature.

Approximation du nombre NE d'échanges	20.10^6		
Approximation du nombre NC de connexions	14.10^6	71%	Quotient NC/NE
Approximation du nombre NEA d'échanges correspondant à un appel	8.10^6	43%	Quotient NEA/NE
Approximation du nombre NCA de connexions		26%	Quotient NCA/NC
correspondant à un appel	3.10^6	44%	Quotient NCA/NEA
Approximation du nombre NCAS de connexions correspondant à un appel sortant	2.10^6	68%	Quotient NCAS/NCA

Tableau 1.3 – Nombres approchés de connexions et d'échanges et quelques pourcentages intéressants calculés sur notre jeu de données Cigale.

Nous nous sommes particulièrement intéressés à deux ensembles de données spécifiques.

Durée des connexions : le premier concerne la *durée des connexions liées à des appels*, *i.e.* en fait à la durée des appels comprise comme laps de temps s'écoulant entre la validation du numéro d'appel et la déconnexion (pour un appel sortant), ou comme laps de temps s'écoulant entre le début de la sonnerie du téléphone et la déconnexion (pour un appel entrant).

Gestion des connexions : le second concerne les prises en charge par les BTS de connexions liées à des appels. Deux cas de figure sont possibles du point de vue d'une BTS b dont le `CI` est codé `ATLAS` :

- prise en charge à une certaine heure (indiquée par `Début connexion`) de la gestion de la connexion dont le `Numéro connexion` est `123456` :
  correspond à une initiation d'appel (`Nombre handovers` renseigné `>`) assurée par la BTS b (`CI` renseigné `ATLAS`), ou à l'aboutissement d'un transfert de connexion par handover (`Nombre handovers` renseigné par un . ou par le nombre de handovers selon les cas) d'une autre BTS b' vers b (`CI` renseigné `ATLAS`) ;
- cession de la gestion à une certaine heure (indiquée par `Début connexion`) de la connexion dont le `Numéro connexion` est `123456` :
  correspond à une terminaison d'appel (`Nombre handovers` renseigné par le nombre de

handovers) assurée par la BTS b (`CI` renseigné `ATLAS`), ou à l'aboutissement d'un transfert de connexion par handover (`Nombre handovers` renseigné par un . ou par le nombre de handovers selon les cas) de b vers une autre BTS b' de `CI` codé `YUKON` (`CI` est renseigné `YUKON`).

Une étude succinte des durées de connexions liées à des appel est reportée à la Section 1.3.

Concernant les données de gestion des connexions par les BTS, on peut notamment en extraire :

- les *nombres d'appels en cours* à chaque instant pour chaque BTS, avec distinction éventuelle selon la nature (sortant, entrant) de la communication. Ces quantités ont le mérite de correspondre exactement au nombre d'usagers connectés au réseau depuis la cellule sectorielle. Voir la Figure 1.1 pour un exemple de nombre d'appels en cours sur une BTS standard parisien.
- les nombres cumulés d'échanges pris en charge ou cédés, pour chaque BTS, avec distinction éventuelle selon la nature de la communication. Voir encore la Figure 1.1 pour une illustration.
  Les nombres d'appels pris en charge permettent entre autres d'évaluer la *quantité de trafic écoulée* par chaque BTS, voir ci-dessous pour davantage de précisions.
- des observations de *comportements individuels* (au sens du téléphone mobile identifié par son `IMEI`) à l'échelle de la journée.

Nous n'avons pas eu le temps de nous pencher sérieusement sur l'étude des comportements individuels autrement que du point de vue des durées de connexions. Il y a là sans doute matière à une étude statistique prometteuse.

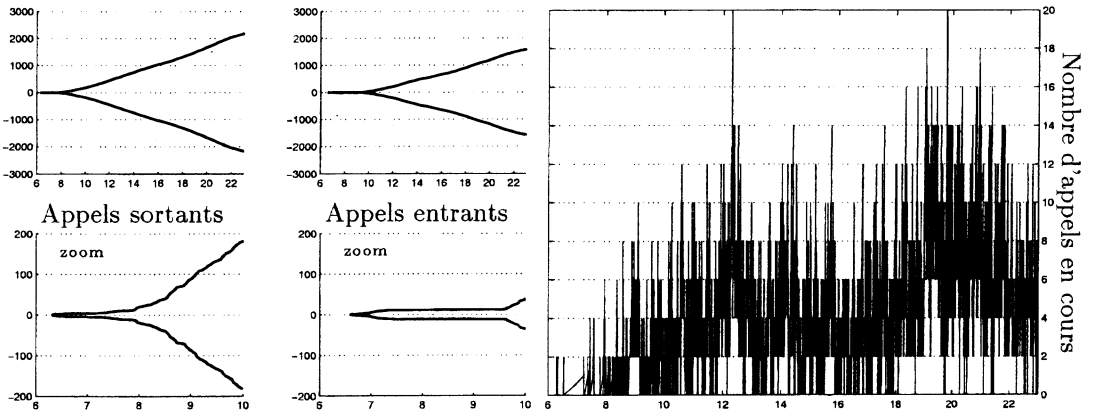


Figure 1.1 – Gestion de connexions par une BTS standard sur notre journée d'observation. A gauche, graphique haut : les courbes supérieure et inférieure représentent respectivement le nombre cumulé de connexions prises en charge ou cedées en fonction du temps (exprimé en heures, en abscisse), en se restreignant aux appels sortants. A gauche, graphique bas : comme précédemment, entre six et dix heures. Au milieu, comme les deux premiers graphiques, restreint aux appels entrants. A droite, nombre de connexions en cours de gestion (tous appels confondus) en fonction du temps.

Quantité de trafic et Erlang

Les quantités de trafic téléphonique ont leur unité propre, appelée *Erlang* en hommage à l'initiateur de la théorie des files d'attente. L'unité Erlang est abrégée E et n'a pas de dimension. *Un Erlang est l'équivalent d'un appel sur un canal radio pendant 3600 secondes.* Dans la mesure où le système de télécommunication qui nous occupe est digital et qu'un même canal peut écouler simultanément plusieurs appels, un certain canal peut écouler une quantité de trafic de, disons, 3 E de diverses façons, dont par exemple évidemment, sur une même heure :

- trois appels simultanés d'une heure chacun,
- six appels de trente minutes chacun, ou encore
- cent quatre-vingts appels d'une minute chacun.

Une formule communément employée permet d'estimer la quantité de trafic en Erlang écoulée sur une durée. Supposons que l'on souhaite évaluer la quantité de trafic $Q_0(b)$ écoulée par la BTS b entre H_0 et $H_0+\Delta_0$ (H_0 et Δ_0 sont exprimées en secondes). Soit A_0 le nombre d'échanges initiés dans cet intervalle de temps et T_0 leur durée moyenne (exprimée en secondes). Alors on approxime $Q_0(b)$ par

$$Q_0(b) = \frac{A_0 T_0}{\Delta_0} \text{ E}. \tag{1.1}$$

Il est clair d'après la précédente sous-section que l'on peut calculer grâce au jeu de données Cigale des quantités de trafic écoulée par chaque BTS parisienne au cours de n'importe quel laps de temps sur notre journée d'observation.

1.2.3. Les relevés HC2

Rappelons que les CRA sont à l'origine des données destinées à la facturation des usagers. Les relevés Cigale fournissent quant à eux une mine importante de données mais sous une forme brute qui requiert beaucoup de soins avant de révéler ses informations.

Contrairement aux deux précédents, les relevés HC2 sont conçus à l'origine à des fins de description quantitative du trafic écoulé. Ils servent notamment à la planification de réseau (*i.e.* au choix de nombre de stations, de leur positionnement et calibrage), à l'identification de zone desservies de façon inappropriée et au suivi global des flux. Il convient enfin de souligner que les relevés HC2 sont hautement confidentiels.

HC2 est un acronyme pour 2ème Heure Chargée. Les relevés HC2 rassemblent les données hebdomadaires HC2 pour chaque cellule. La donnée HC2 notée HC2(b) de la cellule de la BTS b pour la semaine $\mathcal{S} = \{j_1, \ldots, j_7\}$ est calculée comme suit : *primo*, calcul des quantités de trafic écoulée $Q_{1,h}(b), \ldots, Q_{7,h}(b)$ (en Erlang) chaque heure h de 6 heures du matin à 23 heures, pour chaque jour de la semaine $\mathcal{S}$; *secundo*, calcul des maxima quotidiens $Q_1(b), \ldots, Q_7(b)$ de ces quantités de trafic ; *tertio*, choix final de la seconde plus grande de ces quantités, d'où l'appelation « seconde heure chargée ». Ainsi, formellement, notant $A_{k,h}$ le nombre d'échanges pris en charge

entre h et $h+1$ par la BTS b le jour j_k et $T_{k,h}$ la durée moyenne de ces échanges,

$$\begin{aligned} Q_{k,h}(b) &= \frac{A_{k,h}\,T_{k,h}}{3600} \quad (\text{tout } 6 \leq h \leq 23 \textit{ et } 1 \leq k \leq 7), \\ Q_k(b) &= \max_{6\leq h\leq 23} Q_{k,h}(b) \ (\text{tout } 1 \leq k \leq 7) \quad \text{et} \\ \mathrm{HC2}(b) &= \max\left(\left\{Q_k(b) : 1 \leq k \leq 7\right\} \setminus \left\{\max_{1\leq k\leq 7} Q_k(b)\right\}\right). \end{aligned} \tag{1.2}$$

Nous nous sommes procurés les données HC2 sur Paris pour les 21 semaines des mois de mars à juillet de cette année 2002.

1.2.4. Quelles données pour quelle localisation ?

On a déjà argumenté plus tôt pourquoi les relevés CRA sont insatisfaisants pour notre objectif de localisation plus fine du trafic. Que peut-on dire en revanche des extractions Cigale et des relevés HC2 ?

Les relevés HC2 sont inadaptés pour une description horaire

De par leur définition, les relevés HC2 sont révélateurs à l'échelle d'une semaine. Ils rendent compte de la quantité de trafic horaire la plus importante qu'ait eu à gérer chaque BTS, mais hors circonstance exceptionnelle. Cette nuance est heuristiquement la conséquence du choix de la *seconde heure* la plus chargée quotidienne.

Ce choix d'indicateur de trafic est en accord avec le principe industriel de consitution d'un réseau téléphonique mobile qui assure le plus de satisfaction aux usagers sans pour autant être sur-dimensionné. C'est en somme un problème d'ajustement pénalisé comme il sera discuté dans les Chapitres 3, 5 et 7.

Il va en revanche de soi que les relevés HC2 ne sont pas adaptés pour une description à l'échelle, par exemple, de l'heure. La Figure 1.2 témoigne en ce sens : on y représente, pour neuf cellules standard* mais néanmoins choisies* et pour six heures de notre journée d'observation Cigale,* les quantités de trafic écoulées (en Erlang) ainsi que les maxima de ces quantités. On constate bien sûr que la valeur maximale de ces six quantités pour chaque cellule (qui ne coïncide pas avec la quantité de l'heure la plus chargée intervenant dans le calcul du HC2) ne préjuge pas du comportement en termes de trafic écoulé heure par heure.

Existe-t-il une tendance générale pour les trafics horaires ?

L'inadéquation des relevés HC2 pour rendre compte du trafic horaire est d'autant plus prononcée que l'on ne peut pas dégager de tendance générale marquée pour ce dernier. Autrement dit, le trafic horaire maximal écoulé par une BTS sur une journée est statistiquement susceptible de l'avoir été à n'importe quelle heure raisonnable (*i.e.* pas entre huit et dix heures du matin, ni entre neuf et onze heures du soir).

**i.e.* non atypiques : leur comportement est révélateur de celui d'un nombre élevé de cellules sectorielles.

**nous avons en particulier sélectionné des cellules dont les trafics ont globalement les mêmes tendances de croissance et décroissance pour faciliter la lisibilité du graphique.

**sur les 17 heures d'observation : ces heures sont connues comme distinctives pour le jeu complet de données.

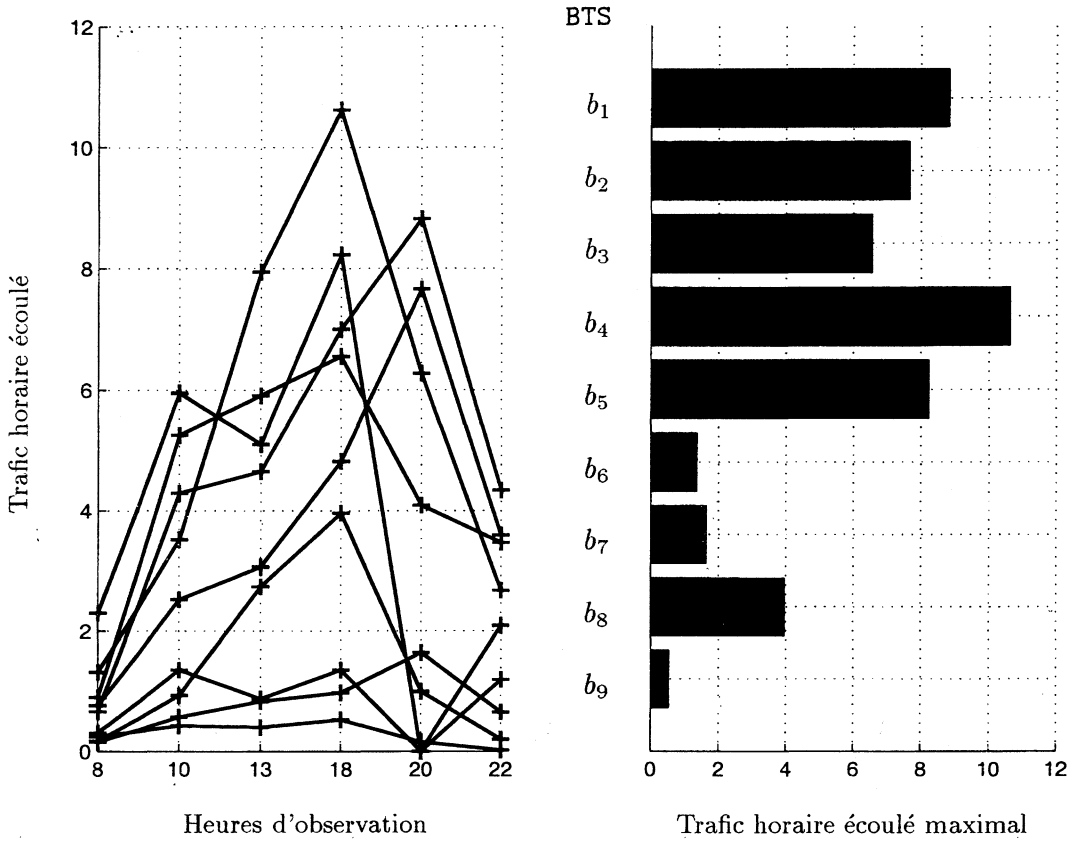


Figure 1.2 – Quantités horaires de trafic écoulées : comparaison des maxima quotidiens et des quantités horaires. A gauche : en abscisses, les six heures sur lesquelles on évalue les quantités de trafic écoulées pour neuf cellules (de BTS notées $b_1, \ldots, b_9$) ; en ordonnées, les quantités de trafic en Erlang. A droite : en abscisses, les quantités maximales de trafic écoulées pour chaque BTS ; en ordonnées, les BTS classées de bas en haut par valeurs croissantes des quantités de trafic écoulées entre 22 et 23 heures (qui correspondent à la dernière colonne du graphique de gauche).

Le graphique droit de la Figure 1.3, où de telles quantités horaires de trafic écoulées sont empiriquement centrées et renormalisées, est destiné à illustrer notre propos. Nous avons sciemment changé de cellules d'observation : les cellules qui participent à la Figure 1.2 ont été choisies afin d'illustrer le manque de représentativité du maximum quotidien de trafic écoulé au regard des trafics horaires. Ce n'est plus le cas pour les douze cellules sélectionnées pour la Figure 1.3, et cependant encore standard. On remarque notamment que le trafic horaire sur la tranche de 8 à 9 heures est toujours en-deçà de la moyenne quotidienne et que le trafic horaire sur les tranches de 10 à 11 heures et de 13 à 14 heures est peu révélateur. En revanche, les valeurs prises sur les tranches de 18 à 19, 20 à 21 et 22 à 23 heures sont assez caractéristiques.

Les enveloppes supérieures et inférieures du graphique droit rendent compte des quantités extrêmes des trafics recentrés et renormalisés sur l'ensemble des BTS parisiennes. Le Tableau 1.4 complète les informations apportées par ces enveloppes en précisant les quartiles et la médiane

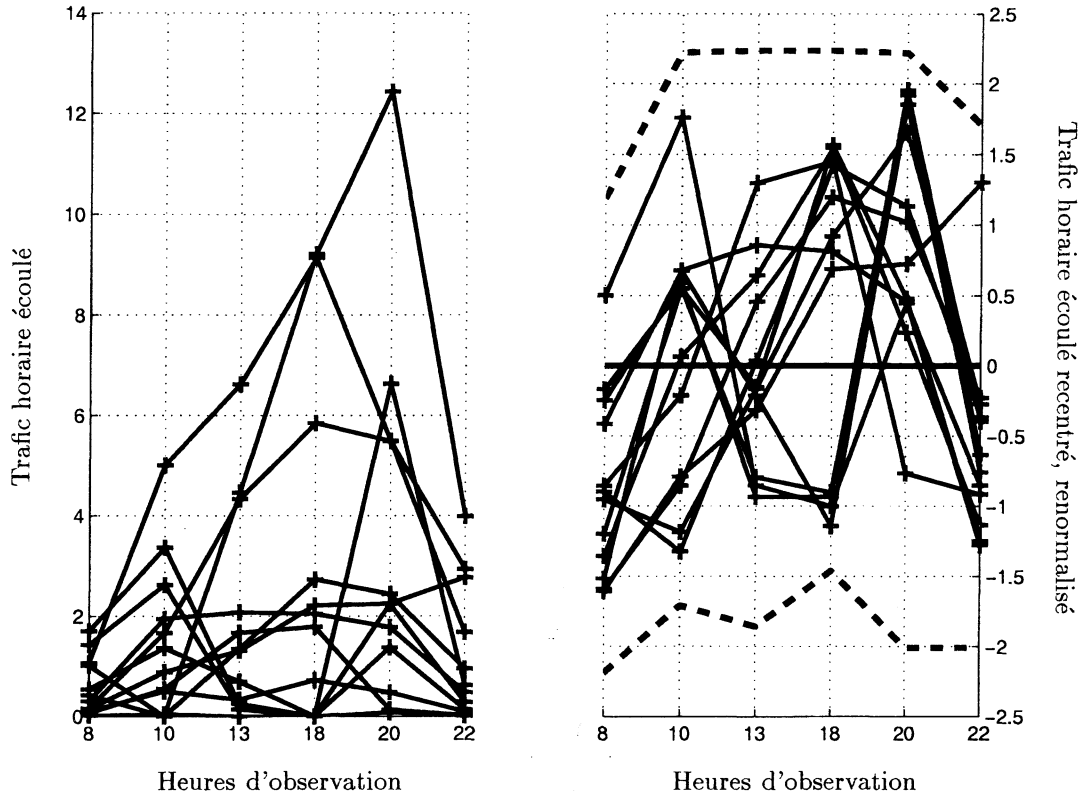


Figure 1.3 – Quantités horaires de trafic écoulées. A gauche, même composition que pour le graphique gauche de la Figure 1.2 pour douze autres cellules. A droite : en abscisses, les heures d'observation ; en ordonnées, les quantités horaires de trafic empiriquement recentrés et renormalisés pour les mêmes cellules qu'à gauche. Les lignes brisées supérieures et inférieures représentent respectivement les maxima et minima sur l'ensemble de toutes les cellules de Paris des trafics recentrés renormalisés.

pour les trafic horaires recentrés et renormalisés, pour chaque tranche horaire d'observation.

Tranche horaire	8–9	10–11	13–14	18–19	20–21	22-23
Quartile inférieur	-1.47	-0.46	-0.29	0.72	0.06	-0.90
Médiane	-1.15	0.03	0.09	1.20	0.85	-0.62
Quartile supérieur	-0.75	0.52	0.50	1.54	1.32	-0.19

Tableau 1.4 – Quartiles pour les répartitions empiriques des trafics horaires recentrés et renormalisés pour chaque plage d'observation.

Quelle localisation ?

C'est finalement la question cruciale à quoi répondre. L'enjeu n'est pas ici, on le comprend bien, de déterminer une *échelle spatiale* de raffinement (question délicate qui sera abordée plus tard) mais bien de décider à quelle *échelle temporelle* on souhaite se placer.

L'ensemble des remarques que nous venons de faire laisse préjuger que notre préférence va à l'échelle temporelle de l'ordre de l'heure. Voici en effet trois motifs en ce sens :

- *primo*, on observe une certaine stationnarité du trafic écoulé sur des plages homogènes de l'ordre d'une à deux heures. De telles plages se prêtent donc bien à la localisation du trafic et à son raffinement.
- *secundo*, une telle description à l'échelle de l'heure n'exclut en rien une description à l'échelle de la journée ni de la semaine. Bien au contraire, elle peut y conduire par synthèse des résultats sur les tranches horaires. De plus, cette approche permet naturellement une meilleure compréhension des mécanismes quotidiens et hebdomadaires.
- *tertio*, si la méthode s'avérait performante, on pourrait aisément automatiser la récolte des données nécessaires au niveau des MSC en adaptant une des simples routines que nous avons écrites et en l'incorporant par exemple à l'interface Cigale. Cela permettrait de résoudre le défaut que comporte ce choix d'échelle, lié à la relative difficulté d'extraire les données.

Néanmoins, la méthode que nous allons développer est caractérisée par une très grande *flexibilité*. Elle ne dépend en particulier pas de la préférence affichée ci-dessus pour une échelle temporelle de l'ordre de l'heure. Aussi, nous l'appliquerons à notre échelle de temps favorite, mais aussi à l'échelle d'une journée complète (grâce à une astuce permettant d'introduire l'heure d'observation des quantités de trafic dans le jeu de données) et encore aux données hebdomadaires HC2. L'étude est reportée au Chapitre 4.

1.3. Ebauche d'une étude statistique des durées d'appel

Cette section est anecdotique. Elle est l'occasion de donner un aperçu statistique assez informel des durées d'appel. Une étude plus poussée rentrerait naturellement dans l'approche suggérée en Section 1.2.2 où l'on extrairait des informations individuelles (*i.e.* mobile par mobile) des relevés Cigale.

Pour commencer, il faut noter que les durées d'appels ne sont pas stationnaires à l'échelle de la journée. Le graphique gauche de la Figure 1.4 illustre notre propos : nous avons découpé de façon arbitraire la journée d'observation en 17 tranches horaires d'une heure, débutant chacune à une heure pleine. Pour toutes ces tranches horaires, nous avons calculé la durée moyenne des appels, d'une part sortants, d'autre part entrants (respectivement colonne gauche et droite). L'asymétrie entre les durées d'appels sortants et entrants est frappante. Un élément d'explication est sans doute le fait que les appels entrants incluent les appels depuis le réseau de téléphonie fixe, alors que ceux-ci sont exclus des appels sortants. On note aussi comme annoncé une tendance prononcée de croissance au fil de la journée des durées d'appels entrants, et celle moins nette des durées d'appels sortants, révélatrices de non stationnarité.

Le graphique gauche de la Figure 1.4 suggère la coexistence de deux populations au sein des durées d'appels, selon que l'appel est sortant ou entrant. Qu'en est-il des durées d'appels sortants uniquement ? Notre étude perd ici un peu en rigueur. Nous allons travailler sur les durées des appels sortants initialisés entre 10 heures et 17 heures, et ce malgré le commentaire préliminaire sur la non stationnarité des durées d'appels à cette échelle. On observe néanmoins que les moyennes des durées correspondantes sur chaque tranche horaire sont peu dissemblables.

La première question est naturellement de tester si une loi exponentielle pour les durées d'appels sortants s'ajuste bien aux observations. Ce n'est pas le cas, comme on le constate sur

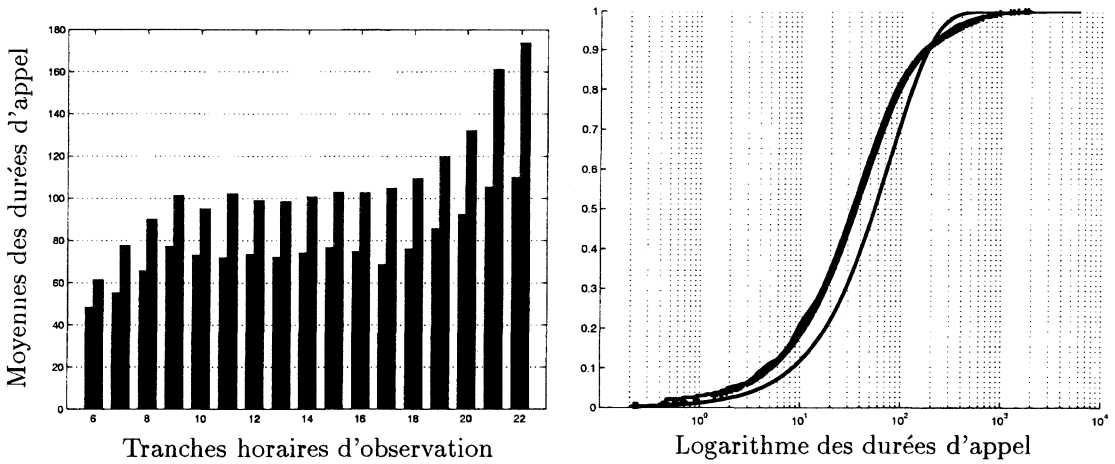


Figure 1.4 – Illustrations d'une étude statistique sommaire des durées d'appels. A gauche, moyennes (en secondes) des durées d'appel sortant (colonnes de gauche) et entrant (colonnes de droite) calculées sur chaque tranche horaire de 6 à 22 heures. A droite, fonction de répartition empirique des durées des appels sortants débutés entre 10 et 17 heures, fonctions de répartition d'une loi exponentielle ajustée (sous la précédente pour 80% des observations) et d'un mélange de deux exponentielles ajusté (presque superposée à la fonction de répartition empirique). L'échelle en abscisses est logarithmique.

le graphique de droite de la Figure 1.4 : la fonction de répartition de l'exponentielle ajustée est située nettement sous la fonction de répartition empirique des observations pour 80% d'entre elles. Il faut préciser ici que nous nous sommes limités à un sous-échantillon (10^4 durées d'appels tirées au hasard uniformément sans remise) moins volumineux que l'original (5.10^5 observations).

La seconde question est donc de tester si un *mélange de deux lois exponentielles* s'ajuste bien. Il s'ajuste mieux puisqu'une telle loi est plus sophistiquée qu'une simple exponentielle. Il mérite cependant d'être souligné qu'un tel modèle de mélange est suggéré par le comportement téléphonique des usagers, qui alternent des appels assez brefs avec des appels sensiblement plus longs.

L'ajustement a été effectué grâce à un algorithme EM (pour Expectation Minimization ; 1000 itérations pour une convergence rapide constatée). La fonction de répartition ajustée colle de très près à la fonction de répartition empirique, voyez le graphique de droite de la Figure 1.4 (les deux courbes sont presque superposées). Nous avons aussi testé l'ajustement par Kolmogorov–Smirnov (nous avons pour cela tiré uniformément 10^3 durées d'appels du jeu original privé des 10^4 données dédiées à l'ajustement). Les résultats sont présentés dans le Tableau 1.5.

Le degré de signification du test ne permet pas de conclure catégoriquement à une bonne adéquation au sens de Kolmogorov–Smirnov. Nous avons constaté que les degrés de signification sont d'autant plus élevés que la plage horaire sur laquelle on ajuste le modèle est réduite. Finalement, on a constaté qu'un mélange de trois lois exponentielles plutôt que deux n'induisait pas d'amélioration sensible au sens du même test de Kologorov–Smirnov. Nous n'avons cependant pas procédé à une estimation d'un meilleur ordre, *i.e.* d'un meilleur nombre de populations pour le mélange d'exponentielles. Il existe une théorie raffinée sur cette question, auquel le Chapitre 7 apporte sa contribution. L'aspect pratique a lui aussi été largement exploré. L'étude des ces

Proportions	Moyennes (secondes)	Degré de signification du test KS
16%	280	
84%	40	11%

Tableau 1.5 – Proportions et moyennes respectives d'un mélange de deux lois exponentielles ajusté aux observations de durées d'appel sortant initialisé entre 10 et 17 heures. Dans ce modèle ajusté, tout se passe comme si l'usager décidait à pile ou face si son appel sera court (84% de chance, durée moyenne 40 s) ou long (16% de chance, durée moyenne 280 s). On indique aussi le degré de signification du test de Kolmogorov–Smirnov pratiqué sur un sous-échantillon indépendant de celui qui a servi à l'ajustement.

questions trouverait évidemment sa place dans un travail ultérieur.

1.4. Les données de recouvrement cellulaire

Les données de trafic discutées plus tôt ont pour complément les *données de recouvrement cellulaire*. Celles-ci permettent d'évaluer les zones de desserte de chaque BTS, de sorte que chaque cellule de Paris intra-muros se voit associer des informations de trafic et une description de sa zone de recouvrement. Nous nous attachons ci-dessous à présenter la méthode d'évaluation des zones de desserte ainsi qu'une brève description quantitative des résultats.

1.4.1. Emission, affaiblissement, réception

Schématiquement, une BTS émet et reçoit des ondes radio codant les échanges protocolaires et le contenu des appels en provenance ou bien à destination des mobiles. Une BTS a une certaine *puissance d'émission* P qui induit un *champ* dont la valeur à proximité de la BTS vaut

$$C = 10 \log_{10} P$$

et dont l'unité est le dB si P est en Watt (W) et le dBm si P est en mW.

Ce champ est sujet à un *affaiblissement* de propagation, de sorte que plus l'on s'éloigne de la BTS, plus le champ émis par elle en cet endroit est faible. Les lois d'affaiblissement dépendent d'un modèle physique de propagation, de la topographie des lieux et du paramétrage de la BTS. Ainsi, une BTS dessert typiquement un *secteur* en forme de pétale (voyez la description quantitative des résultats en Section 1.4.3).

Plus particulièrement, un mobile situé en un point donné ne pourra *décrypter* les informations émises par une BTS que si le champ dû à la BTS en ce point est supérieur à un certain seuil, dit *seuil de décryptage*. Nous donnons dans le Tableau 1.6 les valeurs standard de la puissance d'une BTS et du seuil de décryptage d'un mobile selon les données constructeur.

Puissance BTS	Seuil décryptage
43 dBm	-104 dBm

Tableau 1.6 – Puissance standard d'une BTS et seuil constructeur standard de décryptage pour un mobile.

1.4.2. Méthodologie pour l'évaluation des cellules

Nous allons expliquer ici la méthodologie que nous avons suivie pour obtenir une évaluation des zones de desserte des BTS.

Parcell©

L'application Parcell©* est un outil d'ingénierie interne au groupe France Télécom qu'utilisent quotidiennement les ingénieurs de France Télécom. Elle permet *notamment*, pour n'importe quelle configuration de placement et de paramétrage de BTS, de calculer le champ des affaiblissements associés, *i.e.* l'affaiblissement du champ induit par chaque BTS selon l'endroit où ce champ est évalué. Cette application fait appel à un ensemble de bases de données qui incorporent le modèle de propagation et la topographie des lieux.

Par ailleurs, les affaiblissements sont évalués à l'échelle de la *maille*, qui est un carré dont la longueur d'un côté est un multiple de 25 mètres, typiquememt 25, 50 ou 100 mètres en zone urbaine.

Nous avons ainsi dans un premier temps, avec l'aide précieuse de Joël Tartière de FTR&D/DMR/OIP que je remercie encore au passage, calculé le champ des affaiblissements pour la configuration réelle des BTS (soit paramétrage et emplacement) sur Paris intra-muros.

En résumé, nous disposions à cette étape, pour chaque BTS, de la valeur de l'affaiblissement subi par son champ original en chaque maille parisienne.

Persée

L'application Persée (dénomination interne) est un nouvel outil d'ingénierie interne à France Télécom R&D complémentaire de Parcell©. Persée exploite les données de sortie de Parcell©. Ses fonctions consistent entre autres à (pour ce qui nous concerne) :

- *primo*, uniformiser le pas des mailles du maillage sur lequel on dispose des données d'affaiblissement ;
- *secundo*, basculer le référençage de ces données, afin de transformer les informations d'affaiblissement pour chaque BTS, en chaque maille, en des informations d'affaiblissement en chaque maille, pour toutes les BTS ;
- *tertio*, effectuer une sélection raisonnée des données afin de réduire le volume occupé par elles sans perdre toutefois d'informations utiles. Rappelez-vous en effet que pour Paris intra-muros (environ 3 000 BTS et environ 20 000 mailles de pas divers), Parcell©fournit plus de 5.10^7 valeurs d'affaiblissements. Cette sélection est fondée sur le critère `CMC`, voir (1.3) et le détail ci-dessous.

Le détail du protocole de téléphonie mobile* suggère que l'on peut se contenter, en chaque maille, de ne conserver (à des fins de réduction du volume des données) que les affaiblissements des BTS dont le champ induit en la maille n'est pas trop faible en comparaison du meilleur champ induit. Ainsi, si $\mathbb{BTS}$ dénote l'ensemble des BTS, si $\mathbb{M}$ dénote l'ensemble des mailles, si $A_m(b)$ dénote l'affaiblissement subi par la BTS $b \in \mathbb{BTS}$ en la maille $m \in \mathbb{M}$ et si $C(b)$ dénote le

*Marque déposée par France Télécom.

*Soit, grossièrement, qu'un mobile se contente de répertorier les meilleures serveuses, *i.e.* les BTS dont les champs à l'emplacement où se situe le mobile sont *parmi les plus élevés* – critère dont la formulation est en accord avec le critère `CMC` présentée dans (1.3).

champ original produit par la BTS b (dans la même unité que l'affaiblissement, généralement le dBm), on ne conserve la donnée d'affaiblissement en la maille m pour la BTS b que si

$$\Big\{C(b) - A_m(b)\Big\} - \max_{b' \in \mathbb{BTS}} \Big\{C(b') - A_m(b')\Big\} \geq \texttt{CMC=-4}, \tag{1.3}$$

où `CMC` est la valeur seuil de Champ moins Meilleur Champ (sans unité), prise égale à `-4` sur recommandation avisée.

Concrètement, nous avons choisi un pas uniforme de maillage de 50 mètres puis appliqué Persée pour `CMC=-4`, disposant finalement, pour chacune des 28 431 mailles de Paris intra-muros, de l'ensemble des affaiblissements subis par l'ensemble des BTS en cette maille et satisfaisant la condition de `CMC` donnée par (1.3).

La Figure 1.5 et le Tableau 1.7 offrent respectivement une illustration de l'effet d'affaiblissement en une maille et un extrait de la base de données issues de Parcell©et Persée.

```
3673,113,600475,2424925,-105.020001
3673,135,600475,2424925,-104.399998
3674,135,600525,2424925, -97.279998
3675,135,600575,2424925,-104.779998
3675,136,600575,2424925,-107.500003
3676,135,600625,2424925,-101.529998
3677,135,600675,2424925, -94.149998
3678,135,600725,2424925,-100.149998
3679,323,600775,2424925, -71.969989
3680,135,600825,2424925, -95.899998
```

Tableau 1.7 – Extrait de la base de données issues de Parcell©et Persée. Les colonnes correspondent aux, de gauche à droite : numéro de maille, code BTS, abscisse du centre de la maille, ordonnée du centre de la maille, champ induit par la BTS en la maille (en dBm). Observez qu'une maille peut apparaître à plusieurs lignes, en fait dès que plusieurs BTS satisfont la condition de `CMC` de (1.3).

Evaluation des cellules : la constante de seuillage `SeuilCd`

A ce stade, il ne reste plus qu'à sélectionner en chaque maille les BTS dont on estime qu'elles peuvent être sollicitées par un mobile en cette maille pour assurer temporairement la transmission des informations (soit les échanges protocolaires et l'appel à strictement parler).

On dispose à cet effet d'un nouveau paramètre de seuillage appelé `SeuilCd` (exprimé en dBm) sur lequel on peut jouer. La règle est la suivante : pour une maille donnée, on compare tous les champs reçus en celle-ci au seuil `SeuilCd` et on décide qu'une BTS ne dessert cette maille que si et seulement si son champ en la maille est plus gramd que le seuil `SeuilCd`. Autrement dit, chaque maille $m \in \mathbb{M}$ est desservie par la BTS $b \in \mathbb{BTS}$ à la condition nécessaire et suffisante que

$$C(b) - A_m(b) \geq \texttt{SeuilCd}. \tag{1.4}$$

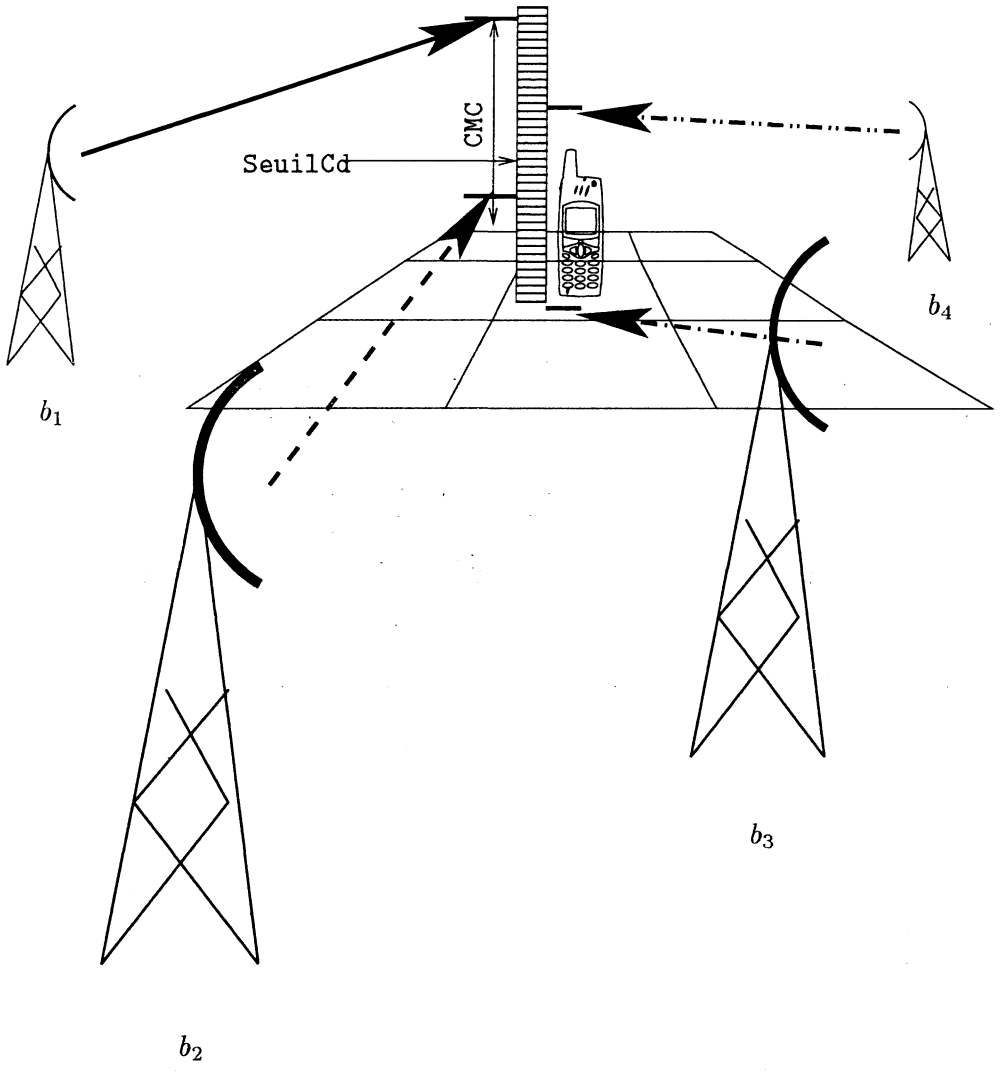


Figure 1.5 – Champs reçus en une maille. Les BTS notées b_1, b_2, b_3 et b_4 émettent toutes à la même puissance. Les flèches qui pointent de chacun d'entre elles vers le jalon disposé à côté du mobile situé au centre de la maille d'intérêt indiquent le champ de réception en la maille. Le champ émis par la BTS b_3 ne satisfait pas la condition (1.3) de `CMC`. La constante de seuillage `SeuilCd` et le critère associé (1.4) excluent ici le champ produit par la BTS b_2.

La constante de seuillage `SeuilCd` peut être vue comme un seuil de décryptage pour un mobile standard, *i.e.* la valeur minimale d'un champ en réception en-deçà de laquelle le mobile n'est plus en mesure de décoder le signal reçu. Cette interprétation suggère un choix de `SeuilCd=-104` en vertu du Tableau 1.6. Pour un tel choix, les BTS `113` et `135` d'une part, `135` et `136` d'autre part ne desservent pas les mailles `3673` et `3675`, respectivement, dans l'extrait présenté dans le Tableau 1.7.

La constante de seuillage admet cependant d'autres interprétations qui la font notamment dépendre de l'environnement (urbain, dense urbain *etc*) et de l'utilisation qu'on souhaite faire des données de recouvrement. Ainsi pour nous, le choix d'une valeur doit aussi dépendre de la pertinence des résultats de couverture qui en découlent. C'est pour cette raison que nous avons joué sur un ensemble de valeurs possibles de `SeuilCd` comprises entre `-80` dBm et `-110` dBm (de deux en deux), calculant pour chacune

- le pourcentage de recouvrement calculé sur les 28 431 mailles ;
- le pourcentage de mailles desservies par une unique BTS, *i.e.* appartenant à une unique cellule ;
- le nombre moyen de BTS servant une maille et l'écart-type associé ;
- le nombre moyen de mailles par cellule et l'écart-type associé.

Les résultats de ce jeu sont présentés dans le Tableau 1.8.

Evaluation des cellules : commentaire du Tableau 1.8

La lecture du Tableau 1.8 nous inspire les quelques commentaires suivants :

- Le pourcentage de mailles couvertes (seconde colonne du tableau) donne une approximation fidèle de la proportion de la superficie de Paris desservie par le réseau. Naturellement, le pourcentage de mailles couvertes est une fonction décroissante de `SeuilCd` : plus ce seuil est haut, moins de BTS offrent un champ en la maille qui lui soit supérieur.
  Plus surprenant au premier abord, quoique justifiable, le nombre de mailles desservies par une unique BTS n'est pas une fonction croissante de `SeuilCd` : lorsque ce seuil devient trop grand, certaines mailles perdent la desserte par leur unique BTS de couverture en vertu du critère `SeuilCd` (1.4), d'où la baisse de la proportion de maille servies uniquement.
- Quant aux nombres moyens et aux écarts-types du nombre de cellules desservant une maille, ils mettent en évidence un taux de recouvrement important, c'est-à-dire un empiettement mutuel significatif des cellules (nous insistons ici sur le fait que le calcul du nombre de cellules par maille est fait indépendamment de la nature macro-/micro-cellulaire correspondante, ce qui induit un biais supérieur). Ces empiettements sont indispensables pour assurer la continuité d'un appel alors que l'usager se déplace : ainsi, si l'usager sort de la zone d'influence d'une première BTS pour aller vers celle d'une seconde BTS, un basculement (appelé handover, voir le Prélude) est effectué, qui assure la prise en charge de la communication par la seconde BTS.
  Il faut néanmoins tâcher de rendre ces empiettements aussi réduits que possible pour au moins deux raisons :
  · afin d'échapper à *l'effet ping-pong*, cas où deux BTS s'échangent de façon répétée la charge d'un appel émis depuis un mobile à leur frontière ;
  · parce qu'un recouvrement multiple est coûteux en termes d'équipement.
  Cependant, le système GSM–GPRS est caractérisé par l'absence d'interférence entre BTS différentes, par opposition aux systèmes de troisième génération pour lesquels l'empiette-

SeuilCd (dBm)	Pourcentage couverture (%)	Pourcentage couverture unique (%)	$M_{c/m}$ – $\sigma_{c/m}$	$M_{m/c}$ – $\sigma_{m/c}$
-80	33.0	23.0	1.4 – 0.7	9.5 – 11.4
-82	39.9	26.4	1.4 – 0.7	11.3 – 13.1
-84	46.8	29.1	1.5 – 0.8	13.2 – 14.9
-86	53.9	31.4	1.6 – 0.8	15.3 – 16.7
-88	60.6	32.8	1.7 – 0.9	17.4 – 18.5
-90	67.3	33.6	1.7 – 0.9	19.7 – 20.2
-92	73.5	33.6	1.8 – 1.0	22.1 – 22.1
-94	79.2	32.9	1.9 – 1.0	24.5 – 23.8
-96	84.0	31.5	2.0 – 1.1	26.8 – 25.4
-98	88.0	29.8	2.1 – 1.1	28.9 – 26.9
-100	91.1	27.8	2.2 – 1.2	30.9 – 28.3
-102	93.6	25.5	2.3 – 1.2	33.0 – 29.8
-104	95.4	23.2	2.4 – 1.3	34.8 – 31.2
-106	96.6	21.3	2.5 – 1.3	36.3 – 32.6
-108	97.3	19.9	2.5 – 1.3	37.6 – 33.7
-110	97.9	18.9	2.6 – 1.4	38.5 – 34.8

Tableau 1.8 – Description statistique élémentaire du recouvrement cellulaire estimé de Paris intra-muros en fonction du choix de la constante de seuillage `SeuilCd`. Contenu des colonnes, de gauche à droite : valeur du paramètre `SeuilCd`; pourcentage de mailles couvertes; pourcentage de mailles desservies par une unique BTS; moyenne (sur les mailles) $M_{c/m}$ du nombre de cellules desservant une même maille; écart-type $\sigma_{c/m}$ du nombre de cellules desservant une même maille; moyenne (sur les cellules) $M_{m/c}$ du nombre de mailles par cellule; écart-type $\sigma_{m/c}$ du nombre de mailles par cellule. La ligne encadrée correspond au choix de `SeuilCd=-104`, seuil standard de décryptage pour un mobile.

ment devra être aussi limité que possible.

- Nous nous intéresserons dans la prochaine section à l'étude des superficies des cellules, dont le nombre moyen de mailles par cellule et l'écart-type correspondant donnent une première idée statistique. Notons cependant déjà que les moyennes et écarts-types laissent entrevoir qu'il existe une grande disparité entre les cellules sectorielles en termes de superficie. Ceci est en partie dû à la coexistence de macrocellules et de microcellules.

Nous décidons de fixer comme constante de seuillage

`SeuilCd=-104`dBm.

Ce choix est essentiellement motivé par le pourcentage de couverture supérieur à 95% qu'il se voit associer. Un tel pourcentage rend assez bien compte en effet de l'expérience quotidienne qu'un usager a du réseau à Paris. La sélection de -104 parmi les valeurs candidates assurant un pourcentage de recouvrement supérieur à 95% est justifiée par sa coïncidence avec le seuil de décryptage standard d'un mobile.

1.4.3. Brève description quantitative des recouvrements cellulaires

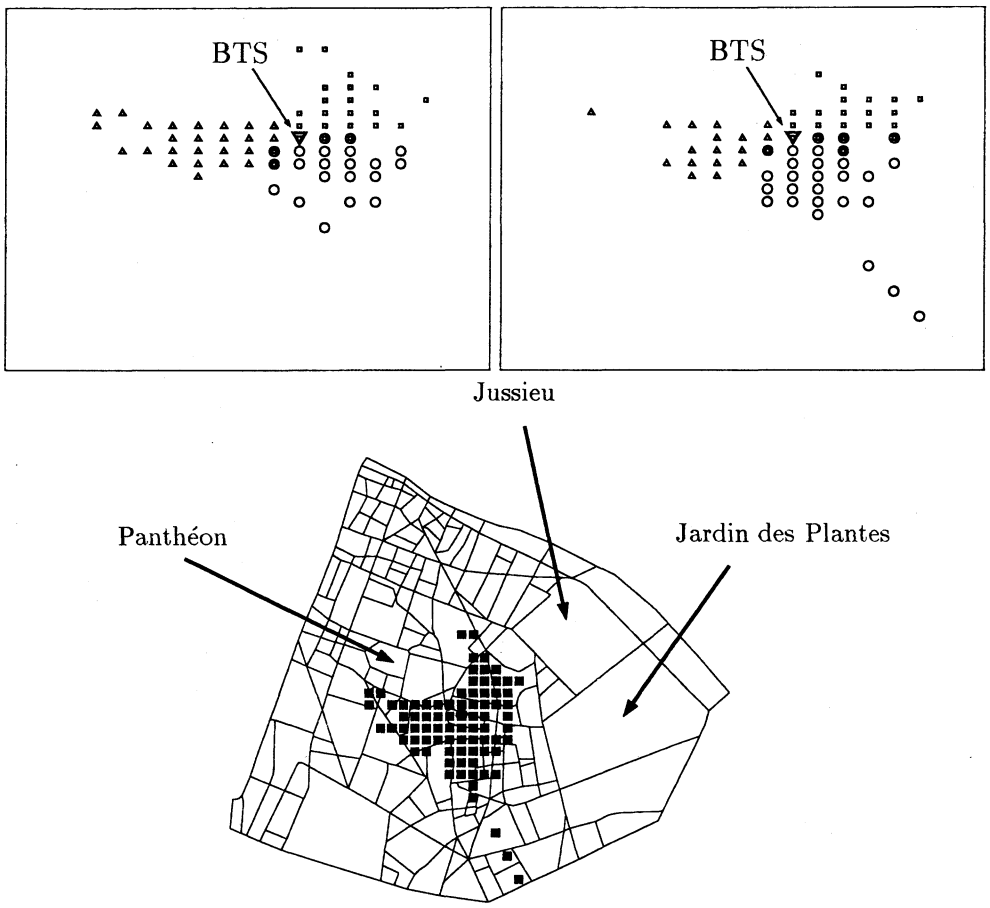


Figure 1.6 – Exemple de cellules standard. Les deux graphiques supérieurs représentent les cellules de chacune des trois BTS 900 MHz (gauche) et 1800 MHz (droite) d'un même site. Le site des BTS est signalé par un triangle pointe vers le bas. Les cellules sont les unions des mailles de 50 mètres de côté dont les centres sont figurés par des triangles pointes vers le haut, des carrés ou des cercles. Le dernier graphique donne une idée de la superficie de l'union des six cellules grâce à la comparaison avec le cinquième arrondissement de Paris. Chaque carré noir vaut pour une maille desservie.

Nous proposons pour commencer de commenter la Figure 1.6. On y découvre les zones de desserte de trois BTS 900 MHz et trois BTS 1800 MHz, respectivement, situées sur un même site parisien.*

On constate notamment que les celluless ont bien une (vague) forme de pétale comme annoncé plus tôt. Les cellules comptent respectivement 21, 21, 27 mailles pour les BTS 900 MHz et 27, 28 et 18 mailles pour les BTS 1800 MHz, soit un total de 69 et 73 mailles, respectivement. Les

*La position apparente du site dans la figure ne correspond pas à sa position réelle pour des raisons de confidentialité.

superficies de couverture sont ainsi du même ordre que la surface d'un carré de 400 mètres de côté.

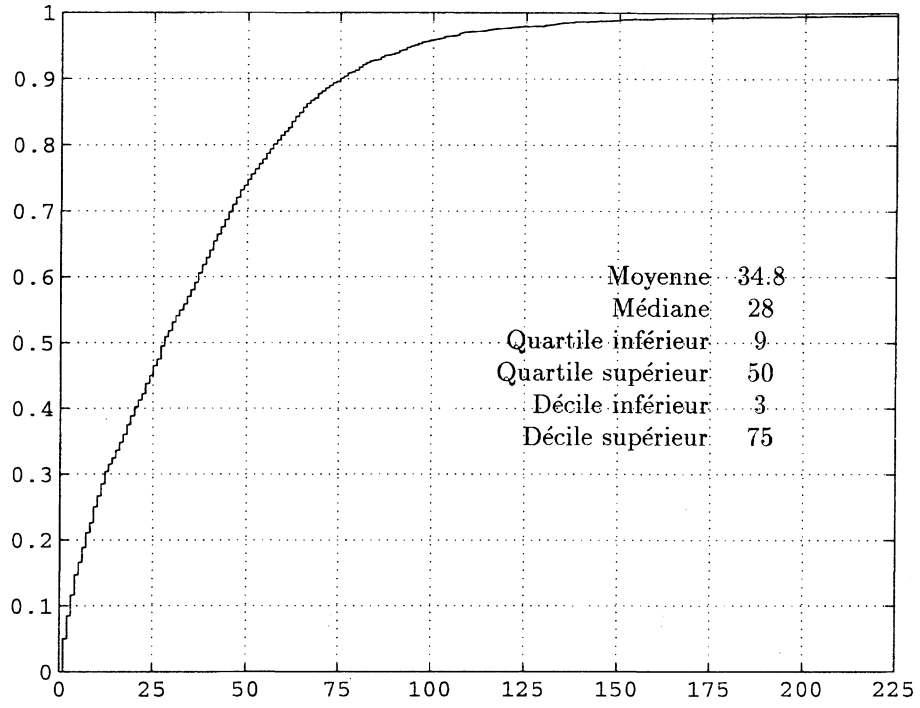


Figure 1.7 – Fonction de répartition empirique du nombre de mailles par cellules sur Paris.

Qu'en est-il plus précisément de la superficie standard d'une cellule ? La fonction de répartition empirique du nombre de mailles par cellule sur Paris donne quelques éléments de réponse, voyez la Figure 1.7.

Ainsi, 90% des cellules couvrent une surface dont l'ordre est inférieur à la surface d'un carré d'environ 435 mètres de côté. En revanche, 10% des cellules couvrent une surface dont l'ordre est inférieur à la surface d'un carré de 85 mètres de côté. Une cellule sur deux a une surface comprise entre celles de carrés de côtés respectifs 150 et 350 mètres environ.

1.5. Méthodes de localisation de trafic

1.5.1. Une méthode antérieure de localisation par triangulation

Des données de trafic et de recouvrement cellulaire telles que celles que nous venons de présenter et de commenter dans les sections précédentes permettent-elles à elles seules d'élaborer une méthode de localisation raffinée du trafic ? La réponse est affirmative, grâce au travail de l'équipe FTR&D/DMR/OIP de Belfort qui consacre une partie de ses activités à ce problème ardu. La méthode élaborée par ses membres requiert l'utilisation de données complémentaires des données Cigale auxquelles nous n'avons pas eu accès et que nous décrivons succintement

plus bas. Nous allons la présenter brièvement et la commenter superficiellement.*

Nous avons bien insisté dans le Prélude sur la différence entre la localisation de trafic (ce à quoi nous nous consacrons ici) et la localisation d'un mobile en cours d'appel. Il existe néanmoins une certaine parenté entre les deux problèmes pour des configurations de trafic peu mobile. La méthode que nous allons résumer ci-dessous est conçue pour analyser le trafic sur *une* cellule particulière. Elle peut-être appliquée au cas par cas pour répondre à des demandes précises sur des cellules problématiques et à l'échelle de l'heure. Elle n'est en revanche pas adaptée à un raffinement global de localisation.

Le principe est le suivant. Soit une certaine cellule d'intérêt associée à la BTS b. Les données de trafic horaires sont naturellement disponibles à l'échelle de la cellule. On peut par exemple en disposer *via* une extraction Cigale. On peut aussi procéder à une observation encore plus méticuleuse en conservant la trace des mesures de champ effectuées par tous les mobiles desservis par la BTS. Les mobiles émettent en effet toutes les 480 millisecondes un rapport de réception des six meilleures serveuses, *i.e.* des six meilleurs champs mesurés par lui à cet instant. La procédure peut être décomposée comme suit.

- Pour tous les mobiles desservis par b et toutes les 480 ms, enregistrement des six meilleurs champs reçus par lui.
- Application d'une méthode de triangulation afin de produire une distribution de probabilité de présence à la maille pour chaque mobile. La dite méthode de triangulation repose notamment sur la comparaison des champs reçus avec les champs prédits à la maille par simulation Parcell©(voir la Section 1.4.2). Elle prend à ce titre en compte la topographie des lieux.
- Basculement des probabilités individuelles de présence à une probabilité collective de présence, qui constitue le résultat final.

La méthode a été testée par création artificielle de *hot-spots*, *i.e.* de surcharges de trafic en des lieux précis par des mobiles tests identifiables. Il semble que les résultats soient satisfaisants.

On peut noter que

- cette méthode est d'autant plus appropriée que le trafic est peu mobile ;
- plus le procédé de triangulation est efficace, plus la localisation est précise ;
- cette méthode est sans doute trop coûteuse en temps de calcul pour être appliquée simultanément à un grand nombre de cellules ;
- cette méthode ne permet pas de dégager d'éléments descriptifs intrinsèques reliant le trafic localisé aux lieux d'observation.

1.5.2. Des données de trafic agrégées

Une caractéristique des données de trafic que nous avons décrites dans la Section 1.2 n'a pas été accentuée : nos données sont par nature *agrégées*. Autrement dit, toute observation liée à un appel est rapportée à la BTS qui gère cet appel, *i.e.* de façon équivalente à la cellule de cette BTS.

Ce phénomène d'agrégation des données est courant dans le domaine de la statistique spatiale. Le lecteur trouvera de nombreux exemples dans la monographie de Cressie (1993). Ceux-ci

*Ce qui suit s'appuie sur l'exposé qu'Olivier Veyrunes nous en a fait. Je profite de cette occasion pour le remercier encore et le prier d'excuser les approximations qui vont suivre.

correspondent souvent à la situation où les observations sont rapportées à une entité géographique (la zone d'agrégation) qui contient le point d'observation. C'est en particulier généralement le cas pour des données épidémiologiques ou environnementales (suivant la hiérarchie schématique ville/agglomération/département/région pour la France).

Les procédures standard de lissage comme le krigeage ou autres méthodes non paramétriques (voir Cressie 1993) ne prennent pas en compte la nature agrégée des observations. Elles sont pourtant parfois appliquées, moyennant le choix d'un point par zone d'agrégation (typiquement, son centre de gravité). Chaque jeu d'observations sur une zone est alors attribué au dit point et la procédure utilisée normalement. Brillinger (1990) a mis en évidence les problèmes soulevés par de telles pratiques. Plus tard, Müller et al. (1997) ont adapté une version d'estimation par moindres carrés localement pondérés avec le même souci d'affranchissement du choix d'un point de référence par zone d'agrégation, puis l'ont appliquée à des données de diagnostic de SIDA dans le comté de San Francisco aux Etats-Unis.

Nous préférons quant à nous à ces méthodes de lissage un modèle de régression associé à une méthode pratique d'estimation de la fonction de régression. Nos motivations, ainsi qu'une évocation de la procédure d'estimation, suivent.

1.5.3. Heuristique pour une méthode originale

Notre intention est d'enrichir la base de données de trafic dont nous disposons d'une base de données de nature sociologique, démographique et culturelle. Nous allons en effet fonder notre approche du problème sur une confrontation de ces données d'origines différentes. Le but ultime est *l'explication* des données de trafic par les données socio–démographiques et culturelles.

Heuristiquement, le trafic local dépend largement, outre de l'horaire d'observation, de la nature du lieu d'observation. On s'attend à un trafic particulier selon que l'on considère une zone résidentielle ou commerciale, un quartier d'affaires ou une zone étudiante par exemple. En particulier, le trafic $Y_h(b)$ géré par la BTS b au cours d'un certain laps de temps de l'heure h peut s'expliquer par la nature sociologique, démographique et culturelle de la cellule sectorielle que dessert b, résumée par le vecteur réel $X(b)$. On exprime ceci grâce au modèle de régression

$$Y_h(b) = f^\star_{\mathrm{m}}\Big(X(b)\Big) + f^\star_{\mathrm{se}}\Big(X(b)\Big)\, e$$

pour deux fonctions de regression de moyenne et écart-type $f^\star_{\mathrm{m}}$ et $f^\star_{\mathrm{se}}$ et un bruit e centré réduit conditionnellement à $X(b)$.

Cette expression fait apparaître que nous nous intéresserons particulièrement aux moyennes et écarts-types des quantités de trafic. Ce choix s'impose d'une part parce qu'on ne peut pas raisonnablement proposer un modèle homoscédastique (*i.e.* où $f^\star_{\mathrm{se}}$ est une constante) : le trafic sur deux zones ne présente pas nécessairement la même variabilité. Il nous semble par ailleurs que ces deux informations sur plusieurs plages horaires témoignent simplement et néanmoins fidèlement de la nature du trafic.

Pour reprendre la discussion informelle là où nous l'avons laissée, le couple moyenne, écart-type est bien sûr plus révélateur que la seule moyenne. Il se prête à une interprétation plus fine : une moyenne modérée ne correspond par exemple pas à une situation unique, et la valeur de l'écart-type associé est informative : s'il est faible, c'est que le trafic est régulièrement modéré; s'il est important, c'est que le trafic est très variable. Ce complément d'information peut être d'une aide précieuse pour la densification de réseau (l'ajout de BTS) ou même la planification

d'un nouveau réseau : on ne couvre pas de la même façon une zone à trafic très variable ou au contraire peu variable, et ce selon la variabilité des autres zones mitoyennes susceptibles d'être desservies par la même cellule.

Finalement, cette approche permet bien de raffiner la localisation du trafic, dans la mesure où, une fois estimée $f_{\mathrm{m}}^{\star}$ et $f_{\mathrm{se}}^{\star}$, il suffit de connaître le vecteur $x(Z)$ d'une zone Z quelconque pour pouvoir lui associer un couple de moyenne, écart-type caractéristiques de son trafic. Ceci s'applique en particulier à des zones Z qui ne sont pas des cellules sectorielles, *i.e.* à des zones pour lesquelles on ne dispose d'aucune observation de trafic (puisque par définition, les données de trafic sont relatives aux cellules).

1.5.4. Mise en garde

Il faut cependant se garder de penser que l'association d'un couple moyenne, écart-type à une zone est *valable* pour n'importe quelle zone. Elle est certes *possible*, mais la taille de la zone ne devrait pas être *sensiblement différente* de celles des cellules sur lesquelles repose la procédure d'estimation. Ce point demande évidemment à être éclairci : d'une part, la surface des cellules peut varier largement, voyez encore la Figure 1.7 ; d'autre part l'expression « sensiblement différente »est intentionnellement peu précise. En l'absence d'observations sur des zones de tailles « sensiblement différentes » de celles des cellules, la conclusion de cette mise en garde est que la procédure statistique que nous allons proposer, aussi satisfaisante soit-elle, devra toujours beaucoup à une utilisation réfléchie.

1.5.5. Esquisse de la méthode originale

Nous avons rencontré à plusieurs reprises des ingénieurs de Orange France issus des départements d'ingénierie et commercial. Il est apparu que les deux points suivants leur tenaient particulièrement à cœur :

- lisibilité aisée des résultats de la procédure de raffinement ;
- mise en évidence d'un nombre réduit d'indicateurs socio–démographiques et culturels pour toute zone, dont les valeurs soient assez révélatrices du trafic associé. A titre de comparaison, treize types géographiques ont été désignés comme représentatifs de l'ensemble des situations possibles sur le territoire français.

Le souhait de lisibilité est par exemple satisfait pour des fonctions de régression constantes par morceaux sur les parties d'une partition, *i.e.* s'écrivant

$$\left(f_{\mathrm{m}}^{\star}(x), f_{\mathrm{se}}^{\star}(x)\right) = \sum_{k=1}^{K} (m_k, \sigma_k) \mathbb{1}\{x \in \tau_k\} \tag{1.5}$$

pour une partition $\tau = (\tau_k)_{1 \leq k \leq K}$ à K pièces de l'espace $\mathcal{X}$ des indicateurs socio-démographiques et culturels et K couples de moyenne et écart-type (m_k, σ_k). Ici, les quantités d'intérêt sont le nombre K de parties de la partition τ, cette même partition et les couples de moyenne et écart-type. Le Chapitre 5 étudie de façon approfondie les propriétés asymptotiques d'un estimateur de ces quantités pour des observations éventuellement dépendantes sous certaines hypothèses assez faibles. Dans le Chapitre 7, nous nous intéressons particulièrement à l'estimation du nombre de pièces de la partition. C'est un problème ardu intrinsèquement captivant et pratiquement intéressant. Ce nombre représenterait ici un nombre de trafics types permettant de décrire fidèlement l'ensemble de tous les trafics possibles.

Ces souhaits nous ont incités à opter pour une procédure d'estimation des fonctions de régression fondée sur des arbres binaires, et plus particulièrement sur la solution CART.* En effet,

- les fonctions de régression produites par CART sont facilement lisibles et interprétables puisqu'elles prennent la forme proposée en (1.5) ;
- les arbres fournissent naturellement une hiérarchisation des variables explicatives socio-démographiques et culturelles par pertinence décroissante en termes d'explication des données de trafic ;
- ils peuvent être combinés* pour produire un meilleur régresseur en termes d'adéquation aux observations et de prédiction, au prix de la dégéresecence de la structure sous-jacente d'arbre en une structure peu lisible. En particulier, les fonctions de régression associées ne sont plus présentables sous la forme proposée en (1.5).

1.6. Ce qu'il faut retenir*

Le problème initial qui nous a été soumis est de proposer une méthode de raffinement de localisation du trafic téléphonique mobile en zone urbaine, soit sur Paris pour ce premier travail (voir le Prélude). Notre démarche statistique s'appuie sur des jeux de données dont quatre sont issus de France Télécom. Il s'agit de trois jeux de données relatifs au trafic et d'un jeu relatif au recouvrement cellulaire. Les deux autres jeux sont de nature socio–démographique et culturel. On les présente dans l'Interlude 2.

Les données de trafic

Nous nous sommes procurés trois jeux de données dissemblables en termes de contenu et d'utilité. Ce sont les relevés CRA, Cigale et HC2.

Les relevés CRA† sont prévus à l'origine à des fins de facturation des usagers. Ils sont insuffisants, en particulier parce qu'ils donnent des informations de localisation restreintes aux instants où les appels sont initialisés.

Au contraire, les relevés Cigale† sont exhaustifs : ils conservent la trace de tous les échanges† entre tous les mobiles† et les trois MSC† pour lesquels on a obtenu ces relevés pour une journée de février 2002. Ces relevés permettent notamment de calculer la durée effective de chaque appel. On peut ainsi étudier statistiquement les durées d'appel : une ébauche de résultats est fournie en Section 1.3. On peut surtout grâce à eux calculer, pour chaque BTS† et à chaque instant, les nombres d'appels en cours, ou bien le nombre cumulé d'échanges pris en charge jusqu'à cet instant. Ce nombre permet quant à lui d'évaluer la quantité de trafic écoulée (en Erlang†, voir Section 1.2.2) par chaque BTS.

Finalement, le relevé HC2† donne pour chaque BTS la seconde plus grande quantité maximale de trafic écoulée en une heure sur chaque jour d'une semaine, pour les semaines des mois de mars à juillet 2002. Ce relevé est le seul des trois présentés ici qui soit conçu à l'origine à des fins de description quantitative du trafic écoulé.

*Celle-ci est présentée dans le Chapitre 3.

*Grâce par exemple aux procédures de Bagging et Boosting présentées elles aussi au Chapitre 3.

*Le lecteur pourra consulter le Glossaire en Annexe pour trouver une définition des acronymes et mots-clefs indiqués par le signe †.

Nous décidons de décrire le trafic sur des plages de temps de l'ordre d'une à deux heures, notamment parce que : *primo*, le trafic est assez stationnaire à cette échelle de temps ; *secundo*, une description à l'échelle de l'heure n'exclut pas une description à l'échelle de la journée ni de la semaine et au contraire, permet une meilleure compréhension des mécanismes quotidiens. Une telle approche se fonde essentiellement sur les relevés Cigale. Nous faisons ce choix bien que les ceux-ci soient peu pratiques à récolter. Une simple routine automatique permettrait en effet de les récupérer au niveau des MSC sous un format adéquat (en particulier moins volumineux que le format standard).

Finalement, la méthode que nous développons est suffisamment *flexible* pour que l'on puisse aussi l'appliquer au jeu de données Cigale complet (*i.e.* sur la journée entière) et aussi aux données hebdomadaires HC2.

Les données de recouvrement cellulaire

Elles sont le complément indispensable des données de trafic. L'estimation des cellules† en termes de mailles†, *i.e.* de carrés élémentaires de côté 50 mètres, est réalisée par simulation grâce à des outils spécifiques de France Télécom R&D. Cette estimation repose partiellement sur le choix d'un paramètre de seuillage. Nous avons choisi une valeur de ce seuil qui assure un recouvrement simulé à plus de 95% des 28 431 mailles de Paris par une cellule au moins du réseau Orange†. Les cellules sont de tailles variables, voyez la Figure 1.7 pour la fonction de répartition empirique du nombre de mailles par cellule.

Heuristique de notre méthode

Une des particularités des données de trafic que nous avons récoltées tient à leur nature *agrégée* : tout appel observé à un certain instant se voit associer, non pas sa localisation, mais la cellule de desserte de la BTS gérant sa connexion à cet instant. D'un autre point de vue, la BTS se voit associer tous les appels qu'elle gère à cet instant. Cette particularité (assez commune en statistique spatiale) interdit l'application des méthodes standard de lissage comme le krigeage. Des solutions ont été élaborées, mais nous leur préférons un modèle de régression et une procédure pratique d'estimation.

Nous partons du constat que le type de trafic sur une zone Z dépend sensiblement de la nature socio–démographique et culturelle de la dite zone (et de l'heure d'observation, si les données sont assemblées sur une plage horaire étendue). Nous nous proposons d'expliquer les quantités de trafic $Y(Z)$ écoulées sur Z à l'aide d'une description $X(Z)$ en ces termes de Z. Le modèle de regression correspondant peut s'écrire

$$Y(Z) = f^{\star}_{\mathrm{m}}\Big(X(Z)\Big) + f^{\star}_{\mathrm{se}}\Big(X(Z)\Big)\, e$$

pour un bruit e centré réduit conditionnellement à $X(Z)$.

Il apparaît que nous nous intéressons particulièrement au couple moyenne, écart-type des quantités de trafic. C'est d'une part parce qu'un modèle homoscédastique (cas où $f^{\star}_{\mathrm{se}}$ est constante) n'est pas justifié ; d'autre part parce que ce couple témoigne simplement et fidèlement de la nature du trafic.

Une fois estimées les fonctions de régression $f^{\star}_{\mathrm{m}}$ et $f^{\star}_{\mathrm{se}}$, il sera possible de prévoir la nature du trafic (au sens de moyenne, écart-type) de toute zone Z dont on connaît la description socio-démographique et culturelle. Il faudra néanmoins procéder avec circonspection, dans la mesure

où la taille de la dite zone ne devrait pas différer sensiblement de celles des cellules qui ont permis l'estimation.

1.7. English summary

We shall tackle the original problem we are proposed, *i.e.* elaborate a method that provides refinement of the localization of the mobile telecommunication traffic in Paris, with a statistical approach. Thus, some datasets are needed. Four of them are provided by France Télécom and concerned with traffic measurements or cellular covering, while the two others left deal with local socio–demographic and cultural characteristics throughout Paris. The latter are introduced and carefully described in the next Interlude 2. We present below the four telecommunication datasets, followed by an informal description of the method we shall propose further.

Traffic databases

A major difficulty of our work has been to obtain a traffic database which was to be both relevant and handleable. We finally worked with three different traffic databases: the variety of the latter illustrates that difficulty.

The first dataset is called CRA dataset. It provides for each mobile phone throughtout a day the time when any call was originated, its duration and the cellular localization at initialization. Localization at initialization is naturally insufficient for our purpose, since mobile calls are usually not handled by a sole BTS.

On the contrary, the Cigale dataset is exhaustive: it contains a large amount of informations on any communicating mobile phone, among which its instantaneous cellular localization. It allows the computation of the instantaneous number of calls handled by any BTS, and also the quantity of traffic (in Erlang) that any BTS deals with.

Finally, the HC2 dataset provides weekly second larger daily maximal quantities of traffic throughout an hour which any BTS deals with (see the formal definition given by (1.2)).

Now, in virtue of

- the (almost) stationarity of the traffic on intervals of one or two hours, and
- the fact that the description of the traffic on such time intervals also can lead to daily and even weekly descriptions (which is obviously not the case from daily to hourly descriptions),

we decide to focus on the localization of the traffic on one to two hours time intervals. Forwardly, this makes the Cigale dataset the most important one, even though it is the less easily extractable and handleable dataset. However, this is not a serious drawback, since minor changes in the standard protocol could automatically yield the needed data under an appropriate format.

Finally, the method we shall elaborate does not crucially rely on the latter focus. So, we shall also apply it to the Cigale dataset through the whole day and to the HC2 dataset.

Cellular covering database

The cellular covering database is the indispensable complement to the former traffic datasets. It provides an estimation of all the cells that cover Paris. Indeed, Paris is divided into 28 431 disjoint elementary squares of mutual length 50 meters, called meshes. Performed simulations

and choice of a threshold yield a description of each cell in terms of covered meshes. The threshold is chosen so that at least 95% of the meshes are covered. The empirical cumulative distribution function of the number of meshes that compose the cells is given in Figure 1.7.

Heuristics of our method

An important feature of our traffic datasets is its *aggregated* nature: in words, the data do not carry exact geographical coordinates of the location where they were obtained, hence our problem. Instead, all the observations related to the same BTS (or equivalently to the same cell) are lumped together in an aggregated quantity. One often meets such a feature in spatial statistics, particularly when coping with epidemiological or environmental data (with the classical hierarchy of aggregation town/zip code area/county/state in the United States), see for instance Cressie (1993). The most common methods in spatial statistics require the choice of a peculiar point in the aggregation zone to which assign the data and then forget the aggregated nature of the original dataset. Various authors have addressed this problem since (Brillinger 1990). In the same spirit, Müller et al. (1997) elaborated a locally weighted least squares procedure in order to get rid of the choice of a point for each aggregation area. However, rather than such smoothing methods, we prefer a regression framework as we shall explain briefly hereafter.

The method we shall construct relies on the following remark: the traffic upon a zone Z crucially depends on the the socio–demographic and cultural characteristics of the zone. Thus, letting $X(Z)$ and $Y(Z)$ denote respectively the vector of the latter characteristics and the traffic associated with Z (on a certain time interval) understood as random variables, one should be able

- *primo* to understand the structural relationship between the response and the measured variables, *i.e.* between $Y(Z)$ and $X(Z)$;
- *secundo* to predict the response variable $Y(Z)$ on the basis of $X(Z)$ for some zone Z which is not a cell (hence particularly, for which no traffic observation is available). Nevertheless, one should not apply that prediction rule to any zone Z: the latter should have a size similar to the typical size of a cell involved in the estimation procedure.

We propose the following regression model

$$Y(Z) = f^{\star}_{\mathrm{m}}\Big(X(Z)\Big) + f^{\star}_{\mathrm{se}}\Big(X(Z)\Big)\, e$$

where e is a noise conditionally centered and variance one given $X(Z)$. The regression functions $f^{\star}_{\mathrm{m}}$ and $f^{\star}_{\mathrm{se}}$ are respectively regression functions of the mean and of the standard error. Indeed, an homoscedastic model (*i.e.* a model as above with $f^{\star}_{\mathrm{se}}$ constant) is inappropriate for our purpose. Besides, we think that the mean and the standard error of the traffic $Y(Z)$ provides a simple yet relevant description of the nature of the typical traffic associated with Z.

2

Interlude : ILOTS15, Contousllots et SIRENE*

Résumé

Nous consacrons cet interlude à la présentation de l'outil de description socio-démographique et culturelle de Paris dont nous nous sommes pourvus pour mener à bien notre étude. Cet outil est le fruit de la fusion de deux bases de données classiques de la statistique économique et sociale française. Une section est dédiée à chacune d'elles. Une troisième synthétise les informations les plus importantes. Un résumé en anglais clôt l'interlude.

Abstract

This interlude is devoted to the presentation of a sociological, demographical and cultural description tool of Paris we shall need for our study. It stems from the melting of two classical french statistical databases. Each of them is described in a section, while a third one summarizes the most important points. An english résumé concludes the interlude.

*Cet interlude s'inspire des documentations fournies par l'INSEE.

Au menu

2.1. **Introduction** 87
2.2. **Les bases de données ILOTS15 et Contoursllots** 87
2.3. **Le répertoire SIRENE** 89
2.4. **Ce qu'il faut retenir** 94
2.5. **English summary** 95
2.6. **Annexe** 96

2.1. Introduction

La méthode statistique que nous avons élaborée en guise de réponse au problème initial repose fondamentalement sur une description sociale, démographique et culturelle de Paris aussi révélatrice que possible de la réalité sociale, démographique et culturelle. Il a fallu déterminer en quoi pouvait tenir une telle description, sans perdre de vue naturellement, d'une part les objectifs finaux et d'autre part, les contraintes purement techniques de faisabilité liées aux types de bases de données existantes sur le sujet. Une fois cernée la nature d'une base de données satisfaisante, nous l'avons construite par croisement de deux bases classiques appelées ILOTS15 et SIRENE. Cet interlude est dédié à la présentation du jeu de données que nous avons conçu par le biais de ces deux bases mentionnées plus haut.

2.2. Les bases de données ILOTS15 et Contoursllots

Une description locale de la population et des logements

La base de données ILOTS15 est le fruit de l'exploitation principale du recensement français du mois de mars 1999 par l'Institut National de la Statistique et des Etudes Economiques (*INSEE* pour la suite). Elle recense 24 données socio-démographique relatives :

- à l'ensemble de la *population* par sexe et tranche d'âge (population féminine exclusivement, masculine exclusivement, totale, sans considération d'âge et par tranches d'âges de 0 à 19, 20 à 39, 40 à 59, 60 à 74, 75 ans et plus) ;
- à l'ensemble des *logements* par catégorie (*principal*, *secondaire*, *occasionnel* et *vacant*) ;
- au *nombre de personnes* vivant dans les résidences principales.

Un *logement* est défini comme un local séparé et indépendant utilisé comme lieu d'habitation. Il est *principal* dès lors qu'il est occupé de façon permanente et à titre principal par un ménage. Il est *secondaire* lorsqu'il est occupé de façon temporaire, *occasionnel* s'il est utilisé une partie de l'année pour des raisons professionnelles et *vacant* s'il n'a pas d'occupants et qu'il n'est pas un logement secondaire.

L'ensemble de ces données est fourni à l'échelle de *l'îlot*. L'îlot est une zone de surface réduite, aussi homogène que possible en termes socio-démographiques (le zonage est fixé en partenariat avec les collectivités locales), peuplée de moins de 800 habitants. En milieu urbain (cas qui nous intéresse), il s'agit de la plus petite surface délimitée par des voies publiques et/ou privées : l'îlot correspond ainsi à ce que l'on entend ordinairement par « pâté de maisons ».

La plupart des communes de plus de 800 habitants (parfois moins) ont été découpées en îlots pour les besoins de la collecte du recensement de la population. Pour toutes les communes

de plus de 10 000 habitants et pour la plupart des communes des agglomérations de 50 000 habitants, les plans des découpages en îlots sont disponibles sous forme numérisée.

Les contours sont disponibles dans la base de données ContoursIlots. Nous donnons à titre illustratif le nombre d'îlots par arrondissements parisiens dans le Tableau 2.1 et une vue globale des îlots parisiens dans la Figure 2.1. On vérifie bien que le nombre d'îlots dont est constitué un

Arrondissement	1 er	2 ème	3 ème	4 ème	5 ème	6 ème	7 ème	8 ème
Nombre d'îlots	148	141	113	177	201	176	221	251
Arrondissement	9 ème	10 ème	11 ème	12 ème	13 ème	14 ème	15 ème	16 ème
Nombre d'îlots	193	164	267	284	334	302	444	496
Arrondissement	17 ème	18 ème	19 ème	20 ème	Total du nombre d'îlots			
Nombre d'îlots	395	399	334	385	5 425			

Tableau 2.1 – Nombre d'îlots par arrondissement à Paris.

arrondissement est très fortement corrélé à sa superficie.

Commande d'une extraction de ILOTS15

Nous avons acquis auprès de l'INSEE une extraction ILOTS15 sur la totalité de Paris. Nous l'avons complétée de l'extraction ContoursIlots correspondante, *i.e.* que nous disposons à l'échelle de Paris dans son ensemble :

- des relevés de population par tranche d'âges et par sexe,
- du nombre de logements par catégorie,
- du nombre de personnes vivant dans leur résidence principale,
- et cela pour tous les îlots parisiens – dont nous connaissons par ailleurs les contours.

Ceci constitue la description en termes de populations et de logement qui est le premier des deux aspects de la description socio-démographique et culturelle que nous souhaitons effectuer de Paris. Elle n'est pour l'instant encore que très partielle et pour la compléter, nous avons recours à une extraction SIRENE.

Règle de proportionnalité

Nous proposons pour conclure une méthode d'approximation de la description sociologique au sens de ILOTS15 d'une zone quelconque de Paris qui ne soit pas nécessairement un îlot ou une union d'îlots. Cette méthode est fondée sur une règle naturelle de proportionnalité.

Soit Z une zone quelconque dans Paris et soit $I_1, \ldots, I_r$ les îlots que cette zone recouvre au moins partiellement. Pour chaque $1 \leq k \leq r$, on note $x(I_k)$ le vecteur des données ILOTS15 de l'îlot I_k et α_k le rapport de la superficie de $Z \cap I_k$ avec la surface totale de l'îlot I_k. On définit alors le vecteur ILOTS15 estimé $x(Z)$ de la zone Z comme la somme pondérée

$$x(Z) = \sum_{k=1}^{r} \alpha_k \, x(I_k).$$

En pratique, la zone dont on veut calculer le vecteur ILOTS15 estimé est décrites en termes de mailles, *i.e.* en termes de zones élémentaires carrées comme introduites dans le Chapitre 1 (cette décomposition est automatique pour les cellules de desserte des BTS). Nous procédons

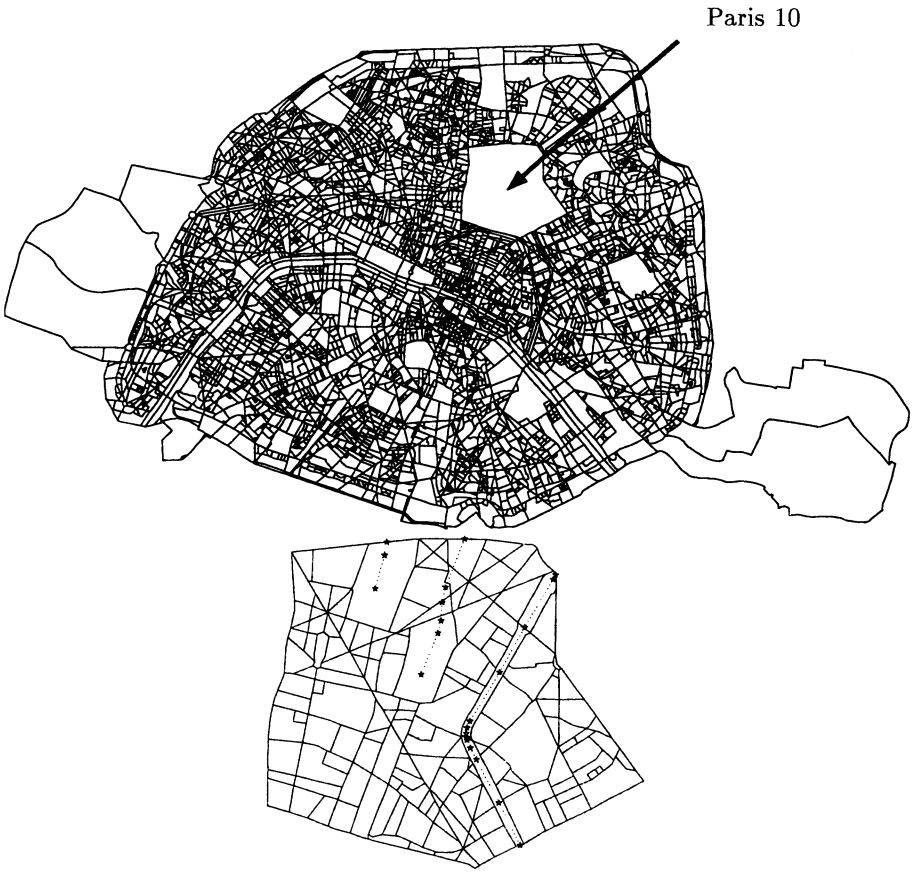


Figure 2.1 – Les îlots parisiens. En haut, tout Paris, 10 ème arrondissement hormis. En bas, le 10 ème arrondissement ; les lignes brisées à étoiles représentent respectivement, de gauche à droite, la Gare du Nord, la Gare de l'Est et le Canal Saint Martin.

alors à l'approximation suivante (qui est d'autant moins grossière que la longueur du côté de la maille est petit) : nous évaluons pour chaque îlot son centre de gravité et calculons sa surface ; puis nous décidons d'affecter à chaque maille (au prorata des surfaces) le vecteur ILOTS15 de l'îlot dont le centre de gravité est le plus proche de celui de la maille en question.

2.3. Le répertoire SIRENE

Le répertoire *SIRENE* (pour Système Informatique pour le Répertoire des ENtreprises et de leurs Etablissements) a été créé par décret en 1973. Sa gestion a été confiée à l'INSEE. Il enregistre l'état civil de toutes les *entreprises* et leurs *établissements* situés en métropole, dans les DOM (Guadeloupe, Guyane, Martinique et Réunion) et à Saint-Pierre et Miquelon et ce, quels que soient leur forme juridique et leur secteur d'activité. Les entreprises étrangères qui ont une représentation ou une activité en France y sont également répertoriées.

Les deux mots-clefs sont ici *entreprise* et *établissement*. Une entreprise est une unité économique juridiquement autonome organisée pour produire des biens ou des services. Un établissement

est une *unité de production* localisée géographiquement, individualisée mais dépendant juridiquement de l'entreprise-mère. Il constitue à ce titre *le niveau le mieux adapté à une approche géographique de l'économie.* Il est relativement homogène et son activité principale apparaît proche du produit.

L'INSEE attribue à chaque entreprise un identifiant numérique SIREN et à chaque établissement un identifiant numérique SIRET. La base de données SIRENE reprend, pour les seuls entreprises et établissements administrativement actifs, les informations contenues dans le répertoire SIRENE en les restructurant et en les complétant. SIRENE rassemble ainsi des informations économiques et juridiques sur plus de 6,7 millions d'établissements et 5 millions d'entreprises appartenant à tous les secteurs d'activité.

Un état civil des entreprises et établissements

Le répertoire SIRENE récolte deux grands types d'informations sur les entreprises et les établissements :

- des données d'identification :
  - pour l'entreprise : le numéro SIREN ; les nom, prénom pour les personnes physiques ; le sigle, la raison sociale ou dénomination pour une personne morale ; le statut juridique ;
  - pour l'établissement : le numéro SIRET ; le statut de siège social ou non ; l'enseigne ; l'adresse ;
- des données de classification économique :
  - pour l'entreprise : le code d'Activité Principale Exercée (APE) ; l'importance de l'effectif salarié ; le chiffre d'affaires ;
  - pour l'établissement : comme au-dessus, chiffre d'affaires hormis.

Ces informations sont mises à jour quotidiennement. De nombreux organismes ont pour mission de déclarer à l'INSEE les inscriptions, radiations et modifications au répertoire. Ce sont en moyenne 10 000 modifications par jour qui sont enregistrées.

Commercialisation de SIRENE

L'INSEE commercialise la base de données sous la forme de fichiers dits NOTICES. Des extractions selon divers critères sont disponibles : c'est par ce biais que nous avons obtenu notre jeu de données SIRENE sous la forme d'une « NOTICES 30 ». Le tarif est *grosso modo* proportionnel au nombre d'établissements en sortie.

L'activité principale de l'établissement est classée suivant 60 divisions puis, à un niveau plus fin, suivant 696 postes. Ce sont respectivement les classifications *APET60* et *APET700* (*APET* pour Activité Principale exercée par l'ETablissement). La classification APET60 est déterminée d'après la norme Nomenclature d'Activités Françaises (*NAF*). Chacune des 60 *divisions* est elle-même raffinée en un certain nombre de *postes*, pour un nombre total de postes de 696. La codification est hiérarchique, dans la mesure où les deux premiers chiffres constituant le code APET700 (format `Chiffre Chiffre Chiffre Lettre`) codent au sens de l'APET60. Pour 13 exemples de divisions APET60 et 102 exemples de postes APET700, voir les Tableaux 2.5, 2.6 et 2.7.

Identification et adresse		
Numéro SIRET	32618432200013	56210768000018
Nom ou raison sociale	ACTION RIVE GAUCHE	SOC PETIT MARGUERY
Enseigne		AU PETIT MARGUERY
Numéro dans la voie	5	9
Type de voie	R	BD
Libellé de voie	DES ECOLES	DU PORT ROYAL
Ligne d'acheminement postal	75005 PARIS	75013 PARIS
Caractér. économiques de l'établissement		
Activité principale APET700	921J	553A
Effectif salarié par tranches	03	11
Nature (commerces par taille)	99	99

Tableau 2.2 – Informations d'intérêt des NOTICES 30. Exemples du cinéma Grand Action (5 rue des Ecoles, 75005 Paris) et du restaurant Au Petit Marguery (9 bd de Port Royal, 75013 Paris). Les activités 921J et 553A sont répertoriées *Projection de films cinématographiques* et *Restauration de type traditionnel* dans la classification APET700 – voir les Tableaux 2.5, 2.6 et 2.7 en Annexe. La Nature codifiée 99 correspond à des locaux de taille réduite.

Commande d'une extraction SIRENE

Nous avons commandé à l'INSEE une extraction SIRENE dont le fichier de sortie contient les informations d'intérêt telles que présentées dans le Tableau 2.2.

Dans un premier temps, nous avons préselectionné, *suivant des appréciations personnelles qualitatives*, 103 codes APET700 sur les 696. Ces codes sont énumérées dans les Tableaux 2.5, 2.6 et 2.7 en Annexe. Le Tableau 2.3 présente les codes APET60 qui apparaissent dans la liste, leur description, ainsi que le nombre de postes APET700 associés et le nombre d'établissements correspondants (enrichis des pourcentages approximatifs respectifs). La Figure 2.2 offre une visualisation des nombres d'établissements selon leur code APET60 ou APET700.

Le choix des 103 divisions APET700 de sélection a été guidé par le souci de ne prendre en compte qu'un nombre limité de codes (rappelez-vous que le tarif d'une extraction est à peu près proportionnel au nombre d'établissements en sortie) dont la variété reflète néanmoins la diversité telle qu'observée par un promeneur dans Paris. *En particulier, une telle sélection préliminaire ne pouvait être effectuée selon nos souhaits sur le seul code APET60.* La classification APET60 est cependant récupérable grâce à la hiérarchisation de codage expliquée plus haut.

Le Tableau 2.4 présente les codes APET60 classés par ordre décroissant de pourcentage d'occurence sur notre jeu de données. On observe que le degré de raffinement du découpage en postes APET700 des divisions APET60 n'est pas révélateur du nombre d'occurences par divisions APET60 dans notre extraction.

Géolocalisation : des adresses postales aux coordonnées spatiales

La dernière étape de la la procédure de fabrication de la base de données socio-démographique et culturelle de Paris dont on a voulu s'équiper est cruciale : il s'agit en effet de traduire les informations de localisation des établissements de nature *postale* en informations de localisation *par*

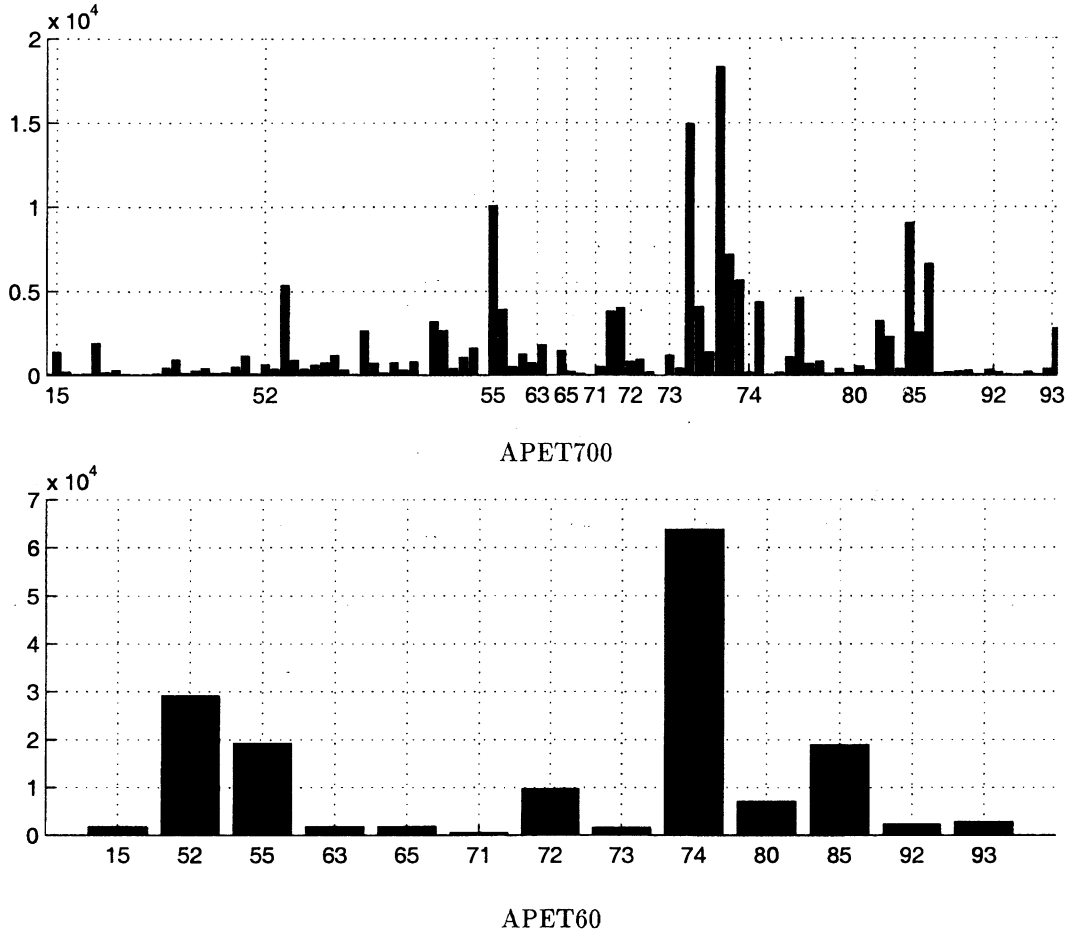


Figure 2.2 – Visualisation de la répartition du nombre d'établissements par APET700. Graphique haut : en abscisses, les APET700 rassemblés autour de l'APET60 dont ils dépendent ; en ordonnées, nombre d'occurences pour chaque APET700 dans notre jeu de données. Graphique bas : en abscisses, les APET60 ; en ordonnées, le nombre d'occurences pour chaque APET60 dans notre extraction (tel que présenté dans le Tableau 2.3.)

Code APET60	Description de l'APET60	Nombre d'APET700 associés	Nombre d'occurences
15	Industries alimentaires	4 (4%)	1 750 (1%)
52	Commerce de détail et réparation d'articles domestiques	38 (38%)	29 146 (18%)
55	Hôtels et restaurants	8 (8%)	19 242 (12%)
63	Services auxiliaires des transports	1 (1%)	1 818 (1%)
65	Intermédiation financière	5 (5%)	1 838 (1%)
71	Location sans opérateur	1 (1%)	529 (0%)
72	Activités informatiques	6 (6%)	9 773 (6%)
73	Recherche et développement	2 (2%)	1 581 (1%)
74	Services surtout aux entreprises	14 (14%)	63 759 (40%)
80	Education	7 (7%)	7 082 (5%)
85	Santé et action sociale	5 (5%)	18 934 (12%)
92	Activités récréatives	11 (11%)	2 270 (1%)
93	Services personnels	1 (1%)	2 864 (2%)
15		103	160 586
Totaux			

Tableau 2.3 – Les codes APET60 de la sélection. On précise pour chacun d'eux le nombre d'APET700 correspondants (et entre parenthèses, le pourcentage approximatif sur les 103 APET700 présents), le nombre d'occurences de chacun des APET60 dans notre extraction (avec entre parenthèses, le pourcentage approximatif sur les 160 586 établissments recensés).

coordonnées géographiques. Cette transformation est couramment appelée *géolocalisation*. Elle est effectuée par croisement avec une gigantesque base de données de géolocalisation de toutes les adresses parisiennes dont dispose Les Pages Jaunes, membre du groupe France Télécom.

Je remercie chaleureusement de nouveau Madame Bénédicte Cherbonnel de FTR&D/DMI/-LAN pour sa bienveillante intervention et son savoir-faire de géolocalisation.

Le fruit de l'extraction SIRENE

En résumé, après transformation des données originales issues du répertoire SIRENE, nous disposons à l'échelle de Paris dans son ensemble :

- pour tous les établissements parisiens dont l'activité principale codée au sens de l'APET700 fait partie des 103 activités selectionnées au préalable,
- de leur numéro SIRET, nom ou raison sociale, adresse postale,
- de leurs coordonnées géographiques,
- de leur activité principale APET700,
- de leur effectif salarié par tranches,
- et de leur nature (*i.e.* de la superficie de leur commerce).

Code APET60	74	52	55	85	72	80	93	92	65	63	15	73	71
% occurence	40	18	12	12	6	5	2	1	1	1	1	1	0
% nombre d'APET700	14	38	8	5	6	7	1	11	5	1	4	2	1

Tableau 2.4 – Les codes APET60 classés par ordre décroissant du nombre d'occurences dans notre extraction. La première ligne donne les codes APET60, la seconde les pourcentages d'occurences de chaque PAET60 sur notre jeu de données et la dernière, le pourcentage de postes APET700 pour chaque APET60.

2.4. Ce qu'il faut retenir

Notre approche du problème initial posé par FTR&D (voir le Prélude) nécessite une base de données socio-démographique et culturelle aussi révélatrice que possible de la réalité sociale, démographique et culturelle loacle de Paris. Nous en avons constitué une à partir de deux bases de données de l'INSEE† : ILOTS15† et SIRENE†.

ILOTS15 participe à l'ensemble en apportant à l'échelle de Paris tout entier :

- les relevés de population par tranche d'âges et par sexe,
- le nombre de logements† par catégorie (principal†, secondaire†, occasionnel† et vacant†),
- le nombre de personnes vivant dans leur résidence principale,
- et cela pour tous les îlots† parisiens – dont nous connaissons par ailleurs les contours.

Un îlot est *grosso modo* un « pâté de maisons ».

Le second pan de l'ensemble est issu du répertoire français SIRENE des entreprises† et établissements†. Par son biais, nous enrichissons la base, à l'échelle de Paris tout entier :

- pour tous les établissements dont l'activité principale† codée au sens de l'APET700† fait partie des 103 activités selectionnées au préalable,
- de leur numéro SIRET†, nom ou raison sociale, adresse postale,
- de leurs coordonnées géographiques,
- de leur activité principale APET700 (696 postes),
- de leur effectif salarié par tranches,
- et de leur nature (*i.e.* de la superficie de leur commerce).

On peut par ailleurs récupérer automatiquement la classification APET60† (60 divisions) de chaque établissement à partir du code APET700 afin de retrouver une répartition plus synthétique des établissements selon leur activité. Le Tableau 2.3 offre un aperçu des 13 codes APET60 associés aux 103 codes APET700 conservés.

Au final, la base offre une description *locale* assez fidèle – du moins on l'espère – de l'aspect social, démographique et culturel de Paris. On insiste sur son caractère local car il est crucial. Ainsi, pour une zone quelconque de Paris (*i.e.* qui ne soit ni une cellule, ni un îlot nécessairement), on peut estimer son caractère :

d'une part, en *évaluant* le nombre de personnes y vivant ainsi que le nombre de logements (par application de règles de proportionnalité de surface commune avec les îlots. Notez que cette règle est justifiée au regard de la faible superficie des îlots et de sa simplicité) ;

d'autre part en calculant *exactement* le nombre d'entreprises par APET700 situées dans la dite zone.

Cette base dont nous venons de décrire la construction est un rouage essentiel de la méthode

proposée en réponse au problème du raffinement de localisation de trafic.

2.5. English summary

As announced in the Prélude, we consider the problem submitted by FTR&D from a statistical point of view. The method we have elaborated requires a socio-demographic and cultural database revealing and relevant of the socio-demographic and cultural local nature of Paris. We constructed such a database on the basis of two databases of the INSEE (Institut National de la Statistique et des Etudes Economiques), namely the ILOTS15 and SIRENE (Système Informatique pour le Répertoire des ENtreprises et de leur Etablissements) datasets.

The ILOTS15 database's contribution consists of, for the whole town of Paris:

- the number of females, or males, or both females and males by ages,
- the number of housings by types (primary, secondary, occasional, unoccupied),
- the number of people living in their primary housing,
- all these on every *îlots* whose contours we precisely know.

An *îlot* has an accurate definition. It is roughly a small neighbourhood, which is often a single block in Paris.

The second contribution is due to the SIRENE database. It contains various informations on every parisian *enterprises* (an unit of economical organization) and *establishments* (a place of business which depends on a sole enterprise).

Now, we have on the whole Paris again:

- for any establishment whose activity is among the 103 APET700 activities we previously selected among 696 (see Table 2.3 for an overview of the 13 APET60 groups in which those 103 activities are classified by the INSEE),
- its official code SIRET, its name, its postal address,
- its geographical coordinates,
- its activity (through the corresponding APET700 code),
- its total staff,
- its size.

Finally, the above database gives a local description quite relevant –we hope so indeed – of the social, demographic and cultural nature of Paris. Let us emphasize its local property because it is essential. Thus, for any zone of Paris (*i.e.* not a cell or an îlot necessarily), we can estimate its character

first, by *estimating* the number of people that live in this zone and also the number of housings (with a proportion rule with respect to the common area wit any îlot),

second, by calculating *exactly* the number of establishments, for any APET700, that lie in the zone.

The latter database is a key-tool of the method we elaborated for localizing the traffic.

2.6. Annexe

Code APET60	Code APET700	Nombre d'occurences	Description du code APET700
15	158B	59	Cuisson de produits de boulangerie
	158C	1 360	Boulangerie et boulangerie-pâtisserie
	158D	192	Pâtisserie
	158K	55	Chocolaterie, confiserie
52	521A	82	Commerce détail (C.d.) de surgelés
	521B	1 904	Alimentation générale
	521C	166	Superettes
	521D	296	Supermarchés
	521E	60	Magasins populaires
	521F	23	Hypermarchés
	521H	41	Grands magasins
	521J	79	Autres Commerces de Détails non spécialisés
	522A	427	C.d. fruits et légumes
	522C	929	C.d. viandes et dérivés
	522E	131	C.d. produits maritimes
	522G	285	C.d. pain, pâtisserie, confiserie
	522J	418	C.d. boissons
	522L	117	C.d. tabac
	522N	135	C.d. produits laitiers
	522P	496	C.d. alimentaires spécialisés divers
	523A	1 161	C.d. produits pharmaceutiques
	523C	105	C.d. articles médicaux
	523E	632	C.d. produits beauté
	524A	372	C.d. textiles
	524C	5 376	C.d. habillement
	524E	897	C.d. chaussures
	524F	362	C.d. maroquinerie et articles voyage
	524H	599	C.d. meubles
	524J	757	C.d. équipement foyers
	524L	1 188	C.d. électroménager
	524N	323	C.d. quincaillerie
	524P	44	C.d. bricolage
	524R	2 661	C.d. livres, journaux, papeterie
	524T	722	C.d. optique et photographie
	524U	154	C.d. revêtements sols et murs
	524V	734	C.d. horlogerie et bijouterie
	524W	313	C.d. articles sports et loisirs
	524X	795	C.d. fleurs
	524Y	6	C.d. combustibles
	524Z	3 196	C.d. divers spécialisé
	525Z	2 663	C.d. biens d'occasion
	527A	397	Réparation d'articles en cuir

Table 2.5 – 42 activités APET700 sur les 103 préselectionnées (à suivre).

Code APET60	Code APET700	Nombre d'occurences	Description du code APET700
55	551A	1 071	Hôtels avec restaurants
	551C	1 599	Hôtels de tourisme sans restaurant
	552A	11	Auberges de jeunesse et refuges
	553A	10 080	Restauration de type traditionnel
	553B	3 915	Restauration de type rapide
	554A	495	Cafés tabac
	554B	1 250	Débits de boisson
	555A	738	Cantines, restaurants d'entreprises
63	633Z	1 789	Agences de voyage
65	651A	9	Banque centrale
	651C	1 466	Banques
	651D	219	Banques mutualistes
	651E	106	Caisses d'épargne
	651F	2	Intermédiations monétaires
71	714B	513	Location d'autre biens personnels et domestiques
72	721Z	3 830	Conseil en système informatique
	722Z	4 021	Réalisation de logiciels
	723Z	831	Traitement de données
	724Z	920	Banque de données
	725Z	171	Entretien, réparation informatiques
	726Z	0	Autres rattachées à l'informatique
73	731Z	1 173	R&D physiques et naturelles
	732Z	408	R&D humaines et sociales
74	741A	14 941	Activités juridiques
	741C	4 074	Activités comptables
	741E	1 376	Etudes de marchés et sondages
	741G	18 300	Conseils d'affaires et gestion
	741J	7 189	Administration d'entreprises
	742A	5 659	Architecture
	742B	172	Métreurs, géomètres
	742C	4 355	Ingénierie
	743A	43	Contrôle technique automobile
	743B	181	Analyses, essais, inspection techniques
	744A	1 109	Gestion supports publicitaires
	744B	4 623	Agences publicitaires
	745A	695	Cabinet recrutement
	745B	826	Travail temporaire

Table 2.6 – 37 activités APET700 sur les 103 préselectionnées (suite).

Code APET60	Code APET700	Nombre d'occurences	Description du code APET700
80	801Z	122	Ens. primaire
	802A	398	Ens. secondaire général
	802C	151	Ens. secondaire technique ou professionnel
	803Z	540	Ens. supérieur
	804A	318	Ecoles de conduite
	804C	3 251	Formation des adultes et continue
	804D	2 302	Autres Ens.
85	851A	386	Activités hospitalières
	851C	9 055	Pr. médicale
	851E	2 565	Pr. dentaire
	851G	6 633	Pr. auxiliaires médicaux
	853G	140	Crèches et garderies
92	921J	199	Projection de films cinématographiques
	922A	245	Activités de radio
	922B	296	Production de programmes télévisés
	922C	80	Diffusion de programmes télévisés
	923D	347	Gestion de salles de spectacles
	923H	200	Bals et discothèques
	923J	60	Autres spectacles
	925A	47	Gestion des bibliothèques
	926A	215	Gestion d'installation sportives
	927A	82	Jeux de hasard et d'argent
	927C	399	Autres activités récréatives
93	930D	2 830	Coiffure

Table 2.7 – 24 activités APET700 sur les 103 préselectionnées (fin).

3

Bagging and boosting CART regression trees: a user approach*

Résumé

Après qu'on a introduit le problème original et son contexte, les diverses données de télécommunication ainsi que les données explicatives socio-démographiques et culturelles dans les chapitres précédents, nous nous intéressons ici à la question pratique de l'estimation statistique du modèle de régression dont on a argumenté qu'il est approprié. La procédure qui s'est imposée est fondée sur des arbres de régression CART. Ceux-ci sont intéressants individuellement, mais aussi collectivement : les procédures d'agrégation Bagging et Boosting permettent en effet d'améliorer leurs performances individuelles.

Nous consacrons ces pages à une description de procédures CART, Bagging et Boosting classiques pour la régression homoscédastique. Nous proposons par ailleurs une adaptation originale de celles-ci dans un cadre de travail plus général de régression hétéroscédastique dans la perspective de leur prochaine application dans le Chapitre 4.

Abstract

We present in this chapter the main statistical procedure we shall apply when tackling the original problem of refinement of the localization of the mobile telecommunication traffic in Paris. The latter is discussed in Prélude, the telecommunication data are introduced and commented in Chapter 1, while the socio-demographic and cultural covariables are presented in Interlude 2. We have indeed argued that a simple regression model may provide interesting answers. We choose the practical estimation method of CART regression trees. Such regression trees can be considered individually or in committee. The Bagging and Boosting procedures allow to build such committees, which are known to perform better than single trees.

The following chapter is devoted to a presentation of some classical CART, Bagging and Boosting procedures for homoscedastic models. We also propose simple variations that allow to cope with heteroscedastic models. The guideline is their forthcoming application in Chapter 4.

*I would like to thank Servane Gey and Gilles Blanchard for many stimulating, helpful discussions about CART, Bagging and Boosting.

Au menu

3.1. Introduction **103**
3.2. Prerequisites **104**
3.2.1. The regression model 104
3.2.2. General principles for estimation methods 105
3.2.3. Tools 108
3.3. The CART regression algorithm **112**
3.3.1. Growing the maximal tree 112
3.3.2. Pruning the maximal tree 115
3.3.3. Selecting a good tree in the forest 118
3.3.4. Variables importance, stability 120
3.4. Bagging and Boosting procedures **121**
3.4.1. Black box modeling culture 121
3.4.2. Bagging 122
3.4.3. Boosting 124
3.4.4. Validation 128

3.1. Introduction

We present in this chapter the main statistical procedure we shall apply when tackling the original problem of refinement of the localization of the mobile telecommunication traffic in Paris, as we discussed in Prélude and Chapter 1. The application itself can be found in Chapter 4.

We argued in Chapter 1 why a simple regression model may provide interesting answers. Responses variables (*i.e.* various traffic data) and predictor variables (or covariables, *i.e.* socio-demographic and cultural data) have been carefully introduced in Chapter 1 and Interlude 2. Thus, the problem at stake is now to determine how to cope with the latter regression model with a view to applications. Various practical methods aim at estimating a regression function. We choose among them the CART regression trees procedure (see the unavoidable Breiman et al. 1984) as the elementary tool (the *weak learner*, according to the terminology introduced further in Section 3.4). We naturally do not pretend that this is the sole best choice. Other methods have proved, sometimes much earlier, their efficiency. Our reasons include the following:

- CART regression trees are flexible;
- CART regression trees are easily computed;
- CART regression trees are easily interpreted (though care is recommended) and particularly allow to calculate some variable importance, that informally quantifies how much a covariable affects the traffic;
- CART regression trees in committee may enhance the results of a single CART regression tree.

We have to emphasize that committees of CART regression trees are not interpretable anymore: they rather appear like *black boxes* (see Section 3.4 again).

This chapter is divided into three more sections. Section 3.2 is dedicated to general principles of estimation and some tools. We particularly introduce two crucial regression models. Indeed, regression methods usually cope with homoscedastic models (and so does the original CART procedure), *i.e.* models

$$Y = f^\star(X) + e.$$

Here, Y is the response, X the covariable, e a noise of conditional mean 0 and variance σ^2 given X and $f^\star$ is the regression function. Heteroscedastic models differ from the previous ones since no assumption is made on the variance of the noise, hence the more general modeling

$$Y = f^\star_{\mathrm{m}}(X) + f^\star_{\mathrm{se}}(X)e,$$

where e has conditional mean 0 and variance 1 given X. Such models fit better in our original problem than homoscedastic ones. Some other important notions as losses, empirical criteria, penalization are introduced and commented.

We propose in Section 3.3 an original, simple adaptation of the original CART procedure of Breiman et al. in order to allow application to heteroscedastic models. The study in Chapter 5 casts some light on our proposition in an idealistic framework where the regression functions are piecewise constant, but under mild assumptions that allow to deal with dependent observations.

We finally introduce in Section 3.4 the Bagging and Boosting iterative procedures on the basis of which committees of CART regression trees are built. Bagging is a perturb and combine procedure originated by Breiman (1996a) in a classification framework. Boosting is an adaptively resampling and combining procedure due to Freund and Schapire (1996) for classification and extended to regression by Drucker (1997). We also have to slightly transform them so that they can apply to heteroscedastic models and the associated modified CART regression trees.

All the algorithms are summarized in concise tables along this chapter.

In conclusion, the title of this chapter emphasizes that we shall not provide any simulation study for our modified procedures for heteroscedastic models. A systematic exploration of their properties is beyond its scope. We shall rather apply them to our original problem and draw attention on some of their features in Chapter 4.

3.2. Prerequisites

3.2.1. The regression model

Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space upon which all the random variables (rv) will be defined in the sequel. E denotes the expectation with respect to (wrt) P. We shall consider the following *regression model*

$$Y = f^\star_{\mathrm{m}}(X) + f^\star_{\mathrm{se}}(X)\, e, \tag{3.1}$$

where the rv $Z = (X, Y)$ takes its values in $\mathcal{X} \times \mathbb{R}$ and e is a noise with mean 0 and variance 1 conditionally given X . We denote $P^\star$ the joint distribution of $Z = (X, Y)$, P the marginal distribution of X and $f^\star = (f^\star_{\mathrm{m}}, f^\star_{\mathrm{se}})$ the regression function. We shall denote $\mathrm{E}_{P^\star}$ the expectation wrt the rv Z.

We aim at estimating the regression function $f^\star = (f^\star_{\mathrm{m}}, f^\star_{\mathrm{se}})$ by mean of a predictor $\widehat{f} = (\widehat{f}_{\mathrm{m}}, \widehat{f}_{\mathrm{se}})$ on the basis of some observations $Z_i = (X_i, Y_i)$ $(i = 1, \ldots, n)$ (*i.e.* by mean of a measurable function $\widehat{f}$ of the observations). We present in Section 3.3 the *practical method* of CART regression trees originated by Breiman et al. in the seventies. CART is based on binary tree-structured partitions and on a penalization device we shall discuss. Then, Section 3.4 is devoted to some recent, very popular, general procedures of enhancement called Bagging and Boosting. However, Section 3.2.2 is concerned with general principles of estimation and not particularly with CART, Bagging and Boosting.

Thus, let $\mathcal{F}$ be a large set of possible predictors that contains $f^\star$. Elements of $\mathcal{F}$ will be denoted f with the general decomposition $f = (f_{\mathrm{m}}, f_{\mathrm{se}})$ except when $f^\star_{\mathrm{se}} = \sigma$ is a constant: in that case, we shall denote $f = f_{\mathrm{m}}$ for sake of simplicity of the notations. In the same spirit, $f^\star$ and $\widehat{f}$ will equal $f^\star_{\mathrm{m}}$ and $\widehat{f}_{\mathrm{m}}$, respectively, when $f^\star_{\mathrm{se}} = \sigma$ is a constant function. In such a case, the model is

$$Y = f^\star(X) + e'$$

with e' noise of conditional mean 0 and variance σ^2 given X.

Remark 3.2.1 (homoscedasticity *vs* heteroscedasticity, to be continued).
The original CART, Bagging and Boosting methods in regression are concerned with *homoscedastic* models, *i.e.* models whose function $f^\star_{\mathrm{se}}$ is constant. Nevertheless, minor changes allow to cope with *heteroscedastic* models, *i.e.* models whose function $f^\star_{\mathrm{se}}$ is not constant. Besides, we are interested in such models for our issues, hence the present general modeling.

Breiman et al. comment the unfortunate effect on tree structures that stems from lack of homoscedasticity of the model in the original CART procedure. See Remark 3.3.2 below for more details.

3.2.2. General principles for estimation methods

Losses and empirical criteria

One usually wishes to make the *mean error* of the estimator of $f^\star$ applied to an unknown example as small as possible. Thus, one introduces the family of *loss functions* L which provide quantitative measurements of the performance of an estimator.

A loss function is usually a function L mapping $\mathcal{F}$ onto $\mathbb{R}$ that achieves its minimum in the sole $f^\star$. One is naturally tempted then to choose an estimator $\widehat{f}$ such that $L(\widehat{f})$ is (possibly almost) minimal. However, a loss function often depends on the unknown distribution $P^\star$, as for instance for the wide family of contrast loss functions of the type

$$L_\gamma(f) = \mathrm{E}_{P^\star}\gamma(f, Z)$$

where γ is a real valued function called *contrast*.

Given a contrast loss function, it is natural to introduce the *relative loss*, namely

$$RL_\gamma(f) = \mathrm{E}_{P^\star}\gamma(f, Z) - \mathrm{E}_{P^\star}\gamma(f^\star, Z).$$

Although the loss and the relative loss only differ by a constant (which is unknown, since it depends on true distribution $P^\star$), the two quantities enjoy distinct statistical properties that justify both their introduction.

In the sequel, we shall focus on two losses, respectively designed for homoscedastic and heteroscedastic models. We denote $z = (x, y)$ for any element of $\mathcal{X} \times \mathbb{R}$.

Homoscedastic model: assume that $f^\star = f^\star_{\mathrm{m}} \in \mathcal{F} = L^2(P)$. The quadratic contrast

$$\gamma_1(f, z) = \Big(y - f(x)\Big)^2$$

yields the classical *quadratic relative loss*

$$RL_{\gamma_1}(f) = RL_1(f) = P\,(f - f^\star)^2 \geq 0,$$

equality if and only if (iff) $f = f^\star$ P-almost everywhere (P-ae).

Heteroscedastic model: assume that $f^\star = (f^\star_{\mathrm{m}}, f^\star_{\mathrm{se}}) \in \mathcal{F} = \mathcal{F}_{\mathrm{m}} \times \mathcal{F}_{\mathrm{se}}$ such that $\mathcal{F}_{\mathrm{m}} \subset L^2(P)$ and elements of $\mathcal{F}_{\mathrm{se}}$ are uniformly bounded away from 0. The following contrast (whose definition stems from the log-likelihood of a Gaussian rv)

$$\gamma_2(f, z) = \frac{\Big(y - f_{\mathrm{m}}(x)\Big)^2}{{f_{\mathrm{se}}}^2(x)} + \log {f_{\mathrm{se}}}^2(x)$$

yields the less classical relative loss

$$RL_{\gamma_2}(f) = RL_2(f) = P\left(\frac{f_{\mathrm{m}} - f^\star_{\mathrm{m}}}{f_{\mathrm{se}}}\right)^2 + P\left(\frac{{f^\star_{\mathrm{se}}}^2}{{f_{\mathrm{se}}}^2} + \log\frac{{f_{\mathrm{se}}}^2}{{f^\star_{\mathrm{se}}}^2} - 1\right) \geq 0,$$

equality iff $f = f^\star$ P-ae, *i.e.* both $f_{\mathrm{m}} = f^\star_{\mathrm{m}}$ and $f_{\mathrm{se}} = f^\star_{\mathrm{se}}$ P-ae.

The heuristics of minimum contrast estimation offers a practical answer to the mentioned problem of dependency of the losses on the distribution $P^\star$. Informally speaking, there is some hope that minimization of the empirical criterion

$$f \mapsto n^{-1} \sum_{i=1}^{n} \gamma(f, Z_i)$$

rather than minimization of its expectation* provides a feasible method of estimation. Particularly, the latter empirical criterion coincides with the least squares when the contrast function is γ_1.

Model complexity and penalization device

The problem of concrete minimization over the whole set $\mathcal{F}$ nevertheless remains. A solution consists of minimizing on subsets $\mathcal{F}_K \subset \mathcal{F}_{K+1}$ ($K \geq 1$) of $\mathcal{F}$ *simpler* than the original set. Here, the index K is relevant of the complexity of $\mathcal{F}_K$.

Now, the more $\mathcal{F}_K$ is sophisticated, the better may be the minimization of the empirical criterion over $\mathcal{F}_K$ since one always has $\inf_{f \in \mathcal{F}_K} U_n(f)$ larger than $\inf_{f \in \mathcal{F}_{K+1}} U_n(f)$ for any empirical criterion U_n, whenever $\mathcal{F}_K$ is a subset of $\mathcal{F}_{K+1}$. Nonetheless, this increasing sophistication has a cost: one has indeed to warrant that overfitting is avoided. Equivalently, the estimator should

i.e. minimization of the original loss functions $f \mapsto L_\gamma(f)$ or equivalently $f \mapsto RL_\gamma(f)$.

not be too adapted to the observations, otherwise it would not apply well to new observations. In other words, if we agree that the empirical criterion is a *bias* term, then minimization of the sole empirical criterion corresponds to minimization of the sole bias, regardless of the subsequent increase of the *variance*. The method of penalization consists of adding a positive term $\mathrm{pen}(n, K)$ that increases with the sophistication K, in order to balance the latter phenomenon. Heuristically, $\mathrm{pen}(n, K)$ simulates a variance term, so that minimization of the penalized empirical criterion $U_n(f) + \mathrm{pen}(n, K)$ ($f \in \mathcal{F}_K$, $K \geq 1$) takes into account both the bias and the variance. Then, the theory ensures that, when correctly balanced, the penalization term yields good estimation.

This procedure has become very popular since it was introduced by Mallows (1973) and Akaike (1974). It has been widely explored and the literature about penalization is abundant. This work itself exploits the penalization device, see the current chapter as well as Chapters 5 and 7. We refer the reader to the comments therein for more references, examples of use and yielded results.

However, let us illustrate the latter remark in a peculiar simple framework that casts some light on the discussion above. Suppose that the class $\mathcal{F}_K$ is constituted of piecewise constant functions on some partition $\partial\mathcal{T}$ with K pieces, *i.e.* of functions of the form

$$f(x) = \sum_{t \in \partial\mathcal{T}} \alpha_t \mathbb{1}\{x \in t\}$$

for any $x \in \mathcal{X}$. With these notations, one has $P(\mathcal{X} \setminus \cup_{t \in \partial\mathcal{T}} t) = P(t \cap t') = 0$ for any pieces $t \neq t'$ of the partition $\mathcal{T}$ (the notations will be justified in Section 3.2.3).

Let us embed $\mathcal{F}$ in the Hilbert space $L^2(\mathbb{P}_n)$, where $\mathbb{P}_n$ is the empirical measure of $X_1, \ldots, X_n$, $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$. We denote $\|\cdot\|_2$ the $L^2(\mathbb{P}_n)$ norm. We finally assume that our framework is homoscedastic. Then, minimization of the sole empirical quadratic criterion

$$f \mapsto n^{-1} \sum_{i=1}^n \Big(Y_i - f(X_i)\Big)^2$$

wrt $f \in \mathcal{F}_K$ yields the estimator $\widehat{f}_{\mathcal{T}}$ given by

$$\widehat{f}_{\mathcal{T}}(x) = \sum_{t \in \partial\mathcal{T}} \widehat{v}(t) \mathbb{1}\{x \in t\} \quad (\text{any } x \in \mathcal{X})$$

where $\widehat{v}(t) = n_t^{-1} \sum_{i=1}^n Y_i \mathbb{1}\{X_i \in t\}$. Here, $n_t = \sum_{i=1}^n \mathbb{1}\{X_i \in t\}$ and $\widehat{v}(t) = 0$ by convention when $n_t = 0$. One may wonder how close $\widehat{f}_{\mathcal{T}}$ is to the regression function $f^\star$. A partial answer follows: first, a Pythagorean argument ensures that

$$\|f^\star - \widehat{f}_{\mathcal{T}}\|_2^2 = \inf\Big\{\|f^\star - f\|_2^2 : f \in \mathcal{F}_K\Big\} + \|\bar{f}_{\mathcal{T}} - \widehat{f}_{\mathcal{T}}\|_2^2,$$

where $\bar{f}_{\mathcal{T}}$ is the $L^2(\mathbb{P}_n)$ projection of $f^\star$ onto $\mathcal{F}_K$. Then

$$\begin{aligned}(\bar{f}_{\mathcal{T}} - \widehat{f}_{\mathcal{T}})^2 &= \sum_{t \in \partial\mathcal{T}} n_t^{-2} \left[\sum_{i=1}^n \Big(Y_i - f^\star(X_i)\Big)^2 \mathbb{1}\{X_i \in t\}\right] \mathbb{1}\{t\} \\ &= \sum_{t \in \partial\mathcal{T}} n_t^{-2} \left[\sum_{i=1}^n e_i^2 \mathbb{1}\{X_i \in t\}\right] \mathbb{1}\{t\},\end{aligned}$$

hence

$$\|\bar{f}_{\mathcal{T}} - \widehat{f}_{\mathcal{T}}\|_2^2 = n^{-1} \sum_{t \in \partial \mathcal{T}} \left[n_t^{-1} \sum_{i=1}^{n} e_i^2 \mathbb{1}\{X_i \in t\} \right].$$

Let us finally take the expectation of the expression above conditionally given $(X_1, \ldots, X_n)$ and then wrt the latter. We get

$$\mathrm{E}\|f^\star - \widehat{f}_{\mathcal{T}}\|_2^2 = \mathrm{E}\inf\left\{\|f^\star - f\|_2^2 : f \in \mathcal{F}_K\right\} + \sigma^2 \frac{K}{n}. \tag{3.2}$$

The left hand term of the equality provides a measurement of a mean distance between $f^\star$ and $\widehat{f}_{\mathcal{T}}$. It is the sum of two terms with opposite variations:

- a bias term $\mathrm{E}\inf\left\{\|f^\star - f\|_2^2 : f \in \mathcal{F}_K\right\}$ that decreases as K increases: it is the averaged best error that can be achieved on the class $\mathcal{F}_K$;
- a variance term $\sigma^2 K/n$ that increases as K increases: it is the difference between the typical error and the best error (the typical error is $\mathrm{E}\|f^\star - \widehat{f}_{\mathcal{T}}\|_2^2$ itself).

Thus, it seems reasonable to care about the variance while estimating, hence the penalization device.

Numerous methods cope with the estimation of $f^\star$ and subsequent difficulties. CART, with possible improvements yielded by Bagging or Boosting, provide some solutions with a view to applications. The next subsection is devoted to the presentation of useful tools we shall need in Sections 3.3 and 3.4, which are dedicated to the CART algorithm itself and the Bagging and Boosting procedures.

3.2.3. Tools

Learning set, test sample, errors and risks

Consider a *learning set* $\mathcal{L}_n$ composed of n observed rv drawn from the same regression model (3.1)

$$\mathcal{L}_n = \left\{Z_i = (X_i, Y_i) : i = 1, \ldots, n\right\} \subset \mathcal{X} \times \mathbb{R}$$

and denote $\mathrm{E}_{P^\star}^n$ the expectation wrt $\mathcal{L}_n$. We wish to estimate the regression function $f^\star$ by mean of a predictor $\widehat{f}$ on the basis of this learning set.

Thus, as suggested in Section 3.2.2, we can replace the contrast loss function we would have liked to minimize directly

$$L_\gamma : f \mapsto L_\gamma(f) = \mathrm{E}_{P^\star}\gamma(f, Z)$$

by the empirical criterion

$$\mathcal{E}_\gamma(\cdot, \mathcal{L}_n) : f \mapsto \mathcal{E}_\gamma(f, \mathcal{L}_n) = n^{-1} \sum_{i=1}^{n} \gamma(f, Z_i),$$

whose expectation under $\mathrm{E}_{P^\star}^n$ equals $L_\gamma(f)$. The right hand quantity above is called *modeling error* of $f \in \mathcal{F}$.

In such a case, a convenient measurement of the quality of the estimation of $f^\star$ by any candidate $\widehat{f}$ is provided by the *risk* of $\widehat{f}$ (or mean relative loss) whose definition is simply

$$\mathrm{R}_\gamma(\widehat{f}) = \mathrm{E}^n_{P\star} RL_\gamma(\widehat{f}).$$

In particular,

Homoscedastic model: the quadratic risk is given by

$$\mathrm{R}_{\gamma_1}(\widehat{f}) = \mathrm{R}_1(\widehat{f}) = \mathrm{E}^n_{P\star} P\,(\widehat{f} - f^\star)^2;$$

Heteroscedastic model: the associated risk is

$$\mathrm{R}_{\gamma_2}(\widehat{f}) = \mathrm{R}_2(\widehat{f}) = \mathrm{E}^n_{P\star}\left\{ P\left(\frac{\widehat{f}_{\mathrm{m}} - f^\star_{\mathrm{m}}}{\widehat{f}_{\mathrm{se}}}\right)^2 + P\left(\frac{f^{\star\,2}_{\mathrm{se}}}{\widehat{f}_{\mathrm{se}}^{\,2}} + \log\frac{\widehat{f}_{\mathrm{se}}^{\,2}}{f^{\star\,2}_{\mathrm{se}}} - 1\right)\right\}.$$

Once again, the definition of the risk involves $P^\star$. Nonetheless, given a *test sample* $\mathcal{C}_m$ independent of $\mathcal{L}_n$ (we write $\mathrm{E}^m_{P\star}$ the expectation wrt $\mathcal{C}_m$),

$$\mathcal{C}_m = \Big\{ Z_i = (X_i, Y_i) : i = n+1, \ldots, m+n \Big\}$$

we can compute for any $f \in \mathcal{F}$ its *test error*

$$\mathcal{E}_\gamma(f, \mathcal{C}_m) = m^{-1} \sum_{i=1}^{m} \gamma(f, Z_{n+i}).$$

This general expression takes the following particular forms in the cases we are interested in

Homoscedastic model: the test error can be written

$$\mathcal{E}_{\gamma_1}(f, \mathcal{C}_m) = m^{-1} \sum_{i=1}^{m} \Big(Y_{n+i} - f(X_{n+i})\Big)^2,$$

whose expectation is

$$\mathrm{E}^m_{P\star} \mathcal{E}_{\gamma_1}(f, \mathcal{C}_m) = L_{\gamma_1}(f);$$

Heteroscedatic model: the test error satisfies

$$\mathcal{E}_{\gamma_2}(f, \mathcal{C}_m) = m^{-1} \sum_{i=1}^{m} \frac{\Big(Y_{n+i} - f_{\mathrm{m}}(X_{n+i})\Big)^2}{f_{\mathrm{se}}{}^2(X_{n+i})} + \log f_{\mathrm{se}}{}^2(X_{n+i})$$

and its expectation is

$$\mathrm{E}^m_{P\star} \mathcal{E}_{\gamma_2}(f, \mathcal{C}_m) = L_{\gamma_2}(f).$$

Now, under appropriate assumptions, *e.g.* as soon as a *Law of large numbers* holds true, one has (the learning sample $\mathcal{L}_n$ is not involved),

$$\mathcal{E}_\gamma(f, \mathcal{C}_m) \underset{m\to\infty}{\longrightarrow} L_\gamma(f) \quad \text{in } \mathrm{P}_{P^\star}\text{–probability.}$$

Besides, if a predictor $\widehat{f}$ has been constructed on the basis of the learning sample $\mathcal{L}_n$, then independence of $\mathcal{L}_n$ wrt the check sample $\mathcal{C}_m$ yields in turn that,

$$\mathcal{E}_\gamma(\widehat{f}, \mathcal{C}_m) \underset{m\to\infty}{\longrightarrow} L_\gamma(\widehat{f}) \quad \text{in } \mathrm{P}_{P^\star}\text{–probability.} \tag{3.3}$$

One can also derive more subtle results that connect the loss $L_\gamma(\widehat{f})$ with $\mathcal{E}_\gamma(\widehat{f}, \mathcal{L}_n)$ and finally allows to control the relative loss $RL_\gamma(\widehat{f})$ in the spirit of (Vapnik 1998). Let us assume that the class of functions $\{\gamma(f, \cdot) : f \in \mathcal{F}_K\}$ is uniformly bounded above and below. Then, under appropriate assumptions that concern the complexity of the class $\mathcal{F}$ and for independent observations $Z_1, \ldots, Z_n$, one can show that there exist some constants A, B depending on $\mathcal{F}_K$ such that

$$\begin{aligned} L_\gamma(\widehat{f}) &\leq \mathcal{E}_\gamma(\widehat{f}, \mathcal{L}_n) + A\left[\frac{\log(2n/B)+1}{n/B} - \log(\eta/4)\right]^{1/2}, \\ L_\gamma(f^\star) &\geq \mathcal{E}_\gamma(f^\star, \mathcal{L}_n) - A\left[-\log(\eta)/2n\right]^{1/2}, \end{aligned} \tag{3.4}$$

where the latter inequalities hold true with probability $1 - 2\eta$ ($\eta < 1/2$). They finally yield

$$RL_\gamma(\widehat{f}) \leq A\left(\left[\frac{\log(2n/B)+1}{n/B} - \log(\eta/4)\right]^{1/2} + \left[-\frac{\log\eta}{2n}\right]^{1/2}\right) \tag{3.5}$$

which provides another simple justification of the minimization of the empirical loss procedure in this nice framework. Indeed, (3.5) particularly implies that the relative loss $RL_\gamma(\widehat{f})$ and the risk $\mathrm{R}_\gamma(\widehat{f})$ of $\widehat{f}$ go to 0 as n tends to infinity.

The omitted proof relies on simple concentration inequalities.* These powerful tools allow to cope with the case of a penalized empirical criterion minimized on nested models $\mathcal{F}_K$. They provide some rules of calibration of the penalization term and subsequent upper bounds for the risk that involve the minimum of $RL_\gamma(f)$ upon each model $\mathcal{F}_K$. This is far beyond our present scope, and we refer to (Massart 2000) for instance in a general context of model selection.

Binary regression trees

From now on, a *node* t will refer to a subset of $\mathcal{X}$, and the special node equal to the whole set $\mathcal{X}$ will be called *root* and denoted t_1. A (binary) tree $\mathcal{T}$ is obtained by repeated splits of nodes of $\mathcal{X}$ into (exactly two, or none: binary trees are *complete*) descendant nodes, beginning with the root t_1. Non-splitted nodes, or terminal nodes (wrt the hierarchical structure of the tree), will be called *leaves*. The set $\partial\mathcal{T}$ of the leaves of $\mathcal{T}$ forms a partition of $\mathcal{X}$.

*Concentrations inequalities are the mathematical devices which make rigorous the following statement of Talagrand (1996b): "A random variable that depends (in a "smooth way") on the influence of many independent variables (but not too much on any of them) is essentially constant." The latter "essential constantness" means that the rv are close to their expected value with high probability.

Denote now $\text{int}(\mathcal{T})$ the set of all the non-terminal nodes.The split of node $t \in \text{int}(\mathcal{T})$ is a *question*

$$q(t,\cdot) \ : \mathcal{X} \to \{0,1\}$$

such that, for any $X \in t$, X goes into the left hand *descendant* node t_L iff $q(t,X) = 0$ and into the right hand one t_R otherwise. Note that t' is a *descendant* of t (or equivalently, t is an *ancestor* of t') iff there is a connected path down the tree leading from t to t'. We shall denote $\mathcal{Q}$ the set of all possible questions. An important example for $\mathcal{Q}$ is

$$\mathcal{Q} = \Big\{ q : \mathcal{X} \to \{0,1\} : \exists C \in \mathbb{R},\ q(x) = \mathbb{1}\{x^{(k)} > C\} \Big\}, \tag{3.6}$$

where $x^{(k)}$ is the k-th coordinate of x (when $\mathcal{X}$ is a product space $\mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(d)}$, where $\mathcal{X}^{(k)} \subset \mathbb{R}$) and C is a given threshold. Such splitting rules yield a partition $\partial\mathcal{T}$ whose subsets t (*i.e.* the leaves of the tree) have boundaries normal to an axis.

Finally, a function $v \ : \partial\mathcal{T} \to \mathbb{R}$ assigns to any leaf a value. Particularly,

Homoscedastic model: $f^\star = f^\star_{\text{m}}$ and $v = v_{\text{m}}$ is a real-valued function;

Heteroscedastic model: $f^\star = (f^\star_{\text{m}}, f^\star_{\text{se}})$ and $v = (v_{\text{m}}, v_{\text{se}})$ has two real-valued coordinates.

Thus, given a tree $\mathcal{T}$ (together with its leaves assigning values function v), one can define its associated regression function

$$f_{\mathcal{T}}(x) = \sum_{t \in \partial\mathcal{T}} v(t)\, \mathbb{1}\{x \in t\} \quad (\text{any } x \in \mathcal{X}) \tag{3.7}$$

which is an estimator by histogram. See Figure 3.1 for a toy example.

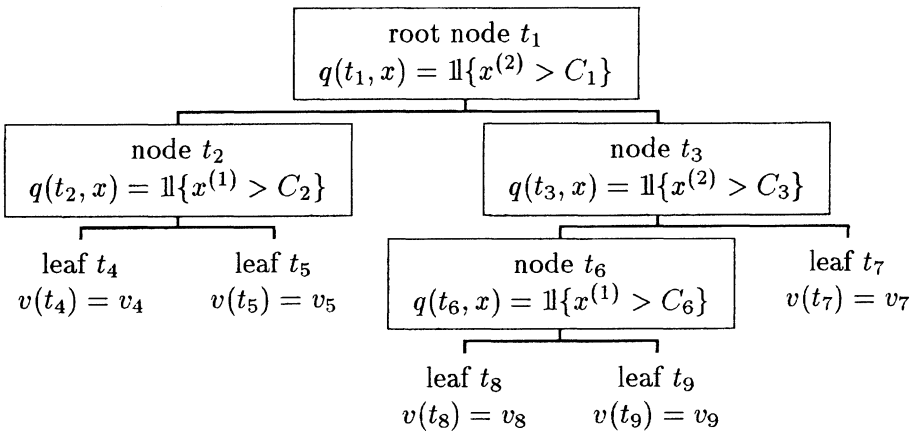


Figure 3.1 – A toy binary regression tree $\mathcal{T}$. Here, $\mathcal{X} = \mathbb{R}^3$, the set $\partial\mathcal{T}$ of the leaves of $\mathcal{T}$ is $\{t_4, t_5, t_7, t_8, t_9\}$ and the set $\text{int}(\mathcal{T})$ of non-terminal nodes is $\{t_1, t_2, t_3, t_6\}$.

Pruning of trees

One can define a *branch* $\mathcal{T}_t$ with root node t of some tree $\mathcal{T}$ as the *subtree* of $\mathcal{T}$ originating in t. Then, *pruning* a branch $\mathcal{T}_t$ of the tree $\mathcal{T}$ consists of deleting from $\mathcal{T}$ all descendants of t.

If $\mathcal{T}'$ is obtained *via* successive pruning procedures, then we say that $\mathcal{T}'$ is a *pruned subtree* of $\mathcal{T}$, which is denoted $\mathcal{T}' \preceq \mathcal{T}$. Such a definition for the relation $\preceq$ provides a non total order relation on the set of all pruned subtrees of $\mathcal{T}$.

Observe finally that the partition associated with the histogram $f_{\mathcal{T}'}$ of the pruned subtree $\mathcal{T}'$ of $\mathcal{T}$ is coarser than the one associated with $f_{\mathcal{T}}$.

3.3. The CART regression algorithm

The CART procedure consists of minimizing the empirical criterion on some subset of the set $\{f_{\mathcal{T}} : \mathcal{T}\}$ of binary trees histograms introduced in Section 3.2.3.

Remark 3.3.1. We emphasize here that the linearity of the latter criterion (wrt the observations Z_i) is absolutely crucial in the sequel. Much of the following does not hold anymore when a non linear criterion is chosen.

This simple algorithm obeys a three steps scheme: the first one consists of *growing a maximal tree*; the second of *pruning* it, getting a *forest* of trees, this step involving a penalization procedure as introduced in Section 3.2.1; and the last one of *choosing* a possibly "best" tree among the forest. The current section is devoted to the description of the algorithm, with a subsection for each step.

3.3.1. Growing the maximal tree

According to Section 3.2.3, the growing procedures requires

- a splitting rule at every intermediate node,
- a stopping rule for deciding when a node is a leaf,
- an assigning rule to associate a value with any leaf.

On the assigning value rule

We first define

$$\mathcal{L}_n|t = \Big\{ Z_i = (X_i, Y_i) \in \mathcal{L}_n : X_i \in t \Big\},$$
$$n_t = \mathrm{card}(\mathcal{L}_n|t),$$
$$\mathcal{E}_\gamma(f, \mathcal{L}_n|t) = n^{-1} \sum_{Z_i \in \mathcal{L}_n|t} \gamma(f, Z_i)$$

and consider some regression tree $\mathcal{T}$ and its regression function $f_{\mathcal{T}}$ whose definition is given by (3.7). Then, the linearity of the empirical criterion ensures that

$$\mathcal{E}_\gamma(f_{\mathcal{T}}, \mathcal{L}_n) = \sum_{t \in \partial\mathcal{T}} \mathcal{E}_\gamma(v(t), \mathcal{L}_n|t).$$

Consequently, the best leaves assigning value function $\widehat{v}$ (in terms of minimization of $\mathcal{E}_\gamma(f_\mathcal{T}, \mathcal{L}_n)$) is such that, for any $t \in \partial\mathcal{T}$,

$$\widehat{v}(t) = \arg\min_u \mathcal{E}_\gamma(u, \mathcal{L}_n|t), \tag{3.8}$$

i.e. particularly:

Homoscedastic model:

$$\widehat{v}(t) = \widehat{v}_{\mathrm{m}}(t) = n_t^{-1} \sum_{Z_i \in \mathcal{L}_n|t} Y_i = n_t^{-1} \sum_{Z_i \in \mathcal{L}_n} Y_i \mathbb{1}\{X_i \in t\}.$$

In other words, $\widehat{v}(t)$ is the empirical mean of the Y_i's whose associated X_i's fall into t;

Heteroscedastic model: $\widehat{v}_{\mathrm{m}}(t)$ is as above and

$$\widehat{v}_{\mathrm{se}}(t)^2 = n_t^{-1} \sum_{Z_i \in \mathcal{L}_n|t} \Big(Y_i - \widehat{v}_{\mathrm{m}}(t)\Big)^2 = n_t^{-1} \sum_{Z_i \in \mathcal{L}_n} \Big(Y_i - \widehat{v}_{\mathrm{m}}(t)\Big)^2 \mathbb{1}\{X_i \in t\}.$$

In other words, $\widehat{v}_{\mathrm{m}}(t)$ and $\widehat{v}_{\mathrm{se}}(t)$ are respectively the empirical mean and the empirical standard error of the Y_i's whose associated X_i's fall into t.

Since for any $t \in \partial\mathcal{T}$, $\widehat{v}(t)$ is calculated on the basis of the sole observations Y_i's such that $X_i \in t$, one often says that the CART regression method is *local*.

We shall denote in the sequel $\widehat{f}_\mathcal{T}$ the regression function based on $\mathcal{T}$ such that $\widehat{f}_\mathcal{T}|t = \widehat{v}(t)$ for any $t \in \partial\mathcal{T}$.

On the splitting rule

Suppose we have grown a tree $\mathcal{T}$ (possibly $\mathcal{T} = t_1$ which is not hard work) and that we wish to split a terminal node t into two descendants t_L and t_R. If we had already done so, we would have $t = t_L \cup t_R$ and

$$\begin{aligned} \mathcal{E}_\gamma(\widehat{v}(t), \mathcal{L}_n|t) &= \mathcal{E}_\gamma(\widehat{v}(t), \mathcal{L}_n|t_L) + \mathcal{E}_\gamma(\widehat{v}(t), \mathcal{L}_n|t_R) \\ &\geq \mathcal{E}_\gamma(\widehat{v}(t_L), \mathcal{L}_n|t_L) + \mathcal{E}_\gamma(\widehat{v}(t_R), \mathcal{L}_n|t_R). \end{aligned}$$

The previous inequality (that relies again on the linearity of the empirical criterion) suggests that a best split makes the difference between the left hand and the right hand terms of the inequality (the *gain*) as large as possible. Thus, we shall denote $\widehat{q}(t, \cdot)$ a best splitting question of the node t, *i.e.*

$$\widehat{q}(t, \cdot) = \arg\max_{q \in \mathcal{Q}} \Big\{ \mathcal{E}_\gamma(\widehat{v}(t), \mathcal{L}_n|t) - \mathcal{E}_\gamma(\widehat{v}(t_L), \mathcal{L}_n|t_L) - \mathcal{E}_\gamma(\widehat{v}(t_R), \mathcal{L}_n|t_R) \Big\} \tag{3.9}$$

where the index q in the arg max above ranges over all possible questions at node t, and t_L and t_R are the subsequent descendants of t.

Remark 3.3.2 (homoscedasticity *vs* heteroscedasticity, continued). When the contrast function is γ_1, one has

$$\begin{aligned} \mathcal{E}_{\gamma_1}(\widehat{v}(t), \mathcal{L}_n|t) &= p(t)\, s^2(t), \quad \text{where} \\ p(t) &= n_t/n \\ \text{and} \quad s^2(t) &= n_t^{-1} \sum_{Z_i \in \mathcal{L}_n|t} \Big(Y_i - \widehat{v}(t)\Big)^2 \end{aligned}$$

is the total squared deviations of the Y_i's in t from their average, *i.e.* the within-node variance. Suppose now that the contrast is γ_1 though the model is possibly heteroscedastic. In such a case, $s^2(t)$ may be large even though $\widehat{v}(t)$ approximates well the regression function on t. Nevertheless, the algorithm will try to split the node t into t_L and t_R, yielding a quantity (the notations naturally extend those above)

$$\mathcal{E}_\gamma(\widehat{v}(t_L), \mathcal{L}_n|t_L) + \mathcal{E}_\gamma(\widehat{v}(t_R), \mathcal{L}_n|t_R) = p(t_L)\, s^2(t_L) + p(t_R)\, s^2(t_R)$$

possibly much lower than the original one, although the associated estimators $\widehat{v}(t_L)$ and $\widehat{v}(t_R)$ of the regression functions on t_L and t_R may be poorer than the original $\widehat{v}(t)$.

Avoidance of this undesirable feature was one of our motivations when introducing the contrast γ_2. Indeed, the γ_2–procedure relies on the simultaneous estimation of the mean and the variance, which would discourage a split of the node t in the previous example.

Remark 3.3.3 (on the set of questions). We have already mentioned that an important set $\mathcal{Q}$ of questions is given by (3.6). It is indeed a very tractable set, equally in computational and interpretation terms. However, it is likely that it may make more sense to split on combinations of variables rather than only on the individual original ones. In other words, it likely happens that in some dataset, the classes are separated by hyperplans that are not necessarily normal to an axis. Breiman et al. suggest to consider best splits over linear combinations of the variables, or to introduce *ad hoc* combinations of variables yielded by examination of the data. We emphasize that the first suggestion is computationally greedy. Besides, we shall see in Section 3.4 that the CART procedure will be used as a *weak learner* whose results are to be enhanced, so it is not worth splitting along combinations of variables.

On the stopping rule

Given the splitting rule we have just described, one could decide to make a node t a terminal node (*i.e.* a leaf) as soon as the best split yields a gain smaller than a given threshold. The main drawback of that solution stems from the necessary choice of such a threshold, which requires exploratory tuning.

Consequently, one usually prefers (and so do we) to decide to make a node t a terminal node (*i.e.* a leaf) as soon as n_t is smaller than a fixed quantity, say 5. Particularly, this rule does not require any preliminary study of the data.

Summary: growing the maximal tree

The scheme in Table 3.1 sums up the procedure of growth of the maximal tree denoted $\mathcal{T}_{\max}$. This completes the growing procedure description. Let us consider now the second step of the algorithm, namely the pruning step.

Input:	The learning set $\mathcal{L}_n$.
Initialization:	Let $\mathcal{N} = \{t_1\}$ and $\mathcal{L} = \emptyset$.
	Split t_1 according to $\widehat{q}(t_1, \cdot)$ into $t_L \cup t_R$.
	Update $\mathcal{N} = \mathcal{N} \setminus \{t_1\}$.
	If $n_{t_L} > 5$, update $\mathcal{N} = \mathcal{N} \cup \{t_L\}$, else update $\mathcal{L} = \mathcal{L} \cup \{t_L\}$.
	If $n_{t_R} > 5$, update $\mathcal{N} = \mathcal{N} \cup \{t_R\}$, else update $\mathcal{L} = \mathcal{L} \cup \{t_R\}$.
Loop:	If $\text{card}(\mathcal{N}) == 0$, exit loop.
	Choose $t \in \mathcal{N}$.
	Split t according to $\widehat{q}(t, \cdot)$ into $t_L \cup t_R$.
	Update $\mathcal{N} = \mathcal{N} \setminus \{t\}$.
	If $n_{t_L} > 5$, update $\mathcal{N} = \mathcal{N} \cup \{t_L\}$, else update $\mathcal{L} = \mathcal{L} \cup \{t_L\}$.
	If $n_{t_R} > 5$, update $\mathcal{N} = \mathcal{N} \cup \{t_R\}$, else update $\mathcal{L} = \mathcal{L} \cup \{t_R\}$.
Termination:	For any $t \in \mathcal{L}$, assign the value $\widehat{v}(t)$.
Output:	The maximal tree $\mathcal{T}_{\max}$.

Table 3.1 – CART: growing the maximal tree.

3.3.2. Pruning the maximal tree

The maximal tree $\mathcal{T}_{\max}$ yields an estimator $\widehat{f}_{\mathcal{T}_{\max}}$ of the regression function. However, we do not *a priori* choose it as our final estimator. Indeed, $\widehat{f}_{\mathcal{T}_{\max}}$ is generally characterized by good fit on the learning set $\mathcal{L}_n$ combined with poor generalization property, or from another point of view, with low bias but high variance.

Thus, $\mathcal{T}_{\max}$ is rather considered as a base for the elaboration of better tree structured, piecewise constant regression functions. The procedure of extraction of better regression functions from $\mathcal{T}_{\max}$ is called *pruning procedure*. It involves penalization, as presented in Section 3.2.2. The next five subsections are devoted to the description of this fast and efficient algorithm.

Penalized empirical criteria

We briefly explained in Section 3.2.2 why it may be interesting to penalize the empirical criterion $\mathcal{E}_\gamma(\cdot, \mathcal{L}_n)$ in order to proceed simultaneously to the reduction of the bias and the variance. We present hereafter the penalization device proposed by Breiman et al. and the corresponding results. A short discussion on this particular choice is postponed in Remark 3.3.5.

Thus, let us introduce the following family $\{\mathcal{PE}_{\gamma,\alpha}\}$ of penalized empirical criteria: for any tree $\mathcal{T}$, for any penalization parameter $\beta \in \mathbb{R}_+$,

$$\mathcal{PE}_{\gamma,\beta}(\mathcal{T}) = \mathcal{PE}_{\gamma,\beta}(\mathcal{T}, \mathcal{L}_n) = \mathcal{E}_\gamma(\widehat{f}_\mathcal{T}, \mathcal{L}_n) + \beta\,|\partial\mathcal{T}|,$$

where $|\partial\mathcal{T}|$ denotes the number of leaves of $\mathcal{T}$. The nonnegative parameter β corresponds to the inverse of a "temperature": heuristically, if one consider a tree $\mathcal{T}$, then the higher β^{-1} (the lower β), the larger are the maximal trees $\mathcal{T}'$ grown from $\mathcal{T}$ while preserving the upper bounding

$$\mathcal{PE}_{\gamma,\beta}(\mathcal{T}') \leq \mathcal{PE}_{\gamma,\beta}(\mathcal{T}).$$

Observe here that $\mathcal{E}_\gamma(\widehat{f}_{\mathcal{T}'}, \mathcal{L}_n)$ and $\beta|\partial\mathcal{T}'|$ respectively decreases and increases when $\mathcal{T}'$ grows from $\mathcal{T}$.

The latter is made precise in the following proposition (for a proof, refer to Breiman et al. 1984):

Proposition 3.3.4. *Set a penalization parameter $\beta \in \mathbb{R}_+$. There exists a unique pruned subtree $\mathcal{T}_\beta$ of $\mathcal{T}_{\max}$ such that*

$$\mathcal{T}_\beta = \arg\min \left\{ \mathcal{PE}_{\gamma,\beta}(\mathcal{T}) : \mathcal{T} \preceq \mathcal{T}_{\max} \right\} \tag{3.10}$$

with the extra constraint

$$\mathcal{PE}_{\gamma,\beta}(\mathcal{T}) = \mathcal{PE}_{\gamma,\beta}(\mathcal{T}_\beta) \Longrightarrow \mathcal{T}_\beta \preceq \mathcal{T}. \tag{3.11}$$

Furthermore, whenever $\beta_1 \leq \beta_2$ are two penalization parameters, one has

$$\mathcal{T}_{\beta_2} \preceq \mathcal{T}_{\beta_1}. \tag{3.12}$$

Remark 3.3.5.

• Once again, the linearity of $\mathcal{E}_\gamma(\widehat{f}_\mathcal{T}, \mathcal{L}_n)$ (wrt the observations Z_i) and subsequent properties are crucial in the proof.

• Breiman et al. have chosen a penalty term which is proportional to the number of leaves. It is certainly a convenient choice regarding the technical facilities it yields. Furthermore, such a penalty term recalls the variance term we exhibited in (3.2). However, it is not clear whether this choice yields as good statistical properties as it might.

Let us cite some recent results in a framework of independent observations and quadratic contrast γ_1 which allow us to think that such a penalty term is appropriate. One can find in (Donoho 1997) some *oracle inequalities* for CART in a framework of fixed design. Heuristically, an oracle inequality ensures that an estimator which is obtained by minimization of an empirical criterion is not only optimal regarding the latter minimization, but also regarding the control of its risk by optimal relative losses on each model. More recently, Gey and Nedelec (2001) have analyzed the pruning procedure in bounded random design and Gaussian fixed design regression models. Their results also yield that a penalization proportional to the number of leaves is appropriate.

We have also tried to cast some light on that choice of penalty in a general framework where the observations are possibly dependent and the contrast is not necessarily the quadratic one, under a certain idealization. The model is roughly the following: there exists a partition $\tau^\star$ of $\mathcal{X}$ such that $f^\star$ is piecewise constant on $\tau^\star = (\tau_j^\star)_{1 \leq j \leq K^\star}$, *i.e.* that there exist some parameters $\theta^\star = (\theta_j^\star)_{1 \leq j \leq K^\star}$ such that

$$f^\star(x) = \sum_{j=1}^{K^\star} \theta_j^\star \mathbb{1}\{x \in \tau_j^\star\} \quad (\text{any } x \in \mathcal{X}).$$

The model $\mathcal{F}_K$ of the nested collection $\{\mathcal{F}_K\}_{1 \leq K \leq K_{\max}}$ is constituted of the piecewise constant functions f_τ on any partition τ of $\mathcal{X}$ with K pieces (it is assumed that $f^\star$ belongs to the collection). The estimator $\widehat{f}$ of $f^\star$ is obtained by minimization of

$$f_\tau \mapsto \mathcal{E}_\gamma(f_\tau, \mathcal{L}_n) + \beta_n |f_\tau|$$

over the whole collection $\cup_{K=1}^{K_{\max}} \mathcal{F}_K$ (here, $|f_\tau|$ denotes the smallest K such that $f_\tau \in \mathcal{F}_K$). We proved under mild assumptions, including the calibration of the sequence $\{\beta_n\}$, that $\widehat{f}$ is weakly consistent, *i.e.* that the estimators of $\tau^\star$ and $\theta^\star$ both converge in probability to the true values (convenient notions of convergence are defined). The reader is referred to Chapter 5 for the whole details of this work.

The pruning procedure initialization

First, one extracts from $\mathcal{T}_{\max}$ the subtree $\mathcal{T}_0$ introduced in Proposition 3.3.4. Indeed, $\mathcal{T}_{\max}$ obviously satisfies (3.10). Besides, for any t node of $\mathcal{T}_{\max}$, denoting $\mathcal{T}_{\max,t}$ the branch of $\mathcal{T}_{\max}$ with root node t, we have

$$\mathcal{E}_\gamma(\widehat{v}(t), \mathcal{L}_n|t) \geq \mathcal{E}_\gamma(\widehat{f}_{\mathcal{T}_{\max,t}}, \mathcal{L}_n|t). \tag{3.13}$$

Now, if we prune each branch $\mathcal{T}_{\max,t}$ such that the above inequality is an equality, we get a subtree of $\mathcal{T}_{\max}$ that satisfies both (3.10) and (3.11), *i.e.* the subtree $\mathcal{T}_0$ itself, according to the terminology of Proposition 3.3.4. We set $\beta_1 = 0$.

This completes the first step of removal of the superfluous branches in terms of gain.

The pruning procedure iterative scheme

Let us assume that we have already constructed two nondecreasing and nonincreasing finite sequences $\{\beta_k\}_{1\leq k\leq K}$ and $\{\mathcal{T}_{\beta_k}\}_{1\leq k\leq K}$ such that, for any $1 \leq k \leq K$,

$$\mathcal{T}_{\beta_k} = \arg\min\left\{\mathcal{PE}_{\gamma,\beta_k}(\mathcal{T}) : \mathcal{T} \preceq \mathcal{T}_{\beta_{k-1}}\right\}, \tag{3.14}$$

$$\mathcal{PE}_{\gamma,\beta_k}(\mathcal{T}) = \mathcal{PE}_{\gamma,\beta_k}(\mathcal{T}_{\beta_k}) \Longrightarrow \mathcal{T}_{\beta_k} \preceq \mathcal{T}. \tag{3.15}$$

(with the convention $\mathcal{T}_{\beta_{-1}} = \mathcal{T}_{\max}$).

Let us denote $\mathcal{T}_{\beta_K,t}$ the branch of $\mathcal{T}_{\beta_K}$ with root the node $t \in \mathrm{int}(\mathcal{T}_{\beta_\|})$. Then, for any node t of $\mathcal{T}_{\beta_K}$, the inequality (3.13) is still satisfied when substituting the branch $\mathcal{T}_{\beta_K,t}$ to $\mathcal{T}_{\max,t}$. Moreover, for a penalization parameter β small enough, one has indeed the stronger inequality

$$\mathcal{PE}_{\gamma,\beta}(t) = \mathcal{E}_\gamma(\widehat{v}(t), \mathcal{L}_n|t) + \beta \geq \mathcal{E}_\gamma(\widehat{f}_{\mathcal{T}_{\beta_K,t}}, \mathcal{L}_n|t) + \beta|\partial\mathcal{T}_{\beta_K,t}| = \mathcal{PE}_{\gamma,\beta}(\mathcal{T}_{\beta_K,t}).$$

Let us choose the smallest β so that the above inequality becomes an equality for at least one node t, *i.e.*

$$\beta_{K+1} = \min\left\{\frac{\mathcal{E}_\gamma(\widehat{v}(t), \mathcal{L}_n|t) - \mathcal{E}_\gamma(\widehat{f}_{\mathcal{T}_{\beta_K,t}}, \mathcal{L}_n|t)}{|\partial\mathcal{T}_{\beta_K,t}| - 1} : t \in \mathrm{int}(\mathcal{T}_{\beta_K})\right\}, \tag{3.16}$$

where the latter ratios are infinite as soon as t is a leaf of $\mathcal{T}_{\beta_K}$. The corresponding tree $\mathcal{T}_{\beta_{K+1}}$ is derived from $\mathcal{T}_{\beta_K}$ by pruning all its branches whose root t achieve the maximum in the previous definition. Moreover, β_{K+1} and $\mathcal{T}_{\beta_{K+1}}$ satisfy both (3.14) and (3.15).

The loop ends when $\mathcal{T}_{\beta_{K+1}}$ coincides with the root t_1 of $\mathcal{T}_{\max}$.

The theorem below summarizes the properties of the sequences we have just constructed (for a proof, refer again to Breiman et al. 1984).

Theorem 3.3.6 (Breiman et al.).
The sequence $\{\beta_k\}$ of penalization parameters increases while the associated sequence $\{\mathcal{T}_{\beta_k}\}$ decreases. Besides, for any penalization parameter β,

$$\beta_k \leq \beta < \beta_{k+1} \Longrightarrow \mathcal{T}_\beta = \mathcal{T}_{\beta_k},$$

and $\max_k \beta_k \leq \beta$ yields $\mathcal{T}_\beta = t_1$.

More comments

We first would like to emphasize two remarkable features of the CART algorithm readily derived from Theorem 3.3.6:

- We do obtain all the pruned subtrees $\mathcal{T}_\beta$ of $\mathcal{T}_{\max}$ defined by (3.10) and (3.11) for *any* penalization parameter β with a *single* course of the whole tree $\mathcal{T}_{\max}$. Besides, the elementary operations are readily computed. Thus, the pruning algorithm is fast.

- The regression function $\widehat{f}_{\mathcal{T}_{\beta_k}}$ based on $\mathcal{T}_{\beta_k}$ achieves the minimum of

$$\widehat{f}_{\mathcal{T}} \mapsto \mathcal{E}_\gamma(\widehat{f}_{\mathcal{T}}, \mathcal{L}_n)$$

  where $\mathcal{T}$ ranges over the set of *all* the pruned subtrees of $\mathcal{T}_{\max}$ *whose number of leaves equal* $|\partial\mathcal{T}_{\beta_k}|$.

  In other words, $\widehat{f}_{\mathcal{T}_{\beta_k}}$ is the sole best (in terms of minimization of $\mathcal{E}_\gamma(f_{\mathcal{T}}, \mathcal{L}_n)$) regression function based on a pruned subtree of $\mathcal{T}_{\max}$ with $|\partial\mathcal{T}_{\beta_k}|$ leaves.

Furthermore, a nice graphical interpretation characterizes the number of leaves that appear in the finite sequence $\{|\partial\mathcal{T}_{\beta_k}| : k\}$. Indeed, if one draws the points

$$\Big(\ell, \min\Big\{\mathcal{E}_\gamma(\widehat{f}_{\mathcal{T}}, \mathcal{L}_n) : \mathcal{T} \preceq \mathcal{T}_{\max} \text{ and } |\partial\mathcal{T}| = \ell\Big\}\Big)$$

for every ℓ between 1 and $|\partial\mathcal{T}_{\max}|$, then the numbers of leaves of the sequence $\{|\partial\mathcal{T}_{\beta_k}| : k\}$ correspond to the above point that belongs to the lower convex hull of the cluster.

Summary: pruning the maximal tree

The scheme in Table 3.2 sums up the procedure of pruning of the maximal tree. We denote therein, for any tree $\mathcal{T}$ and node t,

- $\partial\text{int}(\mathcal{T})$ the set of all nodes of the tree $\mathcal{T}$ whose descendants are leaves;

- $'t$ the ancestor of t, with convention $'t_1 = t_1$.

This completes the pruning procedure description. Let us consider now the last step of the algorithm, namely the final selection step.

3.3.3. Selecting a good tree in the forest

The pruning procedure yields a *forest of trees* $\{\mathcal{T}_{\beta_k}\}$ among which we want to select a "best" tree and its associated regression function. There exist several methods. We shall omit the classical *cross-validation* for it is computationally too expensive (see *e.g.* Breiman et al. 1984 for a presentation of this procedure).

Input:	The learning set $\mathcal{L}_n$ and the maximal tree $\mathcal{T}_{\max}$.
Initialization:	Let $\mathcal{N} = \partial\text{int}(\mathcal{T}_{\max})$, $\mathcal{L} = \emptyset$ and $\mathcal{T} = \mathcal{T}_{\max}$.
Loop:	Choose $t \in \mathcal{N}$. If $\mathcal{E}_\gamma(\widehat{v}(t), \mathcal{L}_n \vert t) == \mathcal{E}_\gamma(\widehat{f}_{\mathcal{T}_t}, \mathcal{L}_n \vert t)$, update $\mathcal{T} = \mathcal{T} \setminus \mathcal{T}_t$, $\mathcal{L} = \mathcal{L} \cup \{'t\}$. Update $\mathcal{N} = \mathcal{N} \setminus \{t\}$. If $\text{card}(\mathcal{N}) == 0$ and $\text{card}(\mathcal{L}) == 0$, exit loop. If $\text{card}(\mathcal{N}) == 0$ and $\text{card}(\mathcal{L}) > 0$, update $\mathcal{N} = \mathcal{L}$.
End of initial.:	Set $\beta_1 = 0$, $\mathcal{T}_0 = \mathcal{T}$ and $K = 1$.
Outer loop (1):	Let $\mathcal{N} = \partial\text{int}(\mathcal{T}_{\beta_K})$, $\mathcal{L} = \emptyset$ and $\mathcal{T} = \mathcal{T}_{\beta_K}$. If $\text{card}(\mathcal{N}) == 0$, exit outer loop. Let β_{K+1} be given by (3.16).
Inner loop:	Choose $t \in \mathcal{N}$. If $\mathcal{PE}_{\gamma,\beta_{K+1}}(t) == \mathcal{PE}_{\gamma,\beta_{K+1}}(\mathcal{T}_{\beta_{K+1},t})$, update $\mathcal{T} = \mathcal{T} \setminus \mathcal{T}_t$, $\mathcal{L} = \mathcal{L} \cup \{'t\}$. Update $\mathcal{N} = \mathcal{N} \setminus \{t\}$. If $\text{card}(\mathcal{N}) == 0$ and $\text{card}(\mathcal{L}) == 0$, exit inner loop. If $\text{card}(\mathcal{N}) == 0$ and $\text{card}(\mathcal{L}) > 0$, update $\mathcal{N} = \mathcal{L}$.
Outer loop (2):	Set $\mathcal{T}_{\beta_{K+1}} = \mathcal{T}$ and $K = K + 1$.
Termination:	Print "et le tremblement éperdu des peupliers blancs d'Asie".
Output:	The sequences $\{\beta_k\}$ and $\{\mathcal{T}_{\beta_k}\}$.

Table 3.2 – CART: pruning the maximal tree.

Choice on the basis of a test sample

When a test sample $\mathcal{C}_m$ is provided, *i.e.* when one has another set of observations which is *independent* of the learning set $\mathcal{L}_n$, the selection can be made on the basis of the test errors $\mathcal{E}_\gamma(\widehat{f}_{\mathcal{T}_{\beta_k}}, \mathcal{C}_m)$. A "best" tree is then a tree whose regression function makes the test error as small as possible, *i.e.* formally

$$\widehat{f}_{\text{best}} = \arg\min \left\{ \mathcal{E}_\gamma(f, \mathcal{C}_m) : f \in \{\widehat{f}_{\mathcal{T}_{\beta_k}} : k \geq 0\} \right\}.$$

Such a choice is justified for instance by (3.3). Indeed, the latter convergence in $\text{P}_{P^\star}$-probability of $\mathcal{E}_\gamma(\widehat{f}, \mathcal{C}_m)$ to $L_\gamma(\widehat{f})$ (as m tends to infinity) allows to think that $\widehat{f}_{\text{best}}$ is a reasonable candidate for the minimization of $L_\gamma(f)$, at least over the family $\{\widehat{f}_{\mathcal{T}_{\beta_k}}\}$.

A robust choice

However, our favourite choice is yet another one, which aims at ensuring *robustness*. Its definition stems from the characteristic *versatility* of the CART procedure: in words, we mean that pruning occurs more frequently when tuning the penalization parameter $\beta \in \mathbb{R}_+$ over a range of small values than over a range of large values. Equivalently, the positive differences $(\beta_{k+1} - \beta_k)$ tend to be smaller for low values of k and larger for high values of k. We emphasize that $(\beta_{k+1} - \beta_k)$ is the length of the interval of penalization parameters β upon which the optimal tree $\mathcal{T}_\beta$ coincides with $\mathcal{T}_{\beta_k}$, *i.e.*

$$\beta_k \leq \beta < \beta_{k+1} \Longleftrightarrow \mathcal{T}_\beta = \mathcal{T}_{\beta_k}.$$

Heuristically, more robustness is ensured when picking up a tree $\mathcal{T}_{\beta_k}$ in the forest with large difference $(\beta_{k+1} - \beta_k)$. Since the notions of large or small values for the latter differences can not be set in generality, one uses renormalized differences, according to the definition below.

Let us set a threshold $\lambda > 0$, define $D_0 = \beta_1$ and for any $k \geq 1$, $D_k = \beta_{k+1} - \beta_k$. We then choose the smallest $k \geq 2$ such that

$$k \frac{D_k}{\sum_{l=0}^{k-1} D_l} \geq \lambda.$$

Remark 3.3.7. This choice has particularly interested us because our implementation for Matlab of the CART procedure is slow for large learning sets. Furthermore, the forthcoming Bagging and Boosting procedures, whose aim is to enhance the results of CART, rely on iterative construction of such CART regression trees. Now, the robust choice presented above performs very quickly, which is of great practical interest. On the contrary, it requires tuning for the parameter λ, which is an undeniable drawback.

3.3.4. Variables importance, stability

Variables importance

One can easily define a simple rule that allows ranking of the covariables according to their *importance*. Informally, the importance quantifies how much a covariable can discriminate two responses. In our framework, one can replace the latter informal definition by the concrete one: how much splitting gain does each covariable yields when constructing our maximal tree ?

Indeed, let us write $\mathcal{X} = \mathcal{X}^{(1)} \times \ldots \times \mathcal{X}^{(d)}$, so that $x^{(k)}$ is the k-th covariable of $x \in \mathcal{X}$. Then a satisfactory definition of the importance $\Delta_\delta(k)$ of the k-th covariable as a sum of best gain at any node of the tree $\mathcal{T}$ is given by

$$\Delta_\delta(k) = \sum_{t \in \mathrm{int}(\mathcal{T})} \delta^{\mathrm{depth}(t)} \max_q \Big\{ \mathcal{E}_\gamma(\widehat{v}(t), \mathcal{L}_n | t) - \mathcal{E}_\gamma(\widehat{v}(t_L), \mathcal{L}_n | t_L) - \mathcal{E}_\gamma(\widehat{v}(t_R), \mathcal{L}_n | t_R) \Big\},$$

where depth(t) denotes the depth of the node t wrt the hierarchical structure of $\mathcal{T}$; the questions in the maximum are of the form $q(x) = \mathbb{1}\{x^{(k)} > C\}$ and yield the two descendants t_L and t_R; $0 < \delta \leq 1$ is an additional parameter.

The parameter δ aims at favouring earlier gains wrt the hierarchical structure of the tree. Indeed, the comparison of $\Delta_\delta(k)$ for various δ gives some interesting details on the importance of the k-th covariable. It may for instance provide some information on the depth where it yields larger gains: low depth corresponds to global importance, while large depth corresponds to local importance. Typical values for δ are 0.8, 0.9 and 1.

Furthermore, only the relative magnitudes of the $\Delta_\delta(k)$ do matter. Thus, one will usually prefer the renormalized vector $100 \times \Delta_\delta / \max_k \Delta_\delta(k)$. However, we do not introduce this renormalization in the definition above, because Δ_δ will be involved in other procedures (see Sections 3.4.2 and 3.4.3) under this form.

This is obviously a solution among many others that enjoys nice properties. It particularly does not require any additional calculus in the CART procedure. Nevertheless, care is needed when interpreting these variables. We shall use the notion of variables importance in Chapter 4.

Stability

We defined in Section 3.2.3 the risk of an estimator of $f^\star$, hence particularly the risk of the estimator provided by the CART algorithm, which is a measurement of the performance of the

algorithm. We emphasized that the risk is not computable, since it depends on the unknown true distribution of the observations. An approach in the spirit of (Vapnik 1998) allows to bound above the risk with large probability by the sum of an expression of the sample size n which goes to zero as n grows to infinity and the empirical modeling error (see 3.4). In this case, the proof relies on some concentration inequalities that ensure some uniform convergence of empirical quantities to their mean. Another approach involves the notion of stability of an algorithm.

The *stability* of an algorithm quantifies how much variations of the input can influence the output. One can indeed identify two sources of randomness: *primo*, in the sampling mechanism; *secundo*, in the intrinsic noise of the observations. Bousquet and Elisseeff (2002) study this notion with a focus on the sampling randomness.* They define three notions of stability that involves the comparison between the estimator constructed on the basis of the whole learning set $\mathcal{L}_n$ and the estimator constructed on the basis of the learning set with a removed observation. They finally get some exponential upper bounds on the risk based on the notion of stability. Informally, the more stable is the algorithm, the closer the modeling error is to its expected loss (*i.e.* its mean loss, which coincides with its risk up to the constant $\mathrm{E}_{P^\star}\gamma(f^\star, Z)$). Thus, more stable estimation procedures perform better.

Does this mean that instable procedures have to be carefully avoided ? Fortunately not, because Breiman (1996b) illustrated the instability of the CART procedure.

Indeed, independently from this theoretical analysis, practical methods have been developed in order to tackle the practical problem of instability of learning algorithms, among which CART, in order to produce stable ones. Breiman (1996a) introduced the Bagging technique and Freund and Schapire (1996) the Boosting procedure which both originally aimed at improving learning algorithms. Those methods are roughly based on the combination of several simple learners in order to build a committee of learners which is typically much superior to any of the single ones.

The next section is devoted to a presentation of the latter procedures in our framework of regression.

3.4. Bagging and Boosting procedures

3.4.1. Black box modeling culture

The title of this section is a reference to (Breiman 2001). In the latter stimulating paper, Breiman confronts the two statistical cultures of data modeling *versus* algorithmic modeling.

Given an original regression problem and its associated data, one can think of the original process as a *black box* that outputs a response variable Y on the basis of an input variable X. The main statistical goal is twofold: *primo* inform, *i.e.* derive from the observations some elements of explanations of how the black box works; *secundo* predict, *i.e.* emulate the black box. According to the author, there are mainly two ways of tackling these problems:

- data modeling, where a probabilistic model for the back box is proposed, fitted on the basis of the observations and tested;
- algorithmic modeling, where an almost as black box is constructed on the basis of the observations and also tested.

*We refer the reader to the paper itself for more details.

In the first case above, interpretability is an important issue. It is not in the second one, where accurate information is mostly wanted.

The P&C procedures (for Perturb and Combine) have been developed with a view to the second approach. Given an automated learning algorithm $\mathcal{W}$ called *weak learner* and characterized by its instability (see Section 3.3.4), they consist of combining a committee of such weak learners constructed on perturbed learning sets, whose complex association yields a better estimator of the original black box. Such weak learners may be neural nets or CART regression trees, as introduced in the previous sections. Bagging (Bootstrap aggregating) weak learners (originated in Breiman 1996a for classification) is an example of P&C procedure. We refer the reader to the recent paper (Bühlmann and Yu 2002a) for a comprehensive bibliography and various theoretical results.

Sequential reweighting schemes, also known as Arcing procedures (for Adaptively resampling and combining, see Breiman 1998 in a framework of classification), have also been proposed by the machine learning community to cope with the emulation of the black box by combining some weak learners constructed by adaptive resampling. Boosting is an example due to Freund and Schapire (1996) in the classification framework (with a comparison to Bagging). Boosting regression trees can be based on a gradient descent approach (see Bühlmann and Yu 2002b and the references therein). It can also follow more closely the original procedure of Freund and Schapire, as proposed by Drucker (1997). The sequel of this section follows the lines of the latter paper when considering homoscedastic models. Let us finally mention the work of Gey and Poggi (2002)*, who address the stability issues of the Boosting procedure of CART regression trees, as well as global performance comparisons.

3.4.2. Bagging

Detailed description

Let us recall that $\mathcal{L}_n$ denotes the learning sample. The Bagging procedure obeys a P&C scheme:

- Perturbation step: the construction of weak learners according to $\mathcal{W}$ is done repeatedly on the basis of bootstrap samples of n observations (n-samples) drawn *uniformly* with replacement from the original set $\mathcal{L}_n$;

- Combination step: the final regressor is an average of the latter weak learners. Particularly, if one has trained K weak learners, whose estimated regression functions are denoted $\widehat{f}^1, \ldots, \widehat{f}^K$, then

  Homoscedastic model: the bagged estimator $\widehat{f}^{\mathrm{ba}}$ is defined by

$$\widehat{f}^{\mathrm{ba}} = K^{-1} \sum_{k=1}^{K} \widehat{f}^k.$$

  Now, when $\mathcal{W}$ is the CART procedure, each $\widehat{f}^k$ is associated with a tree $\mathcal{T}_k$ and a

*I would like to thank the authors and Gilles Blanchard again for the reference (Drucker 1997) and many interesting, helpful discussions about CART, Bagging and Boosting.

leaves assigning value function $\widehat{v}^k$, so that $\widehat{f}^{\mathrm{ba}}(x)$ is the mean

$$\widehat{f}^{\mathrm{ba}}(x) = K^{-1} \sum_{k=1}^{K} \widehat{v}^k(t_k) \tag{3.17}$$

whenever $x \in \bigcap_{k=1}^{K} t_k$ ($t_k \in \partial \mathcal{T}_k$). Furthermore, $\widehat{v}^k(t_k)$ is the empirical mean of all the Y_i's from the k-th bootstrap n-sample of $\mathcal{L}_n$ whose associated X_i's belong to t_k, so that $\widehat{f}^{\mathrm{ba}}(x)$ is the equally weighted mean (see Remark 3.4.2 below) of those empirical means.

Heteroscedastic model: the bagged estimator $\widehat{f}^{\mathrm{ba}} = (\widehat{f}^{\mathrm{ba}}_{\mathrm{m}}, \widehat{f}^{\mathrm{ba}}_{\mathrm{se}})$ is defined as above for $\widehat{f}^{\mathrm{ba}}_{\mathrm{m}}$ and

$$\left[\widehat{f}^{\mathrm{ba}}_{\mathrm{se}}(x)\right]^2 = \left[K^{-1} \sum_{k=1}^{K} \widehat{v}^k_{\mathrm{se}}(t_k)^2\right] + \left[K^{-1} \sum_{k=1}^{K} \widehat{v}^k_{\mathrm{m}}(t_k)^2 - \left(K^{-1} \sum_{k=1}^{K} \widehat{v}^k_{\mathrm{m}}(t_k)\right)^2\right] \tag{3.18}$$

whenever $x \in \bigcap_{k=1}^{K} t_k$ ($t_k \in \partial \mathcal{T}_k$). Furthermore, $\widehat{v}^k_{\mathrm{m}}(t_k)$ and $\widehat{v}^k_{\mathrm{se}}(t_k)^2$ are respectively the empirical mean and variance of all the Y_i's from the k-th bootstrap n-sample of $\mathcal{L}_n$ whose associated X_i's belong to t_k, so that the square of $\widehat{f}^{\mathrm{ba}}_{\mathrm{se}}(x)$ is the equally weighted mean (see Remark 3.4.2 below) of those empirical variances plus the variance of those empirical means.

Some comments

As far as we know, Bagging in heteroscedastic models is an original procedure, as CART regression trees in the latter framework. The definitions above stem from the following elementary lemma

Lemma 3.4.1. *For any $1 \leq k \leq K$, let $(U^k_1, \ldots, U^k_{n_k})$ be n_k rv, $\overline{U}^k$ be their empirical mean and S^k their empirical variance. Define $N = \sum_{k=1}^{K} n_k$ and $\pi_k = n_k/N$ $(1 \leq k \leq K)$. Then obviously,*

$$\overline{U} = \sum_{k=1}^{K} \pi_k \overline{U}^k$$

$$\text{and} \quad S = \left[\sum_{k=1}^{K} \pi_k S^k\right] + \left[\sum_{k=1}^{K} \pi_k \left(\overline{U}^k\right)^2 - \left(\sum_{k=1}^{K} \pi_k \overline{U}^k\right)^2\right]$$

where $\overline{U}$ and S are respectively the empirical mean and the empirical variance of the aggregated family of observations $\{U^k_1, \ldots, U^k_{n_k} : 1 \leq k \leq K\}$

The previous lemma casts some light on (3.17) and (3.18): those formulas correspond to the case where all the leaves of all the trees $\mathcal{T}_k$ contain the same number of observations from the bootstrap n-samples.

Remark 3.4.2. Let us emphasize two points:

- One could naturally apply exactly the formulas given in Lemma 3.4.1 simply by calculating

the proportions π_k. However, quick experiments seemed to show that the exact formulas do not significantly enhance the quality of the final bagged estimator.

• The total variance equals the mean of the partial variances iff all the partial means coincide, and it is larger otherwise. So, the CART procedure will generally overestimate the regression function $f^\star_{\text{se}}$, as will the Bagging procedure applied to CART regression trees.

Bagging variable importance

Assume that the bagged weak learner procedure $\mathcal{W}$ includes calculus of variables importance. For instance, CART provides the vector Δ_δ of variables importance as described in Section 3.3.4. Then the bagging procedure refines the calculus thanks to its bootstrap scheme that averages the single vectors of importance Δ_δ^k into a final vector $\Delta_\delta^{\text{ba}}$:

$$\Delta_\delta^{\text{ba}} = K^{-1} \sum_{k=1}^{K} \Delta_\delta^k.$$

Heuristically, $\Delta_\delta^{\text{ba}}$ is less instable than Δ_δ computed on a single weak learner. Thus, it provides safer informations.

Bagging as a benchmark

Input:	The learning set $\mathcal{L}_n$, the CART procedure $\mathcal{W}$, $0 < \delta \leq 1$.
Initialization:	Set $k = 1$, $\Delta_\delta = 0$.
Loop:	If $k > K$, exit loop. Draw a bootstrap n-sample uniformly with replacement from $\mathcal{L}_n$. Apply $\mathcal{W}$ and get a regression tree $\widehat{f}^k$ and a vector of variables importance Δ_δ^k. Update $k = k + 1$, $\Delta_\delta = \Delta_\delta + \Delta_\delta^k$.
Termination:	Aggregate $\widehat{f}^1, \ldots, \widehat{f}^K$ according to (3.17) and (3.18).
Output:	The bagged estimator $\widehat{f}^{\text{ba}}$ and the vector of variable importance $\Delta_\delta^{\text{ba}}$.

Table 3.3 – Bagging CART regression trees.

In the classical framework of homoscedastic models, bagged regression trees are known to improve (sometimes dramatically) the performance of a single tree. Thus, Bagging is a benchmark for any other method, including Boosting. We shall use Bagging for heteroscedastic models in Chapter 4. It will enhance the results of single CART procedures. Comparison with Boosting will be performed. Furthermore, we shall derive from Bagging some interesting variables importance.

We sum up the Bagging procedure applied to CART regression trees in Table 3.3.

3.4.3. Boosting

Outline

While the Bagging bootstrap resample rule at each iteration is uniform, the Boosting procedure relies on an adaptive bootstrap resample rule, whose weights at each iteration depend on past errors. That is, Boosting is an Arcing procedure:

- Adaptive resampling step: the construction of weak learners according to $\mathcal{W}$ is done repeatedly on the basis of bootstrap n-samples drawn with replacement from the original set $\mathcal{L}_n$. The weights of the bootstrap draws are initially uniform, then updated depending on the performance of the estimator at the previous step. An updating aims at focusing the new weak learner on those examples that have been previously poorly predicted, while almost forgetting the others (thanks to large weights for the first ones and small for the others). In other words, it makes the weak learner to localize on some parts of the set $\mathcal{X}$.
- Combination step: it relies on a weighted median of the already built estimators. The corresponding weights allow to take into account the local nature of the estimators. Indeed, the bootstrap procedure yields that each of them performs well locally, though poorly globally.

Here again, it seems that Boosting in heteroscedastic models is an original procedure, as Bagging and CART regression trees in the latter framework.

We adapt the procedure proposed by Drucker (1997) for homoscedastic model.

Detailed description

At initialization, an estimator $\widehat{f}^1$ is constructed on the basis of a bootstrap uniform n-sample drawn from $\mathcal{L}_n$ with replacement. Let us assume that we have already built k estimators. We denote ω_i^k the weight associated with the observation $Z_i \in \mathcal{L}_n$ for any $1 \leq i \leq n$ at the step k. Particularly, $\omega_i^1 = n^{-1}$ for any i. As announced earlier, the updating of the weights depends on the performances of $\widehat{f}^k$. We have to consider separately our two crucial examples:

Homoscedastic model: we compute for each $Z_i \in \mathcal{L}_n$ the *regret* $0 \leq r_i^k \leq 1$ of the i-th observation for the k-th predictor

$$r_i^k = \frac{\gamma_1(\widehat{f}^k, Z_i)}{\max_{1\leq j\leq n} \gamma_1(\widehat{f}^k, Z_j)} = \frac{\left(\widehat{f}^k(X_i) - Y_i\right)^2}{\max_{1\leq j\leq n} \left(\widehat{f}^k(X_i) - Y_i\right)^2}. \tag{3.19}$$

The regret quantifies the quality of the prediction wrt the underlying contrast in proportion to the worse case. The averaged regret $\overline{r}^k$ is the weighted mean of the latter,

$$\overline{r}^k = n^{-1} \sum_{i=1}^{n} \omega_i^k r_i^k. \tag{3.20}$$

It yields the nonnegative confidence ζ^k in the predictor $\widehat{f}^k$, according to

$$\zeta^k = \frac{\overline{r}^k}{1 - \overline{r}^k}. \tag{3.21}$$

Observe that low ζ^k corresponds to high confidence in the prediction. Now, if ζ^k equals 1 or is larger, or equivalently if $\overline{r}^k$ equals $1/2$ or is larger, the iterative procedure is stopped. Otherwise, the weights are adapted along the rule

$$\omega_i^{k+1} \propto \omega_i^k \left(\zeta^k\right)^{1-r_i^k}. \tag{3.22}$$

then renormalized in order to sum to one.

Remark 3.4.3.

- The latter original procedure (*i.e.* for homoscedastic models) follows the lines of the AdaBoost algorithm by Freund and Schapire (1996) in a classification framework. There, the regrets, averaged regret and confidence have a more natural interpretation.
- The updating rules are derived both from individual (through r_i^k) and collective (through ζ^k) performances.
- Besides, let us emphasize that $0 \leq \zeta^k < 1$, so that the smaller the regret r_i^k, the larger the updated weight ω_i^{k+1} and the more likely the corresponding observation will belong to the next bootstrap learning n-sample.

Heteroscedastic model: we also compute for each $Z_i \in \mathcal{L}_n$ the *regret* $0 \leq r_i^k \leq 1$ of the i-th observation for the k-th predictor. Its definition is less obvious than for homoscedastic models, since it requires recentering in order to get nonnegative regrets. Let us denote for convenience, for any $x \in \mathcal{X}$, $t = \{x\}$,

$$\begin{aligned} g(x, \mathcal{L}_n) &= \inf_{(m,\sigma^2)} \left\{ n^{-1} \sum_{i=1}^{n} \left[\frac{(Y_i - m)^2}{\sigma^2} + \log \sigma^2 \right] \mathbb{1}\{X_i = x\} \right\} \\ &= n^{-1} \sum_{i=1}^{n} \left[\frac{\left(Y_i - \widehat{v}_{\mathrm{m}}(t)\right)^2}{\widehat{v}_{\mathrm{se}}(t)^2} + \log \widehat{v}_{\mathrm{se}}(t)^2 \right] \mathbb{1}\{X_i \in t\} \\ &= \frac{n_t}{n} \left(1 + \log \widehat{v}_{\mathrm{se}}(t)^2\right), \\ \text{and} \quad h^k(x, \mathcal{L}_n) &= n^{-1} \sum_{Z \in \mathcal{L}_n} \gamma_2(\widehat{f}^k, Z) \mathbb{1}\{X = x\} \\ &= n^{-1} \sum_{Z \in \mathcal{L}_n} \left\{ \frac{(Y - \widehat{f}_{\mathrm{m}}^{\,k}(X))^2}{\widehat{f}_{\mathrm{se}}^{\,k}(X)^2} + \log \widehat{f}_{\mathrm{se}}^{\,k}(X)^2 \right\} \mathbb{1}\{X = x\}. \end{aligned}$$

Then, one can define two regrets according to the definitions below (up to renormalization constants so that $r_{\varepsilon,i}^k$ is bounded above by 1 with equality for at least one example), where $\varepsilon = 0$ for the first definition, $\varepsilon = 1$ for the second one,

$$r_{\varepsilon,i}^k \propto h^k(X_i, \mathcal{L}_n) - \left\{ (1-\varepsilon) \min_{1 \leq j \leq n} h^k(X_j, \mathcal{L}_n) + \varepsilon \min_{1 \leq j \leq n} g(X_j, \mathcal{L}_n) \right\}. \tag{3.23}$$

Here again, an averaged regret $\overline{r_\varepsilon}^k$ allows to define a confidence ζ_ε^k in the predictor $\widehat{f}^k$, with

$$\overline{r_\varepsilon}^k = n^{-1} \sum_{i=1}^{n} \omega_i^k r_{\varepsilon,i}^k \quad \text{and} \quad \zeta_\varepsilon^k = \frac{\overline{r_\varepsilon}^k}{1 - \overline{r_\varepsilon}^k}. \tag{3.24}$$

Finally, whenever $\zeta_\varepsilon^k < 1$, the weights are updated along

$$\omega_{\varepsilon,i}^{k+1} \propto \omega_{\varepsilon,i}^k \left(\zeta_\varepsilon^k\right)^{1 - r_{\varepsilon,i}^k}, \tag{3.25}$$

then renormalized in order to sum to one; otherwise, the iterative procedure is stopped.

Remark 3.4.4.

- The last point in Remark 3.4.3 is naturally still valid for this example.
- A simple graph casts some light on the behaviours of the two versions of regret above, see Figure 3.2. One can indeed see in the latter that for any $1 \leq i \leq n$,

$$r_{0,i}^k \vee \rho_1 \leq r_{1,i}^k,$$

where ρ_1 is a nonnegative constant independent of i which is positive iff

$$\min_{1 \leq j \leq n} g(X_j, \mathcal{L}_n) < \min_{1 \leq j \leq n} h^k(X_j, \mathcal{L}_n).$$

Otherwise, there exists some j such that

$$\begin{aligned} \widehat{f}_{\mathrm{m}}^{\,k}(X_j) &= \widehat{v}_{\mathrm{m}}(\{X_j\}) \quad \text{and} \\ \widehat{f}_{\mathrm{se}}^{\,k}(X_j) &= \widehat{v}_{\mathrm{se}}(\{X_j\}). \end{aligned}$$

Heuristically, the regrets $r_{0,i}^k$ make the Boosting procedure to focus more intensively on those examples that have been poorly predicted by the weak learner $\widehat{f}^k$ than does the same procedure for $r_{1,i}^k$.

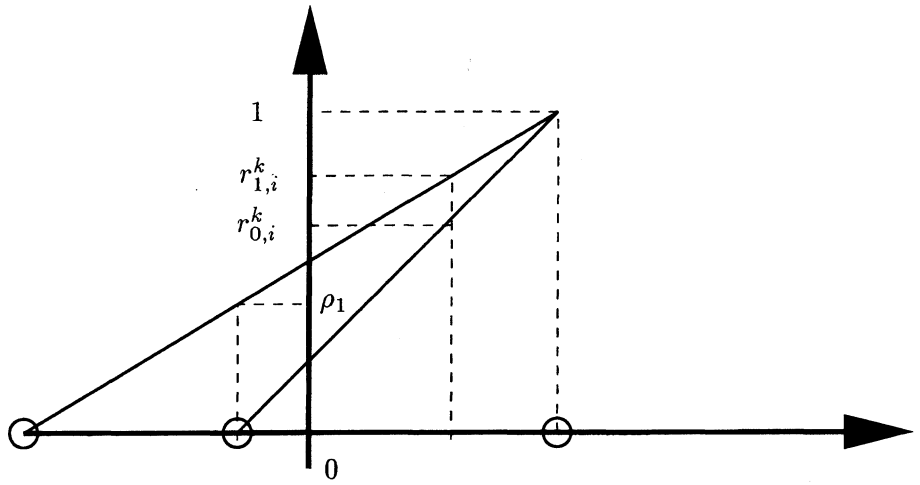


Figure 3.2 – Comparison of the two regrets for heteroscedastic models.

Assume now that we have constructed K elementary predictors. We then use the same definition for the final estimator than Drucker's (who once again drew his inspiration from the classification framework), both for homoscedastic and heteroscedastic models. Let x be an input: then $\widehat{f}^1(x), \ldots, \widehat{f}^K(x)$ are its individual predictions, and the output prediction $\widehat{f}^{\mathrm{bo}}(x)$ satisfies

Homoscedastic model:

$$\begin{aligned} \widehat{f}^{\mathrm{bo}}(x) = \widehat{f}_{\mathrm{m}}^{\mathrm{bo}}(x) &= \inf \left\{ y \in \mathbb{R} : \sum_{k=1}^{K} (-\log \zeta^k) \mathbb{1}\{\widehat{f}^k(x) \leq y\} \geq \frac{1}{2} \sum_{k=1}^{K} (-\log \zeta^k) \right\} \\ &= \inf \left\{ y \in \mathbb{R} : \sum_{k=1}^{K} (-\log \zeta^k) \mathbb{1}\{\widehat{f}_{\mathrm{m}}^{\,k}(x) \leq y\} \geq \frac{1}{2} \sum_{k=1}^{K} (-\log \zeta^k) \right\}. \quad (3.26) \end{aligned}$$

Heteroscedastic model: the same definition than above defines $\widehat{f}_{\mathrm{m}}^{\mathrm{bo}}(x)$, while the following holds true for $\widehat{f}_{\mathrm{se}}^{\mathrm{bo}}(x)$

$$\widehat{f}_{\mathrm{se}}^{\mathrm{bo}}(x) = \inf\left\{ y \in \mathbb{R} : \sum_{k=1}^{K} (-\log \zeta^k) \mathbb{1}\{ \widehat{f}_{\mathrm{se}}^{\ k}(x) \leq y \} \geq \frac{1}{2} \sum_{k=1}^{K} (-\log \zeta^k) \right\}. \tag{3.27}$$

When the weak learner procedure $\mathcal{W}$ is CART, then one can naturally replace $\widehat{f}_{\mathrm{m}}^{\ k}(x)$ and $\widehat{f}_{\mathrm{se}}^{\ k}(x)$ by $\widehat{v}_{\mathrm{m}}^{k}(t_k)$ and $\widehat{v}_{\mathrm{se}}^{k}(t_k)$, as soon as $x \in t_k \in \partial \mathcal{T}_k$ (which is the k-th tree associated with the k-th CART regression function $\widehat{f}^k$).

We sum up the Boosting procedure in Table 3.4. Finally, let us stress two points:

- *primo*, the Boosting procedure inherits the CART property of overestimation of the regression function $f_{\mathrm{se}}^{\star}$, as does Bagging.
- *secundo*, the nonuniform reweighting scheme does not allow a variable importance calculus as in Bagging.

Input:	The learning set $\mathcal{L}_n$, the CART procedure $\mathcal{W}$.
Initialization:	Set $k = 1$, $\omega_i^k = n^{-1}$ for all i.
Loop:	If $k > K$, exit loop. Draw a bootstrap n-sample along (ω_i^k) with replacement from $\mathcal{L}_n$. Apply $\mathcal{W}$ and get a regression tree $\widehat{f}^k$. Compute r_i^k, $\overline{r}^k$, ζ^k and ω_i^{k+1} according to (3.19, 3.20, 3.21, 3.22) or (3.23, 3.24, 3.25). If $\zeta^k \geq 1$, update $K = k$, exit loop. Update $k = k + 1$.
Termination:	Aggregate $\widehat{f}^1, \ldots, \widehat{f}^K$ according to (3.26) and (3.27).
Output:	The boosted estimator $\widehat{f}^{\mathrm{bo}}$.

Table 3.4 – Boosting CART regression trees.

3.4.4. Validation

Let us recall that $\mathcal{C}_m$ denotes a test sample, which is independent of the learning set $\mathcal{L}_n$. The validation simply relies on the calculation of the empirical contrast $\mathcal{E}_{\gamma_2}(\widehat{f}, \mathcal{C}_m)$ for $\widehat{f}$ equal to $\widehat{f}^{\mathrm{ba}}$ or $\widehat{f}^{\mathrm{bo}}$. See Section 3.2.3 for a glimpse of justification and Chapter 4 for some examples.

4

Eléments de raffinement de localisation

Résumé

Ce chapitre est dédié à l'application à notre problème initial de la méthode que nous avons soigneusement construite dans les trois chapitres précédents. Nous tâchons de mettre en valeur sa flexibilité et d'évaluer ses performances. Nous l'utilisons pour cela sur six jeux de données différents. Les résultats sont présentés de façon qu'une lecture transversale en soit aisée. Nous concluons par une discussion de ceux-ci. Un bref résumé en anglais conclut le chapitre.

Abstract

This chapter is finally dedicated to the application of the method we have elaborated along the previous chapters to our original problem. We aim at showing that it is flexible and we tackle the evaluation of how well it performs. Thus we apply it to various datasets. The yielded results are presented in such a way that comparison between examples is easy. We conclude by a discussion on those results. The whole is summarized in english.

Au menu

4.1. Introduction **131**
4.1.1. Vue d'ensemble 131
4.1.2. Préparatifs 132
4.2. Tranche horaire matinale **137**
4.3. Tranche horaire de la mi-journée **141**
4.4. Tranche horaire de l'après-midi **145**
4.5. Tranche horaire de la soirée **149**
4.6. Journée complète **153**
4.7. Données HC2 . . . . . **155**
4.8. Synthèse **159**
4.8.1. Performances des procédures Bagging et Boosting 159
4.8.2. Importance des variables INSEE 159
4.9. English summary **163**
4.9.1. Outline 163
4.9.2. Some results 164

4.1. Introduction

4.1.1. Vue d'ensemble

Ce chapitre est consacré aux applications de la méthode que nous avons élaborée plus tôt. Six sections suivent, chacune dédiée à l'application à un jeu de données particulier. Ainsi,

- la Section 4.2 consiste en l'exposition des résultats obtenus lorsque les données téléphoniques sont les quantités de trafic (en Erlang, calculées à partir du jeu de données Cigale, voir la Section 1.2.2 du Chapitre 1) écoulées sur des périodes de dix minutes entre neuf et onze heures — on parlera de la tranche horaire de la matinée ;
- la Section 4.3 concerne les résultats de l'application aux mêmes données entre onze et treize heures — on parlera de la tranche horaire de la mi-journée ;
- la Section 4.4 correspond encore à ces résultats pour les mêmes données, cette fois prises entre seize et dix-huit heures — on parlera de la tranche horaire de l'après-midi ;
- la Section 4.5 propose ces résultats pour les données restreintes à la plage de vingt à vingt-deux heures — on parlera de la tranche horaire de la soirée ;
- la Section 4.6 offre un aperçu des résultats de la méthode appliquée aux données comme définies plus haut, mais sur une plage horaire plus étendue, s'étalant entre douze et vingt-deux heures — on parlera (ce n'est pourtant pas très révélateur) des données de la journée complète ;
- la Section 4.7 est enfin dévolue à l'exposé des résultats de la méthode lorsqu'appliquée aux données de trafic hebdomadaire HC2, voyez la Section 1.2.3 du Chapitre 1 — on parlera des données HC2.

Ces diverses applications permettent d'illustrer le fonctionnement de notre méthode. On s'attachera particulièrement

- à donner un aperçu des estimateurs qu'elle délivre en sortie ;
- à leurs performances grâce à une procédure de validation que nous décrivons plus bas ;
- à une formulation finale d'un résultat inédit concernant l'importance des variables explicatives socio-démographiques et culturelles ;
- à la mise en lumière de la *flexibilité* de la méthode que nous avons élaborée, *i.e.* de sa capacité à être exploitée dans des cadres de travail variés.

On s'efforcera de présenter dans la Section 4.8 une vue d'ensemble synthétique sur la somme des informations contenues dans les sections précédentes.

Un bref résumé en anglais clora ce chapitre.

4.1.2. Préparatifs

A propos de l'implémentation

Nous avons entièrement programmé la procédure élémentaire originale de CART pour les arbres de régression dans le modèle hétéroscédastique. Le logiciel `Matlab` nous a semblé tout indiqué à l'époque de l'implémentation, en raison notamment de notre familiarité avec lui. Le code exploite autant que possible les affinités vectorielles de `Matlab`. Les routines les plus sollicitées ont été converties en langage C afin de limiter les temps d'exécution. Néanmoins, la procédure est assez lente. Ainsi, la construction de la forêt d'arbres de régression CART comme présentée dans le Chapitre 3, plus particulièrement dans les Tableaux 3.1 et 3.2 du même chapitre, nécessite entre 30 et 400 secondes sur les divers exemples ci-dessous. Cette lenteur nous a malheureusement empêché de procéder à un certain nombre de simulations dont les résultats auraient certainement apporté de nombreux complèments et informations. Les résultats qui suivent sont peu spectaculaires mais de notre avis encourageants. Aussi, nous nous consacrerons plus tard à une nouvelle programmation dans un langage finalement mieux adapté.

Données de trafic

On calcule facilement les quantités de trafic écoulé sur $\Delta_0 = 10$ minutes grâce à la formule (1.1) du Chapitre 1, et ce toutes les Δ_0 minutes entre six heures du matin et onze heures du soir (la plage totale d'observation Cigale) et pour toutes les cellules de Paris.

Etant donnée une des plages horaires d'intérêt de matinée, mi-journée, après-midi et soirée, on peut alors constituer une matrice de données d'entrée qui compte autant de lignes qu'il y a de cellules et dont chaque ligne contient les douze quantités de trafic écoulé en Δ_0 minutes entre $H \in \{9, 11, 16, 20\}$ et $H + 12 \times \Delta_0$ heures.

Remarque 4. Nous soulignons par ailleurs que nous admettons l'absence d'effet saisonnier sur les plages horaires de matinée, mi-journée, après-midi et soirée pour les quantités de trafic ecoulé sur dix minutes.

Pour l'étude à la journée exposée dans la Section 4.6, la matrice d'entrée est la concaténation verticale de cinq matrices du même type que celles décrites dans le paragraphe précédent, construites pour les plages horaires de midi à deux heures, deux à quatre heures, quatre à six heures, six à huit heures et huit à dix heures du soir. Cette fois-ci, chaque cellule se voit associer exactement cinq lignes dans la matrice d'entrée, à raison de douze quantités de trafic par ligne.

La matrice d'entrée pour l'application aux données HC2 est prête d'emblée : on compte une ligne par cellule exactement, chaque ligne étant composée de vingt-et-une quantités de trafic (une par semaine d'observation de mars à juillet 2002).

Remarque 5. Nous avons intentionnellement employé l'expression de « matrice de données » afin de mettre en lumière le fait que nous associons bien à chaque cellule ses 12 observations sur la tranche horaire (pour les quatre premiers exemples), et non pas leur moyenne. Si nous avions fait ce choix, la variabilité des quantités de trafic écoulé n'aurait été que le fruit de la variabilité intercellulaire. Au contraire donc, la variabilité est pour nous la somme de cette variabilité intercellulaire et d'une variabilité intracellulaire (ou autrement dit, d'une variabilité due à l'échantillonage sur chaque cellule).

Données INSEE

Pour commencer, il faut préciser que nous nous sommes souciés de réduire à moindre frais les temps de calculs des constructions élémentaires des arbres de régression CART. Ceux-ci dépendent proportionnellement du produit du nombre de cellules par le nombre de variables explicatives socio-démographiques et culturelles. Par ailleurs, le nombre de variables explicatives dans un cadre de régression doit être aussi réduit que possible pour un nombre d'observation des variables à expliquer fixé. Aussi, nous avons ainsi entrepris de condenser les données explicatives tout en tâchant de conserver autant de leur pouvoir d'explication que possible.

Les données socio-démographiques et culturelles ont été soigneusement présentées dans l'Interlude 2. Chaque cellule se voit associer, d'une part les 24 données issues de la base ILOTS15, à raison de

- 3 fois 6 données de population par sexe et tranche d'âge (population féminine exclusivement, masculine exclusivement, totale, sans considération d'âge et par tranches d'âges de 0 à 19, 20 à 39, 40 à 59, 60 à 74, 75 ans et plus) ;
- 5 données relatives à l'ensemble des logements par catégorie (sans distinction, principal, secondaire, occasionnel et vacant) ;
- 1 donnée relative au nombre de personnes vivant dans les résidences principales ;

et d'autre part, les 206 quantités calculées à partir du relevé SIRENE, à raison de

- 103 nombre d'établissements localisés dans la cellule pour les 103 code APET700 présentés dans les Tableaux 2.5, 2.6 et 2.7 de l'Interlude 2 ;
- 103 effectifs salariés cumulés approximatifs pour chaque code APET700.*

La réduction des données INSEE repose fondamentalement sur deux constats.
Primo, les populations féminines et masculines sont positivement très corrélées avec la population totale pour toutes les tranches d'âges. D'autre part, les nombres de logements sans distinction et le nombre de logements principaux coïncident presque sur Paris.

> Nous tirons partie de ces remarques en décidant de ne conserver que les populations totales par tranches d'âges (soit 6 indicateurs) et les nombres totaux de logements (soit 1 indicateur).

Secundo, nous avons déjà évoqué dans l'Interlude 2 que les codages APET60 et APET700 sont hiérarchiques, *i.e.* que les deux premiers chiffres qui constituent les codes APET700 (format `Chiffre Chiffre Chiffre Lettre`) codent au sens de l'APET60.

*Le répertoire SIRENE ne communique que des tranches d'effectif salarié que l'on ne peut pas sommer directement. Nous avons choisi d'affecter à chaque établissement la moyenne des nombres minimal et maximal de sa tranche. Nous associons alors à une zone la médiane de ces effectis moyens pour tous les établissements de même code APET700 localisés dans la zone.

> Nous tirons partie à notre tour de cette hiérarchie en décidant de considérer les pseudo–codes APET obtenus par lecture seule des trois premiers chiffres du code APET700 (soit 41 codes, d'où le double d'indicateurs en comptant aussi les effectifs approximatifs associés).

Cette solution a le mérite d'être consistante avec les notions d'APET60 et APET700 dont elle apparaît comme une sorte d'interpolation. Ces choix de réduction font tomber à 89 le nombre de variables explicatives.

Enfin, on introduira une variable explicative supplémentaire dans la Section 4.6, qui rendra compte de la tranche horaire d'observation des quantités de trafic écoulé.

Remarque 6. Notre implémentation de l'algorithme CART nous contraint à considérer l'indicateur de tranche horaire comme une variable ordinale et non catégorielle. Concrètement, les séparations de variables suivant cet indicateur seront nécessairement de la forme $\mathbb{1}\{H > C\}$, là où l'on aurait préféré la forme générale $\mathbb{1}\{H \in \cup_i \{H_i\}\}$.

Dans la suite, nous dénoterons `#Chiffre Chiffre Chiffre` l'effectif cumulé approximatif du pseudo–code APET codé `Chiffre Chiffre Chiffre`. La population totale sera notée `PT` et les populations totales par tranches d'âges seront notées respectivement `PT0`, `PT20`, `PT40`, `PT60` et `PT75` (dans le même ordre que l'énumération ci-dessus). Le nombre total de logements sera noté `LOG` et la variable horaire pour la Section 4.6 sera notée `H`.

Composition des sections suivantes

Les six prochaines sections sont composées suivant un même canevas qui en facilite une lecture transversale.

Ainsi, pour chacun des exemples d'application de la méthode, nous avons procédé aux opérations résumées dans le Tableau 4.1. Il faut préciser que la procédure de remise à jour des poids que nous avons sélectionnée pour le Boosting correspond au choix de $\varepsilon = 0$ dans l'équation (3.23) de définition du regret, *i.e.* au Boosting *radical* pour lequel l'algoritme d'apprentissage pour la nouvelle itération se concentre le plus fortement sur les observations mal prédites à l'itération courante.

Ainsi, nous pouvons produire dans chaque section :

- Un exemple d'arbre de régression CART construit sur les données de la section. Il est accompagné par un graphique constitué du nuage de points des couples de moyenne et écart-type assigné à chaque feuille, ainsi que par un tableau contenant, pour chaque feuille de l'arbre, le couple associé et la proportion de cellules lui appartenant.
- Un graphique présentant l'évolution en fonction du nombre d'itérations des contrastes de validation moyennés pour dix procédures Bagging et pour dix procédures Boosting appliquées à des échantillons Bootstrap (uniformes, avec remise) de l'ensemble d'apprentissage. Une barre horizontale révèle le contraste de validation moyen de dix arbres de régression CART seul construits sur chacun des échantillons Bootstrap.
- Deux tableaux contenant les pourcentages moyens (et les écarts-types correspondants) d'amélioration de CART seul par les procédures Bagging et Boosting au sens du contraste de validation et de l'amélioration de l'écart norme 2 de validation que nous définissons ci-dessous.

Entrée :	Un ensemble d'apprentissage $\mathcal{A}_{m+n}$
Initialisation :	Soit $k = 1$.
Boucle :	Si $k > 10$, sortie de la boucle.
	Tirage aléatoire uniforme sans remise dans $\mathcal{A}_{m+n}$ d'un ensemble d'apprentissage $\mathcal{L}_n^k$ à hauteur de 80% du jeu complet.
	Ensemble de validation $\mathcal{C}_m^k = \mathcal{A}_{m+n} \setminus \mathcal{L}_n^k$.
	Construction d'un arbre CART $\mathcal{T}_k$ sur $\mathcal{L}_n^k$.
	50 itérations Bagging Ba_k sur $\mathcal{L}_n^k$.
	50 itérations Boosting Bo_k sur $\mathcal{L}_n^k$.
	Calcul des contrastes de validation des $\mathcal{T}_k$, Ba_k et Bo_k sur $\mathcal{C}_m^k$.
	Calcul des écarts norme 2 de validation pour la moyenne.
	Mise à jour $k = k + 1$.
Conclusion :	Calcul des pourcentages moyens de réduction du contraste de validation d'un arbre CART seul par Bagging, Boosting.
	Calcul des pourcentages moyens de réduction des écarts norme 2 de validation pour la moyenne.
	Extraction de l'importance des variables évaluée par la procédure CART seul.
	Extraction de l'importance agrégée des variables dues aux procédures Bagging.
Sortie :	Les six graphiques et tableaux de la section correspondant aux données d'entrée, ainsi que les tableaux d'importance des variables en Annexe.

Tableau 4.1 – Opérations auxquelles on procède dans chaque exemple pour préparer les six sections suivantes et l'Annexe. Pour la section dévolue aux données Cigale à la journée, on n'a pu faire qu'une itération des procédures Bagging et Boosting à cause des durées de calcul.

- Enfin, deux graphiques côte à côte où l'on représente, pour une des 10 itérations exposées dans le Tableau 4.1, les moyennes empiriques des observations (en abscisse) contre les moyennes prédites (en ordonnée) pour chaque cellule de l'ensemble de validation et pour les régresseurs Bagging (graphique gauche) et Boosting (graphique droit).

En guise de conclusion de cette section, nous précisons en quoi consistent les deux évaluations de l'amélioration des performances d'un arbre CART seul due aux procédures Bagging et Boosting.

Réduction du contraste de validation : à la $k$ième itération de l'algorithme du Tableau 4.1, on évalue (avec les notations du Chapitre 3 et celles du Tableau 4.1) les contrastes de validation des fonctions de régression associées à $\mathcal{T}_k$, Ba_k et Bo_k (que l'on note $\widehat{f}_k$, $\widehat{f}_k^{\mathrm{ba}}$ et $\widehat{f}_k^{\mathrm{bo}}$), soit respectivement

$$\mathcal{E}_{\gamma_2}(\widehat{f}_k, \mathcal{C}_m^k), \quad \mathcal{E}_{\gamma_2}(\widehat{f}_k^{\mathrm{ba}}, \mathcal{C}_m^k) \quad \text{et} \quad \mathcal{E}_{\gamma_2}(\widehat{f}_k^{\mathrm{bo}}, \mathcal{C}_m^k).$$

De plus, si $\mathcal{C}_m^k = \{Z_i = (X_i, Y_i) : i = n+1, \ldots, n+m\}$ et si $\{X_i : i = n+1, \ldots, n+m\} = \cup_{l=1}^{L} t_l$, alors le minimum absolu du contraste de validation est

$$\mathcal{E}_{\gamma_2}(\widehat{\varphi}_k, \mathcal{C}_m), \quad \text{où} \quad \widehat{\varphi}_k(x) = \sum_{l=1}^{L} \widehat{v}(t_l) \mathbb{1}\{x \in t_l\}$$

(pour la définition de $\widehat{v}$, voir (3.8)).

On peut dès lors définir l'amélioration en contraste due au Bagging pour cette itération (pour le Boosting, faire la substitution évidente) comme le rapport

$$\frac{\mathcal{E}_{\gamma_2}(\widehat{f}_k, \mathcal{C}_m^k) - \mathcal{E}_{\gamma_2}(\widehat{f}_k^{\mathrm{ba}}, \mathcal{C}_m^k)}{\mathcal{E}_{\gamma_2}(\widehat{f}_k, \mathcal{C}_m^k) - \mathcal{E}_{\gamma_2}(\widehat{\varphi}_k, \mathcal{C}_m)} \times 100. \tag{4.1}$$

La réduction finale est la moyenne sur les dix itérations des quantités ci-dessus.

Réduction des écarts norme 2 de validation pour la moyenne : cette quantité mesure un gain apporté par les procédures Bagging et Boosting relativement à l'estimateur naïf qui associe à toute cellule de l'ensemble de validation la moyenne des quantités de trafic sur toutes les cellules de l'ensemble d'apprentissage, que l'on note $\overline{Y}_n$. Elle est l'équivalente de la précédente lorsque l'on remplace la fonction contraste γ_2 par γ_1 et l'arbre CART $\mathcal{T}_k$ par un arbre à unique feuille.

Ainsi, la réduction de l'écart norme 2 de validation pour la procédure Bagging est donnée simplement par la moyenne sur les itérations des quantités

$$\frac{\sum_{i=n+1}^{n+m} \left(Y_i - \overline{Y}_n\right)^2 - \sum_{i=n+1}^{n+m} \left(Y_i - \widehat{f}_k^{\mathrm{ba}}(X_i)\right)^2}{\sum_{i=n+1}^{n+m} \left(Y_i - \overline{Y}_n\right)^2 - \sum_{i=n+1}^{n+m} \left(Y_i - \widehat{v}(\{X_i\})\right)^2} \times 100, \tag{4.2}$$

où $\widehat{v}(t)$ est la moyenne des Y_i tels que $X_i \in t$. La réduction pour le Boosting se déduit par substitution évidente.

4.2. Tranche horaire matinale

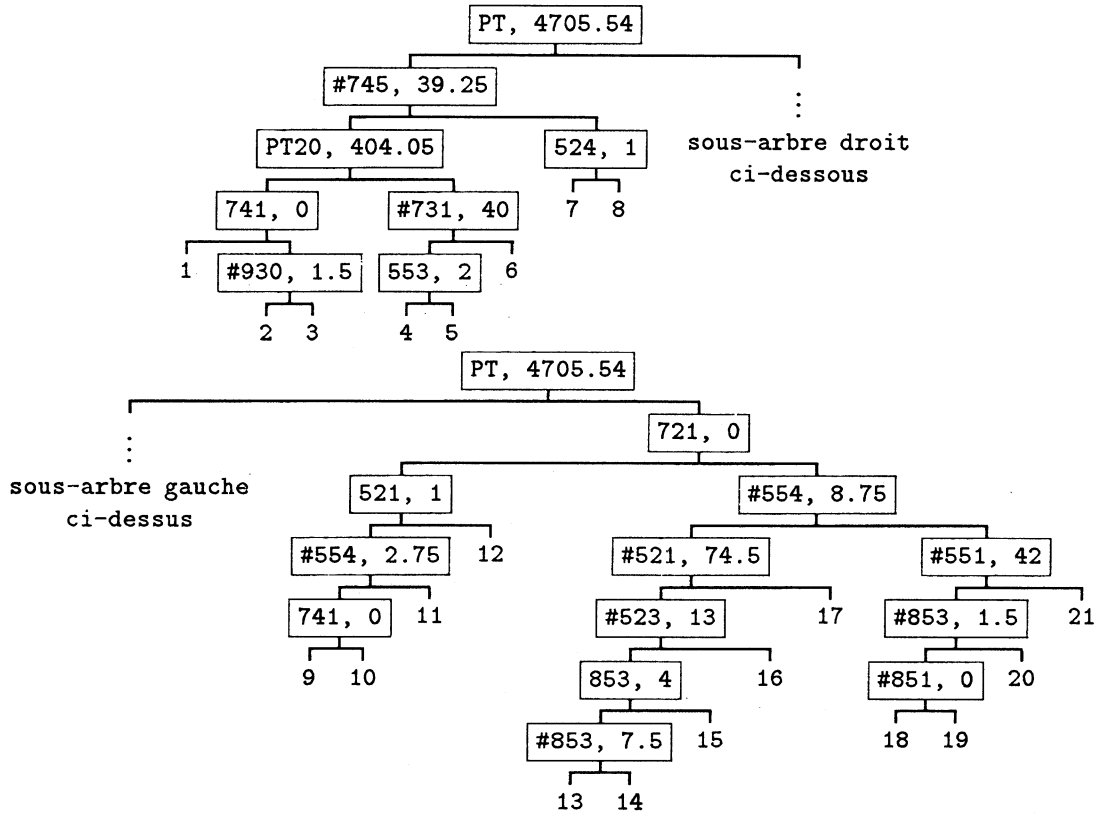
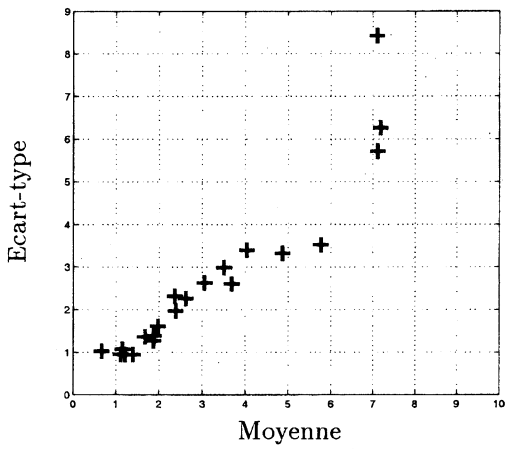


Figure 4.1 – Arbre de régression CART construit à partir des observations Cigale pour la matinée et les données INSEE associées. On indique à chaque nœud le code de l'indicateur INSEE de coupure et le seuil correspondant. On compte 21 feuilles. Les couples de moyenne et écart-type pour chaque feuille sont présentés dans la Figure 4.2. Les importances des variables calculées sur cet arbre sont exposées dans le Tableau C.2 de l'Annexe C.



f	m_f	σ_f	%	f	m_f	σ_f	%
1	2.4	2.3	6	12	1.2	0.9	1
2	1.2	1.1	12	13	2.4	2.0	21
3	1.9	1.3	6	14	3.5	3.0	3
4	2.6	2.3	11	15	1.1	0.9	1
5	2.0	1.6	17	16	3.7	2.6	6
6	4.9	3.3	1	17	0.7	1.0	1
7	7.2	6.3	1	18	1.4	0.9	1
8	1.9	1.4	1	19	3.1	2.6	3
9	4.0	3.4	3	20	5.8	3.5	2
10	7.1	5.7	2	21	7.1	8.4	1
11	1.7	1.4	1				

Figure 4.2 – Les couples de moyenne et écart-type pour les 21 feuilles de l'arbre de régression CART construit à partir des observations Cigale restreintes à la matinée et les données INSEE associées. On peut voir l'arbre dans la Figure 4.1. A gauche, les points du plan associés. A droite, les couples de moyenne m_f et écart-type σ_f pour chaque feuille f, ainsi que les proportions respectives de représentants par feuille.

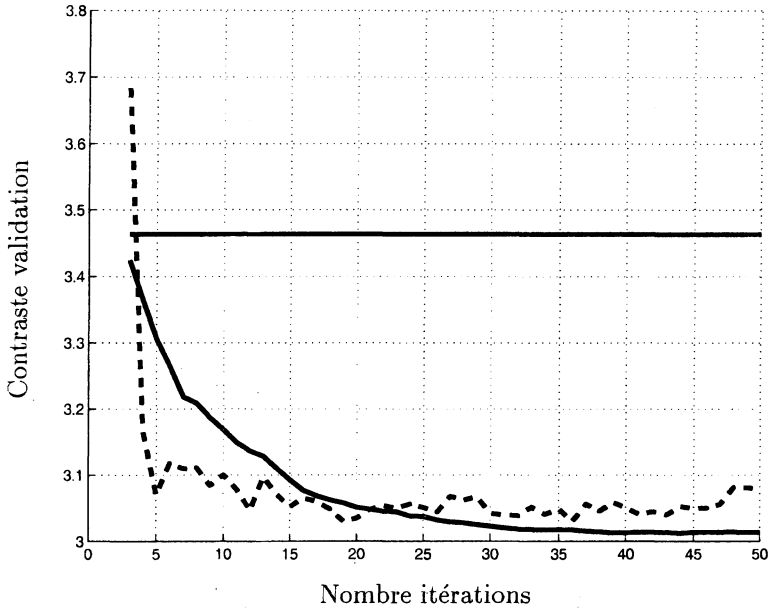


Figure 4.3 – Contraste empirique de validation moyenné sur dix procédures CART (barre horizontale), Bagging (tracé continu) et Boosting (tracé discontinu) pour un jeu d'observations de données Cigale restreintes à la matinée. Le nombre d'itérations est en abscisse. On omet les contrastes empiriques pour les deux premières itérations par souci de clarté de la figure. Les améliorations moyennes apportées par les deux procédures vis-à-vis d'un seul arbre de régression CART sont données dans les Tableaux 4.2 et 4.3.

Bagging	Boosting
14%±8	12%±8

(contraste)

Tableau 4.2 – Réductions du **contraste de validation** d'un seul arbre de régression CART dues aux procédures Bagging et Boosting pour les observations Cigale restreintes à la matinée. On les évalue par moyenne sur dix procédures chacune. Le premier nombre indique le gain moyen en pourcents ; le second (précédé par le signe ±), l'écart-type correspondant.

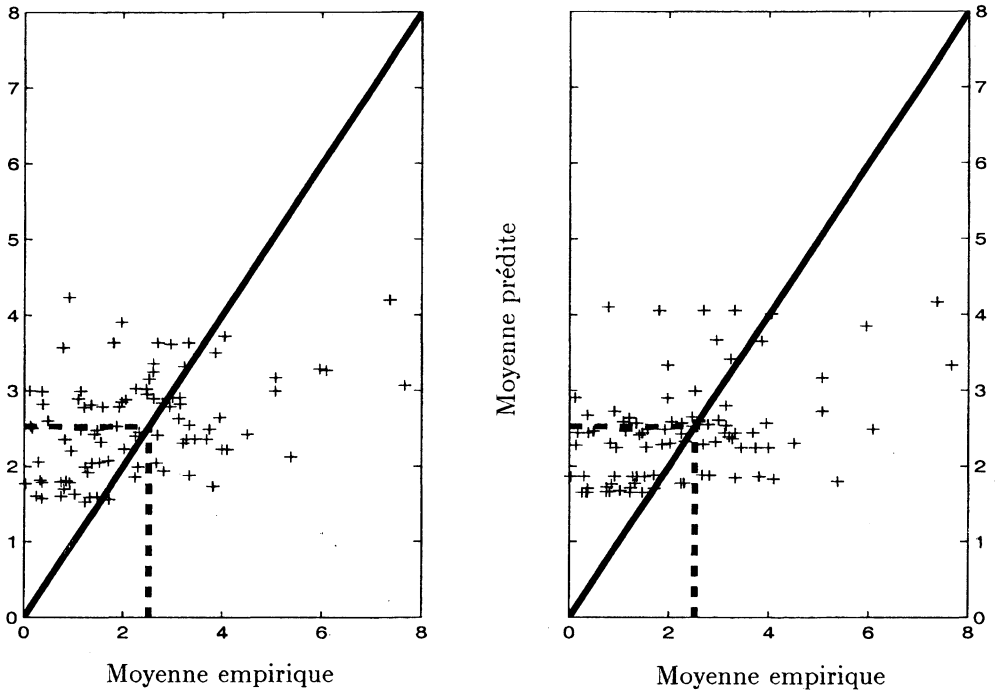


Figure 4.4 – Aperçu des résultats d'une unique procédure **Bagging** et d'une **Boosting** sur les dix chacune réalisées pour l'étude des observations Cigale restreintes à la matinée. En abscisse, les moyennes empiriques sur chaque cellule réservée à la procédure de validation ; en ordonnée, les moyennes prédites à chacune par la procédure **Bagging** (gauche) et la procédure **Boosting** (droite). La diagonale représente une portion de la droite de pente 1. Les traits discontinus indiquent la valeur de la moyenne des quantités de trafic sur toutes les observations de l'ensemble d'apprentissage. Les réductions des **écarts norme 2 de validation pour la moyenne** sont présentées dans le Tableau 4.3. On peut voir l'évolution des contrastes empiriques de validation moyennés sur les dix procédures dans la Figure 4.3. Les importances $\Delta_{\delta}^{\mathrm{ba}}$ ($\delta = 1,\ 0.9,\ 0.8$) des variables socio-démographiques et culturelles évaluées par moyennage des importances agrégées à chaque procédure Bagging sont exposées dans le Tableau C.3 de l'Annexe C.

Bagging	Boosting
10%±4	10%±5
(moyenne)	

Tableau 4.3 – Réductions des **écarts norme 2 de validation pour la moyenne** entre le modèle à moyenne unique et les modèles dus aux procédures Bagging et Boosting pour les observations Cigale restreintes à la matinée. On les évalue par moyenne sur dix procédures chacune. Le premier nombre indique le gain moyen en pourcents ; le second (précédé par le signe ±), l'écart-type correspondant.

4.3. Tranche horaire de la mi-journée

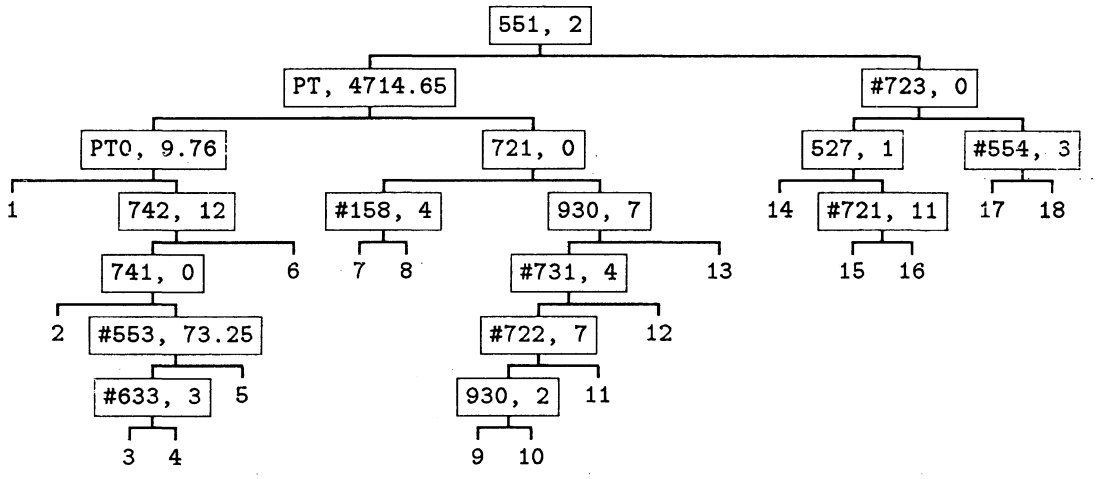
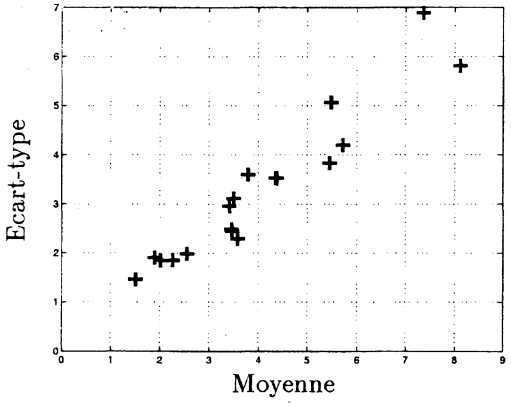


Figure 4.5 – Arbre de régression CART construit à partir des observations Cigale pour la mi-journée et les données INSEE associées. On indique à chaque nœud le code de l'indicateur INSEE de coupure et le seuil correspondant. On compte 18 feuilles. Les couples de moyenne et écart-type pour chaque feuille sont présentés dans la Figure 4.6. Les importances des variables calculées sur cet arbre sont exposées dans le Tableau C.5 de l'Annexe C.



f	m_f	σ_f	%	f	m_f	σ_f	%
1	1.5	1.5	3	10	3.4	3.0	7
2	3.8	3.6	5	11	4.4	3.5	9
3	2.6	2.0	22	12	2.0	1.8	3
4	3.5	2.5	11	13	5.7	4.2	5
5	4.4	3.5	2	14	3.5	2.4	8
6	5.4	3.8	2	15	3.6	2.3	1
7	5.5	5.1	5	16	8.1	5.8	1
8	1.9	1.9	1	17	3.5	3.1	4
9	2.3	1.9	5	18	7.4	6.9	6

Figure 4.6 – Les couples de moyenne et écart-type pour les 18 feuilles de l'arbre de régression CART construit à partir des observations Cigale restreintes à la mi-journée et les données INSEE associées. On peut voir l'arbre dans la Figure 4.5. A gauche, les points du plan associés. A droite, les couples de moyenne m_f et écart-type σ_f pour chaque feuille f, ainsi que les proportions respectives de représentants par feuille.

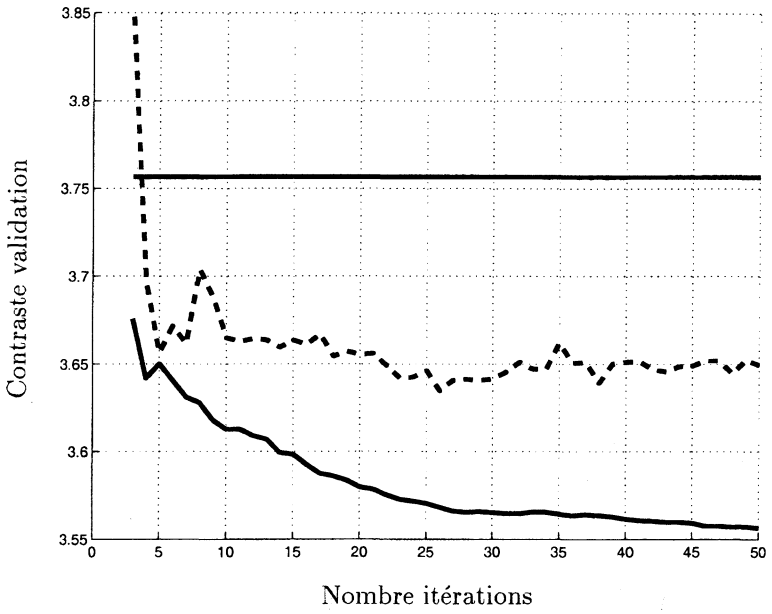


Figure 4.7 – Contraste empirique de validation moyenné sur dix procédures CART (barre horizontale), Bagging (tracé continu) et Boosting (tracé discontinu) pour un jeu d'observations de données Cigale restreintes à la mi-journée. Le nombre d'itérations est en abscisse. On omet les contrastes empiriques pour les deux premières itérations par souci de clarté de la figure. Les améliorations moyennes apportées par les deux procédures vis-à-vis d'un seul arbre de régression CART sont données dans les Tableaux 4.4 et 4.5.

Bagging	Boosting
10%±7	7%±7

(contraste)

Tableau 4.4 – Réductions du **contraste de validation** d'un seul arbre de régression CART dues aux procédures Bagging et Boosting pour les observations Cigale restreintes à la mi-journée. On les évalue par moyenne sur dix procédures chacune. Le premier nombre indique le gain moyen en pourcents ; le second (précédé par le signe ±), l'écart-type correspondant.

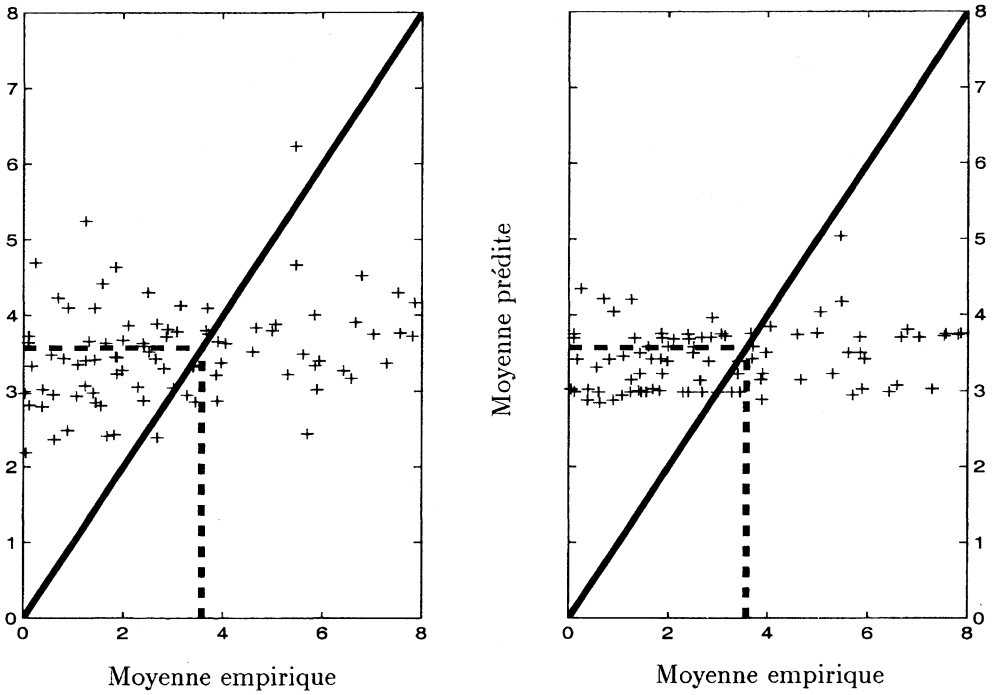


Figure 4.8 – Aperçu des résultats d'une unique procédure **Bagging** et d'une **Boosting** sur les dix chacune réalisées pour l'étude des observations Cigale restreintes à la mi-journée. En abscisse, les moyennes empiriques sur chaque cellule réservée à la procédure de validation ; en ordonnée, les moyennes prédites à chacune par la procédure **Bagging** (gauche) et la procédure **Boosting** (droite). La diagonale représente une portion de la droite de pente 1. Les traits discontinus indiquent la valeur de la moyenne des quantités de trafic sur toutes les observations de l'ensemble d'apprentissage. Les réductions des **écarts norme 2 de validation pour la moyenne** sont présentées dans le Tableau 4.5. On peut voir l'évolution des contrastes empiriques de validation moyennés sur les dix procédures dans la Figure 4.7. Les importances $\Delta_{\delta}^{\mathrm{ba}}$ ($\delta = 1,\ 0.9,\ 0.8$) des variables socio-démographiques et culturelles évaluées par moyennage des importances agrégées à chaque procédure Bagging sont exposées dans le Tableau C.6 de l'Annexe C.

Bagging	Boosting
10%±3	9%±3
(moyenne)	

Tableau 4.5 – Réductions des **écarts norme 2 de validation pour la moyenne** entre le modèle à moyenne unique et les modèles dus aux procédures Bagging et Boosting pour les observations Cigale restreintes à la mi-journée. On les évalue par moyenne sur dix procédures chacune. Le premier nombre indique le gain moyen en pourcents ; le second (précédé par le signe ±), l'écart-type correspondant.

4.4. Tranche horaire de l'après-midi

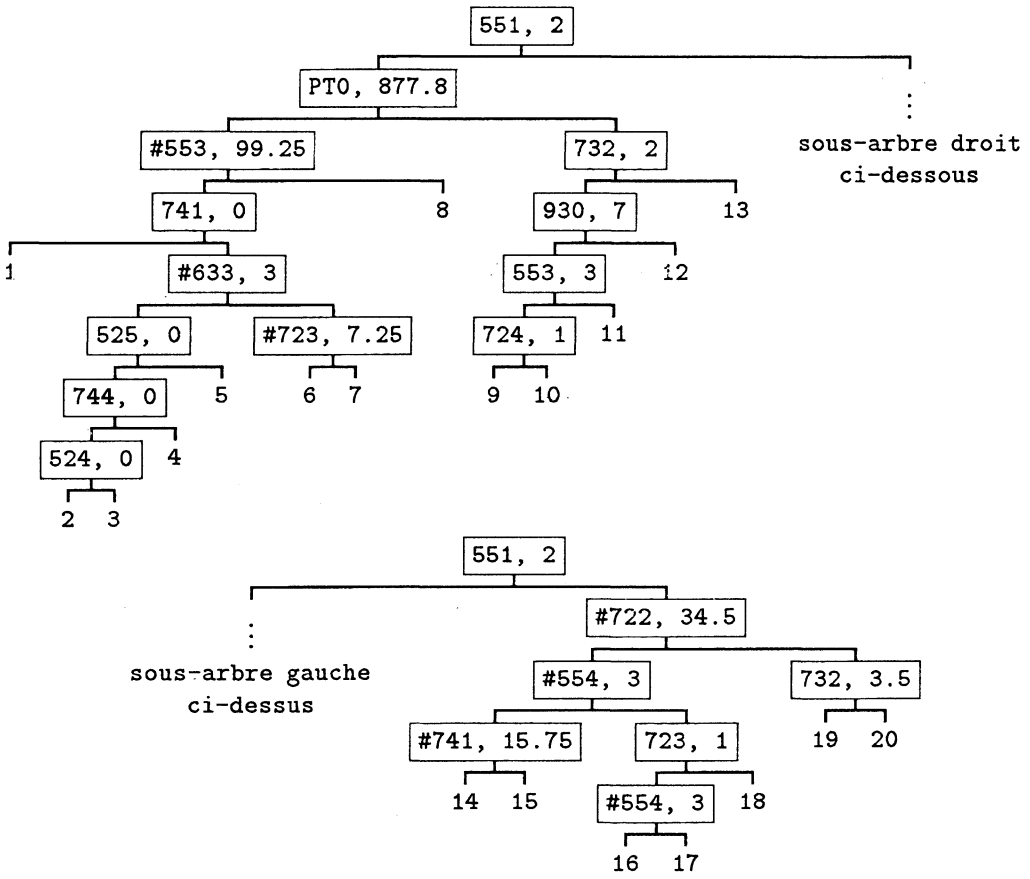
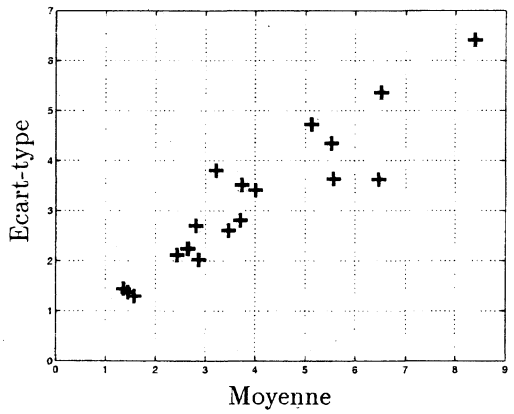


Figure 4.9 – Arbre de régression CART construit à partir des observations Cigale pour l'après-midi et les données INSEE associées. On indique à chaque nœud le code de l'indicateur INSEE de coupure et le seuil correspondant. On compte 20 feuilles. Les couples de moyenne et écart-type pour chaque feuille sont présentés dans la Figure 4.10. Les importances des variables calculées sur cet arbre sont exposées dans le Tableau C.8 de l'Annexe C.



f	m_f	σ_f	%	f	m_f	σ_f	%
1	3.7	3.5	7	11	4.0	3.4	15
2	1.4	1.4	1	12	6.5	5.4	4
3	3.2	3.8	4	13	2.8	2.7	3
4	2.6	2.3	7	14	3.7	2.8	5
5	2.9	2.0	13	15	1.6	1.3	1
6	3.5	2.6	12	16	5.6	3.6	4
7	6.5	3.6	1	17	2.7	2.2	1
8	5.5	4.3	2	18	8.4	6.4	3
9	5.1	4.7	10	19	9.2	8.2	4
10	2.4	2.1	1	20	1.4	1.4	1

Figure 4.10 – Les couples de moyenne et écart-type pour les 20 feuilles de l'arbre de régression CART construit à partir des observations Cigale restreintes à l'après-midi et les données INSEE associées. On peut voir l'arbre dans la Figure 4.5. A gauche, les points du plan associés. A droite, les couples de moyenne m_f et écart-type σ_f pour chaque feuille f, ainsi que les proportions respectives de représentants par feuille.

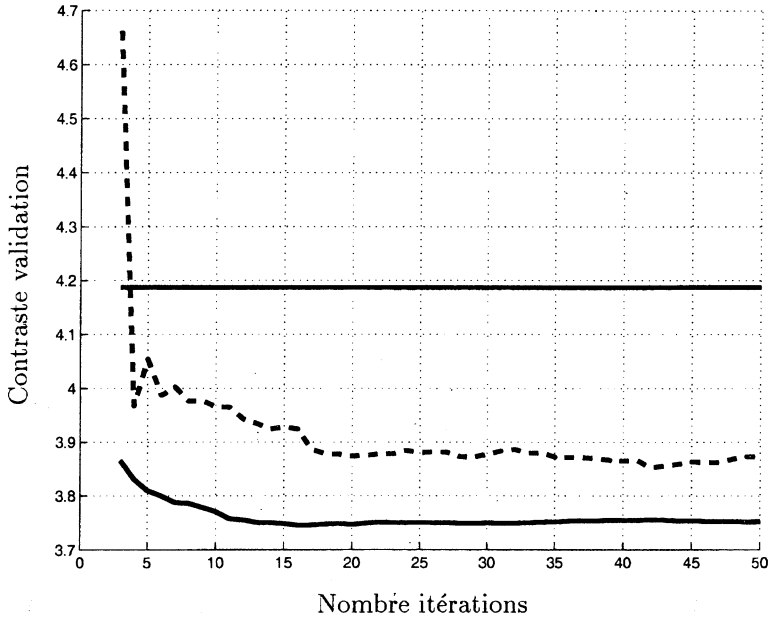


Figure 4.11 – Contraste empirique de validation moyenné sur dix procédures CART (barre horizontale), Bagging (tracé continu) et Boosting (tracé discontinu) pour un jeu d'observations de données Cigale restreintes à l'après-midi. Le nombre d'itérations est en abscisse. On omet les contrastes empiriques pour les deux premières itérations par souci de clarté de la figure. Les améliorations moyennes apportées par les deux procédures vis-à-vis d'un seul arbre de régression CART sont données dans les Tableaux 4.6 et 4.7.

Bagging	Boosting
12%±16	8%±17

(contraste)

Tableau 4.6 – Réductions du **contraste de validation** d'un seul arbre de régression CART dues aux procédures Bagging et Boosting pour les observations Cigale restreintes à l'après-midi. On les évalue par moyenne sur dix procédures chacune. Le premier nombre indique le gain moyen en pourcents ; le second (précédé par le signe ±), l'écart-type correspondant.

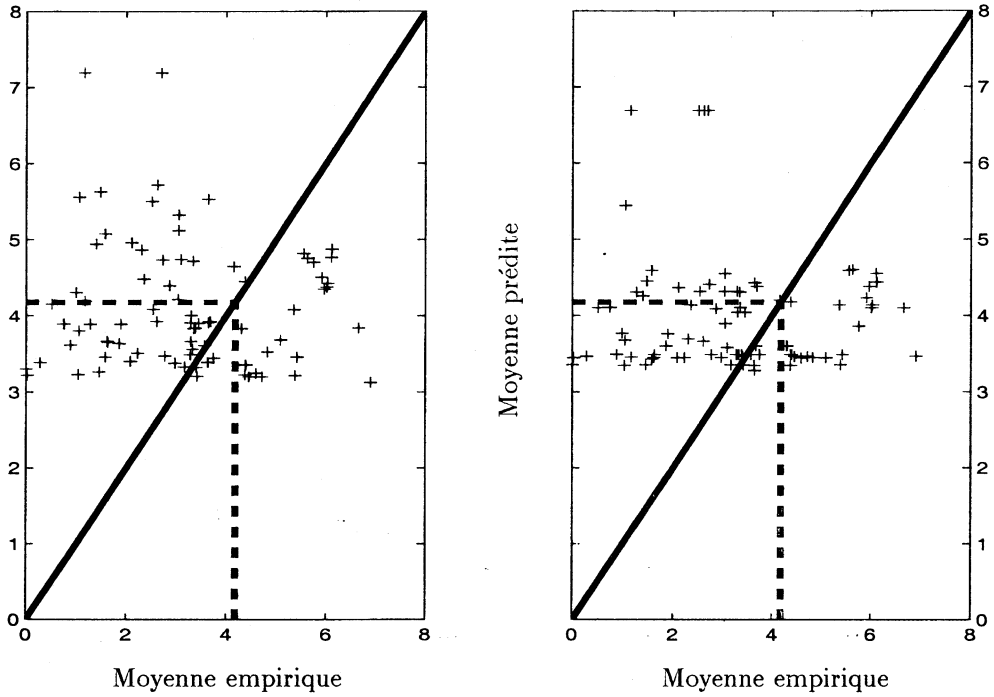


Figure 4.12 – Aperçu des résultats d'une unique procédure **Bagging** et d'une **Boosting** sur les dix chacune réalisées pour l'étude des observations Cigale restreintes à l'après-midi. En abscisse, les moyennes empiriques sur chaque cellule réservée à la procédure de validation ; en ordonnée, les moyennes prédites à chacune par la procédure **Bagging** (gauche) et la procédure **Boosting** (droite). La diagonale représente une portion de la droite de pente 1. Les traits discontinus indiquent la valeur de la moyenne des quantités de trafic sur toutes les observations de l'ensemble d'apprentissage. Les réductions des **écarts norme 2 de validation pour la moyenne** sont présentées dans le Tableau 4.7. On peut voir l'évolution des contrastes empiriques de validation moyennés sur les dix procédures dans la Figure 4.11. Les importances $\Delta_{\delta}^{\mathrm{ba}}$ ($\delta = 1,\ 0.9,\ 0.8$) des variables socio-démographiques et culturelles évaluées par moyennage des importances agrégées à chaque procédure Bagging sont exposées dans le Tableau C.9 de l'Annexe C.

Bagging	Boosting
11%±6	8%±4
(moyenne)	

Tableau 4.7 – Réductions des **écarts norme 2 de validation pour la moyenne** entre le modèle à moyenne unique et les modèles dus aux procédures Bagging et Boosting pour les observations Cigale restreintes à l'après-midi. On les évalue par moyenne sur dix procédures chacune. Le premier nombre indique le gain moyen en pourcents ; le second (précédé par le signe ±), l'écart-type correspondant.

4.5. Tranche horaire de la soirée

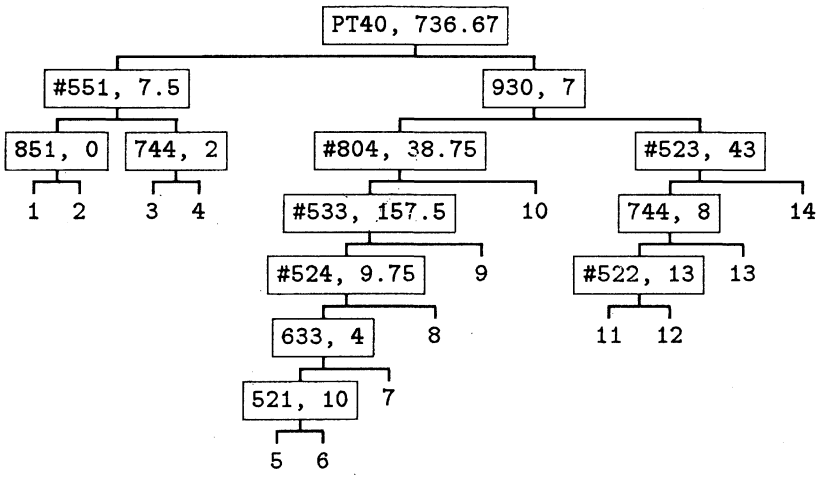
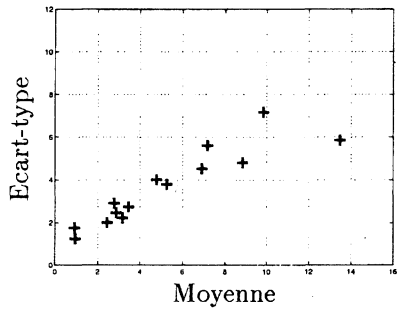


Figure 4.13 – Arbre de régression CART construit à partir des observations Cigale restreintes à la soirée et les données INSEE associées. On indique à chaque nœud le code de l'indicateur INSEE de coupure et le seuil correspondant. On compte 14 feuilles. Les couples de moyenne et écart-type pour chaque feuille sont présentés dans la Figure 4.14. Les importances des variables calculées sur cet arbre sont exposées dans le Tableau C.11 de l'AnnexeC.



f	m_f	σ_f	%	f	m_f	σ_f	%
1	2.7	2.9	9	8	2.8	2.5	3
2	2.4	2.0	18	9	8.8	4.8	2
3	3.1	2.2	7	10	0.9	1.2	1
4	5.3	3.8	3	11	6.9	4.5	7
5	4.8	4.0	38	12	9.8	7.2	3
6	7.2	5.6	3	13	13.5	5.8	2
7	3.4	2.7	3	14	0.9	1.8	1

Figure 4.14 – Les couples de moyenne et écart-type pour les 14 feuilles de l'arbre de régression CART construit à partir des observations Cigale restreintes à la soirée et les données INSEE associées. On peut voir l'arbre dans la Figure 4.13. A gauche, les points du plan associés. A droite, les couples de moyenne m_f et écart-type σ_f pour chaque feuille f, ainsi que les proportions respectives de représentants par feuille.

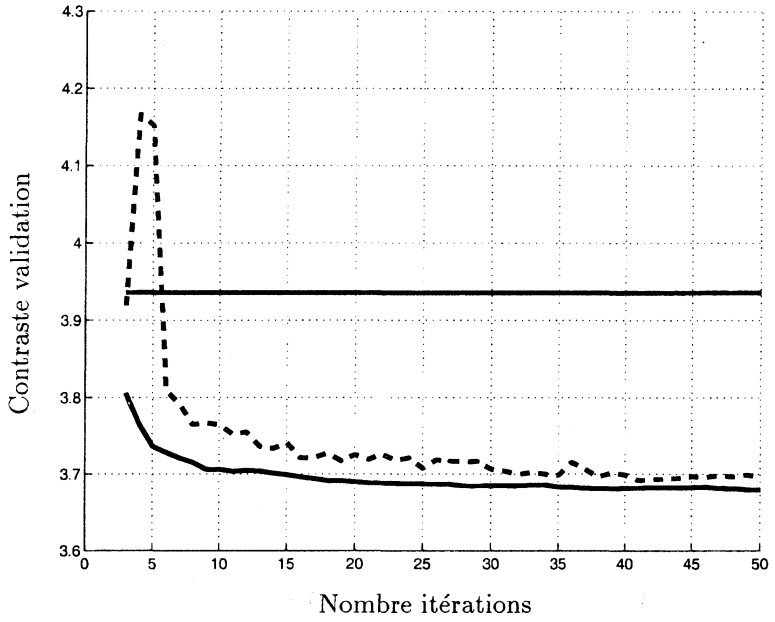


Figure 4.15 – Contraste empirique de validation moyenné sur dix procédures CART (barre horizontale), Bagging (tracé continu) et Boosting (tracé discontinu) pour un jeu d'observations de données Cigale restreintes à la soirée. Le nombre d'itérations est en abscisse. On omet les contrastes empiriques pour les deux premières itérations par souci de clarté de la figure. Les améliorations moyennes apportées par les deux procédures vis-à-vis d'un seul arbre de régression CART sont données dans les Tableaux 4.8 et 4.9.

Bagging	Boosting
10%±14	9%±14

(contraste)

Tableau 4.8 – Réductions du **contraste de validation** d'un seul arbre de régression CART dues aux procédures Bagging et Boosting pour les observations Cigale restreintes à la soirée. On les évalue par moyenne sur dix procédures chacune. Le premier nombre indique le gain moyen en pourcents ; le second (précédé par le signe ±), l'écart-type correspondant.

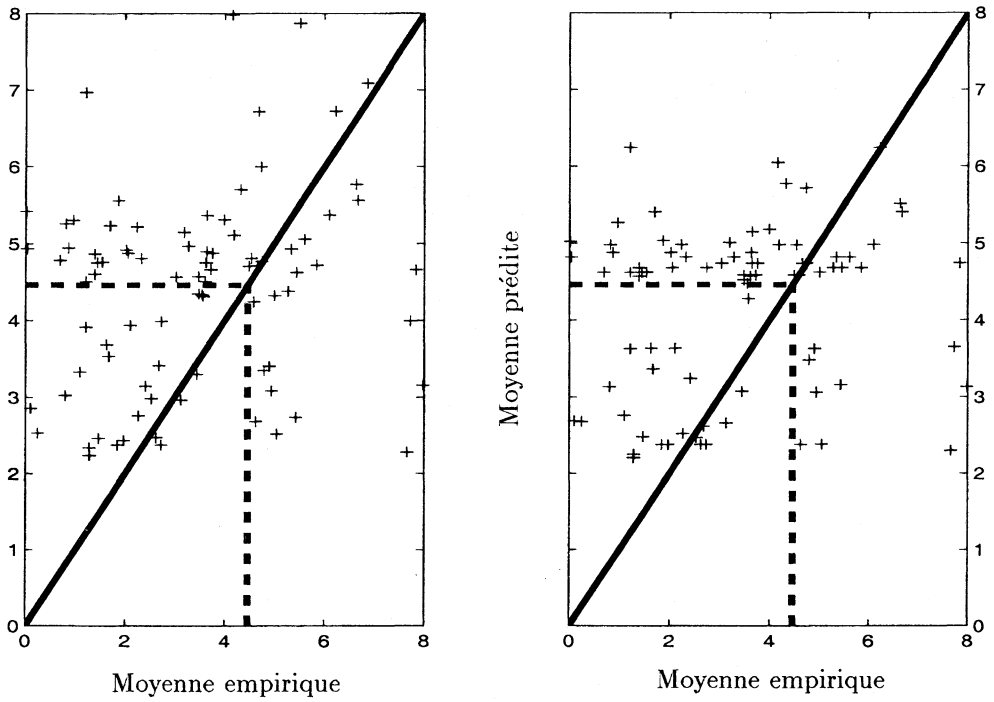


Figure 4.16 – Aperçu des résultats d'une unique procédure **Bagging** et d'une **Boosting** sur les dix chacune réalisées pour l'étude des observations Cigale restreintes à la soirée. En abscisse, les moyennes empiriques sur chaque cellule réservée à la procédure de validation ; en ordonnée, les moyennes prédites à chacune par la procédure **Bagging** (gauche) et la procédure **Boosting** (droite). La diagonale représente une portion de la droite de pente 1. Les traits discontinus indiquent la valeur de la moyenne des quantités de trafic sur toutes les observations de l'ensemble d'apprentissage. Les réductions des **écarts norme 2 de validation pour la moyenne** sont présentées dans le Tableau 4.9. On peut voir l'évolution des contrastes empiriques de validation moyennés sur les dix procédures dans la Figure 4.15. Les importances $\Delta_{\delta}^{\mathrm{ba}}$ ($\delta = 1,\ 0.9,\ 0.8$) des variables socio-démographiques et culturelles évaluées par moyennage des importances agrégées à chaque procédure Bagging sont exposées dans le Tableau C.12 de l'Annexe C.

Bagging	Boosting
22%±6	21%±7
(moyenne)	

Tableau 4.9 – Réductions des **écarts norme 2 de validation pour la moyenne** entre le modèle à moyenne unique et les modèles dus aux procédures Bagging et Boosting pour les observations Cigale restreintes à la soirée. On les évalue par moyenne sur dix procédures chacune. Le premier nombre indique le gain moyen en pourcents ; le second (précédé par le signe ±), l'écart-type correspondant.

4.6. Journée complète

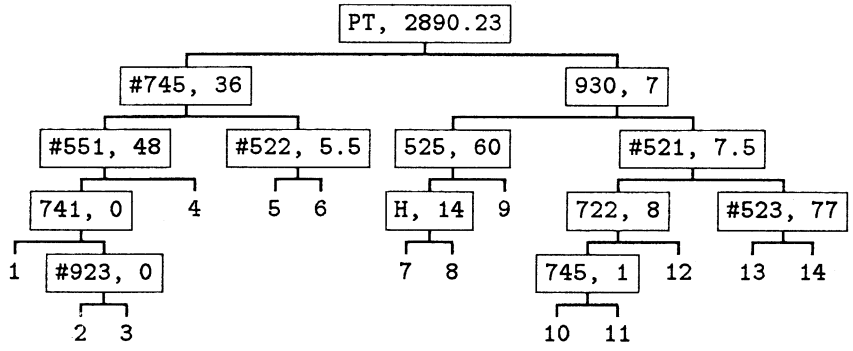
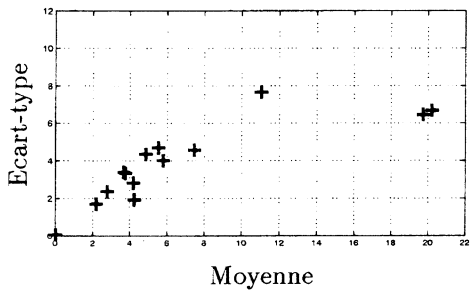


Figure 4.17 – Arbre de régression CART construit à partir des observations Cigale pour la journée complète et les données INSEE associées. On indique à chaque nœud le code de l'indicateur INSEE de coupure et le seuil correspondant. On compte 14 feuilles. Les couples de moyenne et écart-type pour chaque feuille sont présentés dans la Figure 4.21. Les importances des variables calculées sur cet arbre sont exposées dans le Tableau C.14 de l'Annexe C.



f	m_f	σ_f	%	f	m_f	σ_f	%
1	3.7	3.4	8	8	4.9	4.3	39
2	2.8	2.4	3	9	20.2	6.7	0
3	4.2	2.8	6	10	7.4	4.6	2
4	5.8	4.0	2	11	2.2	1.7	1
5	4.2	1.9	1	12	11.0	7.7	3
6	19.7	6.4	0	13	5.5	4.7	9
7	3.8	3.3	25	14	0.0	0.1	0

Figure 4.18 – Les couples de moyenne et écart-type pour les 14 feuilles de l'arbre de régression CART construit à partir des observations Cigale pour la journée complète et les données INSEE associées. On peut voir l'arbre dans la Figure 4.17. A gauche, les points du plan associés. A droite, les couples de moyenne m_f et écart-type σ_f pour chaque feuille f, ainsi que les proportions respectives de représentants par feuille.

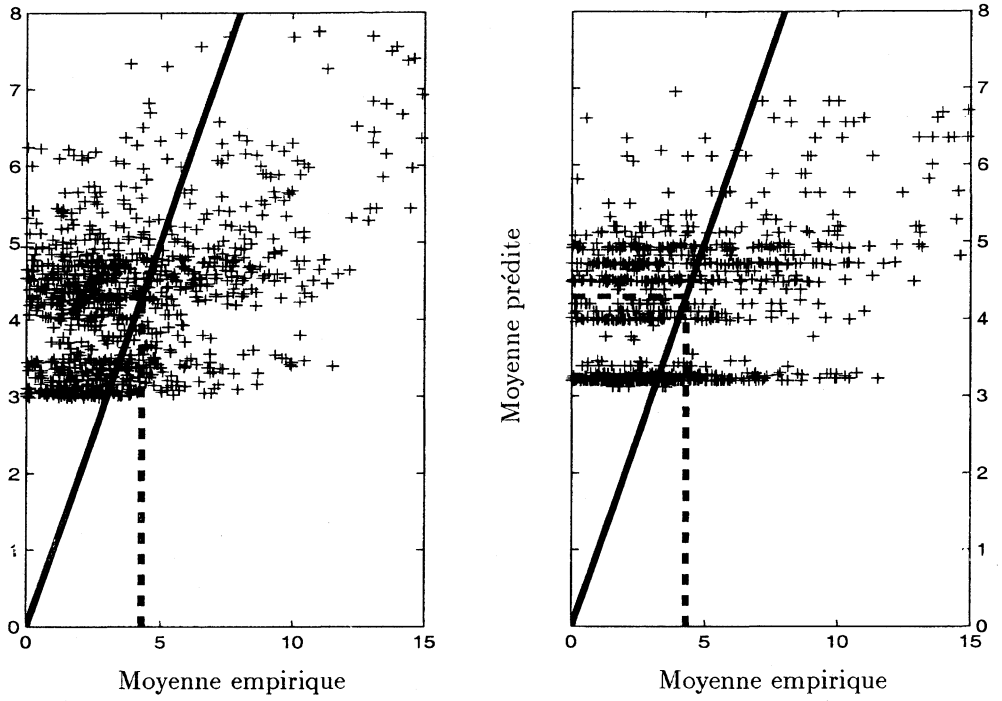


Figure 4.19 – Aperçu des résultats des uniques procédure **Bagging** et **Boosting** réalisées pour l'étude des observations Cigale à la journée. En abscisse, les moyennes empiriques sur chaque cellule réservée à la procédure de validation ; en ordonnée, les moyennes prédites à chacune par la procédure **Bagging** (gauche) et la procédure **Boosting** (droite). La diagonale représente une portion de la droite de pente 1. Les traits discontinus indiquent la valeur de la moyenne des quantités de trafic sur toutes les observations de l'ensemble d'apprentissage. Les réductions des **écarts norme 2 de validation pour la moyenne** sont présentées dans le Tableau 4.10. Les importances $\Delta_\delta^{\mathrm{ba}}$ ($\delta = 1,\ 0.9,\ 0.8$) des variables socio-démographiques et culturelles évaluées lors de l'unique procédure Bagging sont exposées dans le Tableau C.15 de l'Annexe C.

Bagging	Boosting
23%±?	18%±?
(moyenne)	

Tableau 4.10 – Réductions des **écarts norme 2 de validation pour la moyenne** entre le modèle à moyenne unique et les modèles dus aux procédures Bagging et Boosting pour les observations Cigale restreintes à la soirée. On ne peut donner que l'amélioration due à chaque unique procédure. Aussi, le premier nombre indique le gain en pourcents ; l'absence d'écart-type est signalée par le point d'interrogation.

4.7. Données HC2

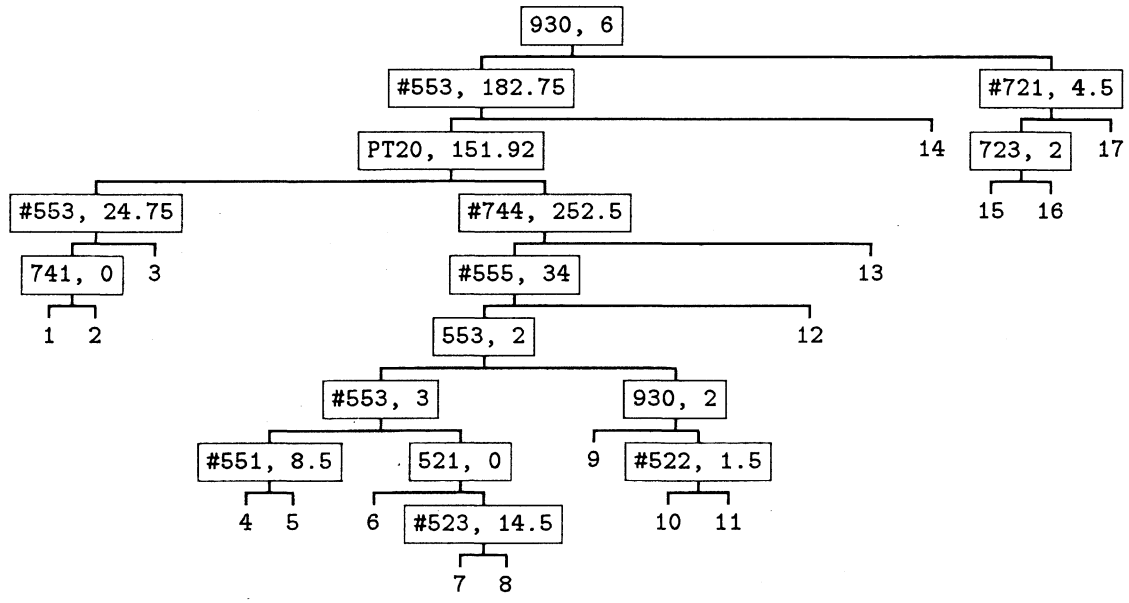
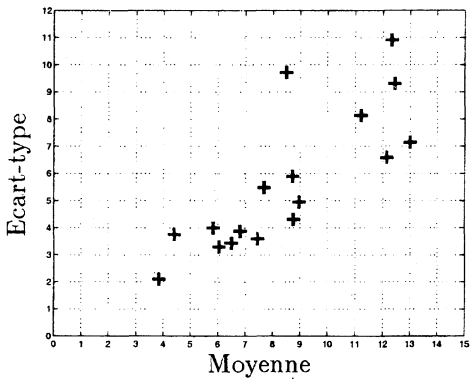


Figure 4.20 – Arbre de régression CART construit à partir des observations de trafic HC2 pour 1614 cellules parisiennes et les données INSEE associées. On indique à chaque nœud le code de l'indicateur INSEE de coupure et le seuil correspondant. On compte 17 feuilles. Les couples de moyenne et écart-type pour chaque feuille sont présentés dans la Figure 4.21. Les importances des variables calculées sur cet arbre sont exposées dans le Tableau C.17 de l'Annexe C.



f	m_f	σ_f	%	f	m_f	σ_f	%
1	4.4	3.8	3	10	7.4	3.6	15
2	3.8	2.1	7	11	7.7	5.5	5
3	6.5	3.4	1	12	8.7	5.9	8
4	5.8	4.0	17	13	12.3	10.9	1
5	9.0	5.0	2	14	12.4	9.3	3
6	8.5	9.7	3	15	8.8	4.3	7
7	6.8	3.9	6	16	13.0	7.1	2
8	11.2	8.1	1	17	12.1	6.6	0
9	6.0	3.3	18				

Figure 4.21 – Les couples de moyenne et écart-type pour les 17 feuilles de l'arbre de régression CART construit à partir des observations de trafic HC2 pour 1614 cellules parisiennes et les données INSEE associées. On peut voir l'arbre dans la Figure 4.20. A gauche, les points du plan associés. A droite, les couples de moyenne m_f et écart-type σ_f pour chaque feuille f.

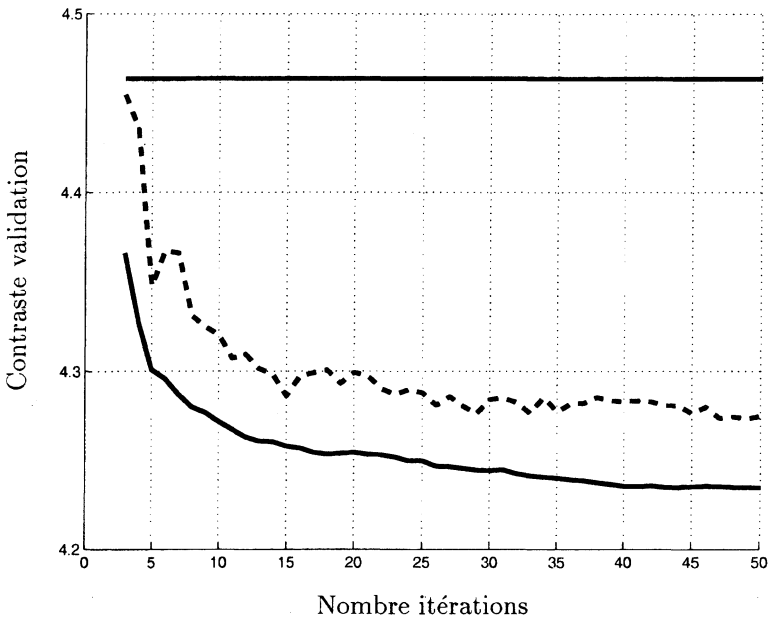


Figure 4.22 – Contraste empirique de validation moyenné sur dix procédures CART (barre horizontale), Bagging (tracé continu) et Boosting (tracé discontinu) pour le jeu d'observations HC2. Le nombre d'itérations est en abscisse. On omet les contrastes empiriques pour les deux premières itérations par souci de clarté de la figure. Les améliorations moyennes apportées par les deux procédures vis-à-vis d'un seul arbre de régression CART sont données dans les Tableaux 4.11 et 4.12.

Bagging	Boosting
6%±5	5%±6

(contraste)

Tableau 4.11 – Réductions du **contraste de validation** d'un seul arbre de régression CART dues aux procédures Bagging et Boosting pour les observations de trafic HC2. On les évalue par moyenne sur dix procédures chacune. Le premier nombre indique le gain moyen en pourcents ; le second (précédé par le signe ±), l'écart-type correspondant.

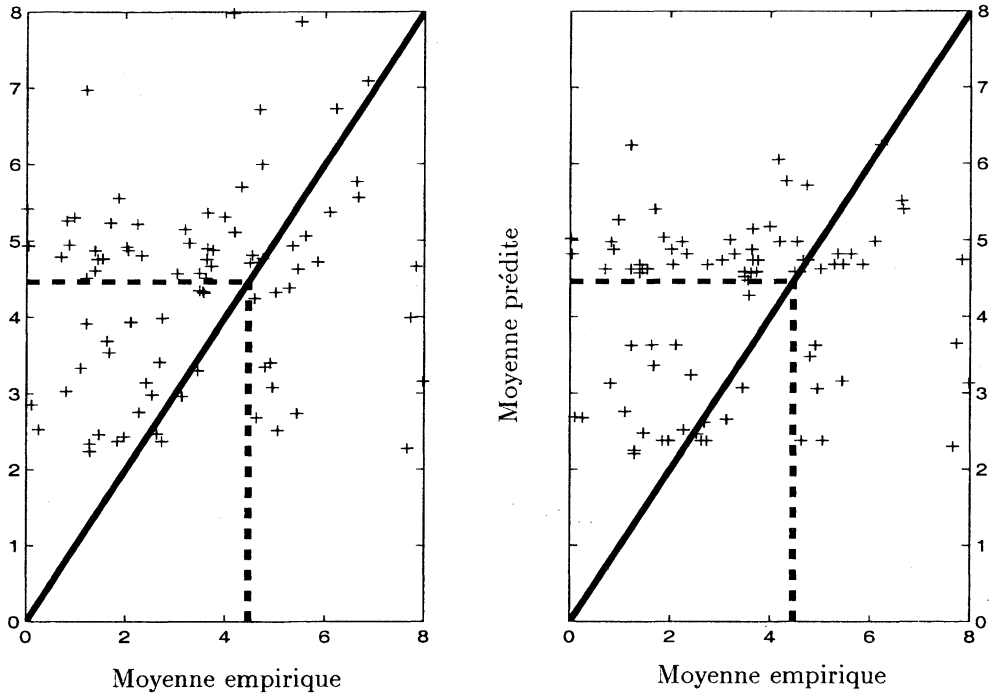


Figure 4.23 – Aperçu des résultats d'une unique procédure **Bagging** et d'une **Boosting** sur les dix chacune réalisées pour l'étude des observations de trafic HC2. En abscisse, les moyennes empiriques sur chaque cellule réservée à la procédure de validation ; en ordonnée, les moyennes prédites à chacune par la procédure **Bagging** (gauche) et la procédure **Boosting** (droite). La diagonale représente une portion de la droite de pente 1. Les traits discontinus indiquent la valeur de la moyenne des quantités de trafic sur toutes les observations de l'ensemble d'apprentissage. Les réductions des **écarts norme 2 de validation pour la moyenne** sont présentées dans le Tableau 4.12. On peut voir l'évolution des contrastes empiriques de validation moyennés sur les dix procédures dans la Figure 4.22. Les importances $\Delta_{\delta}^{\mathrm{ba}}$ ($\delta = 1,\ 0.9,\ 0.8$) des variables socio-démographiques et culturelles évaluées par moyennage des importances agrégées à chaque procédure Bagging sont exposées dans le Tableau C.18 de l'Annexe C.

Bagging	Boosting
17%±6	16%±5
(moyenne)	

Tableau 4.12 – Réductions des **écarts norme 2 de validation pour la moyenne** entre le modèle à moyenne unique et les modèles dus aux procédures Bagging et Boosting pour les observations de trafic HC2. On les évalue par moyenne sur dix procédures chacune. Le premier nombre indique le gain moyen en pourcents ; le second (précédé par le signe ±), l'écart-type correspondant.

4.8. Synthèse

4.8.1. Performances des procédures Bagging et Boosting*

On note que les deux procédures améliorent systématiquement les performances d'un arbre seul pour le contraste de validation moyenné, ce qui est la moindre des choses. On note aussi que la procédure Bagging est toujours supérieure à la procédure Boosting au sens toujours du contraste de validation moyenné.

Nous pouvons témoigner que les performances pour le contraste de validation des procédures Bagging et Boosting dépendent de façon importante du choix des échantillons d'apprentissage et de validation. Il arrive parfois que les procédures Bagging et Boosting détériorent à hauteur de un pourcent les performances d'un arbre seul. A l'opposé, il arrive que le gain en performance soit de l'ordre de 50 %. Les valeurs assez élevées des écarts-types des améliorations sont à cet égard révélatrices.

Les performances des procédures Bagging et Boosting dépendent aussi sensiblement des valeurs choisies pour le paramètre seuil λ présenté dans la Section 3.3.3 du Chapitre 3. Rappelons que ce paramètre régit le choix d'un bon arbre comme sortie de chaque apprentissage élémentaire CART au cours des itérations Bagging et Boosting. Une procédure Monte Carlo sur l'algorithme Bagging a *suggéré* un comportement satisfaisant selon le choix du seuil λ : heuristiquement (nous prenons des précautions car nous n'avons pu faire assez d'itérations pour assurer ce qui suit) des valeurs élevées ou au contraire faibles pour le seuil (qui se traduisent respectivement par le choix d'arbres trop grands ou trop petits et donc par des phénomènes de sur– ou sous-ajustement) entraînent une détérioration de l'amélioration moyenne, alors que des valeurs modérées semblent stabiliser l'amélioration moyenne. Quant au calibrage du seuil, il est malaisé. Nous préconiserons sans doute plus tard une autre méthode de sélection d'un bon arbre dans une forêt CART.

Enfin, les graphiques où l'on représente les couples de moyenne empirique contre moyenne estimée sur l'ensemble de validation laissent penser qu'un léger phénomène de sur-estimation de la moyenne se manifeste. Le calcul des réductions des écarts norme 2 de validation pour la moyenne rassurent quant à l'apport des procédures Bagging et Boosting relativement au modèle naïf pour lequel les trafics partagent une même moyenne (estimée par la moyenne de toutes les observations de trafic sur l'ensemble d'apprentissage). Pour ce critère, le gain est d'autant plus important que le trafic est inhomogène, *i.e.* que les améliorations sont sensiblement meilleures sur la journée complète et pour les données HC2.

4.8.2. Importance des variables INSEE

Les différentes mesures de l'importance

Nous allons tâcher ici de résumer l'ensemble des tableaux rassemblés en Appendice C.

Rappelons pour commencer que nous avons appliqué notre méthode à six jeux différents d'observations de trafic, qui correspondent d'une part à quatre tranches horaires de deux heures dites de matinée (neuf à onze heures), mi-journée (onze à treize heures), après-midi (seize à dix-huit heures) et soirée (vingt à vingt-deux heures) ; d'autre part à la journée dite complète (en fait restreinte de midi à vingt-deux heures par souci d'économie de temps de calcul) et au jeu de données HC2 dans son intégralité.

*La lenteur de nos algorithmes a rendu difficile une étude approfondie de propriétés fines des procédures Bagging et Boosting. Nous souhaitons y remédier prochainement.

Pour chacune de ces applications, nous avons calculé

- les corrélations sur une échelle de -100 à 100 pour tous les indicateurs INSEE vis-à-vis de la moyenne et de l'écart-type, calculées sur l'ensemble des données ;
- les importances Δ_1, $\Delta_{0.9}$ et $\Delta_{0.8}$ des variables INSEE échelonnées entre 0 et 100 (seuls les rapports des importances sont significatifs) comme résultats d'une unique procédure CART (voir la Section 3.3.4 du Chapitre 3) ;
- les mêmes importances Δ_1^{ba}, $\Delta_{0.9}^{\text{ba}}$ et $\Delta_{0.8}^{\text{ba}}$ pour tous les indicateurs INSEE échelonnées entre 0 et 100, résultants d'une procédure d'agrégation Bagging (voir la Section 3.4.2 du Chapitre 3).

Ces différentes quantités sont de nature différente bien qu'elles contribuent toutes trois à l'évaluation de l'importance des variables INSEE, *i.e.* à la quantification pour chaque variable socio-démographique et culturelle de son caractère explicatif du trafic. Ainsi

- les corrélations sont de nature *globale*. Il est possible que les corrélations calculées à partir de certaines cellules uniquement soient très différentes des corrélations calculées sur le jeu total. En particulier, une corrélation faible ne révèle pas nécessairement une absence d'influence, mais peut par exemple être, heuristiquement, le fruit de deux corrélations de même ordre et de sens opposés sur deux morceaux d'une partition du jeu de données en deux parties distinctes.
- les importances Δ_δ et $\Delta_\delta^{\text{ba}}$ sont au contraire de nature *locale* du fait même de leur définition.* Il faut noter que le paramètre $0 < \delta \leq 1$ a par ailleurs pour fonction de tempérer cette nature locale en pénalisant l'importance évaluée sur de petits groupes de données. Cet effet est annulé pour la valeur $\delta = 1$. C'est pour cette raison que nous avons choisi de calculer les importances agrégées $\Delta_\delta^{\text{ba}}$ pour les trois valeurs de δ égales à 1, 0.9 et 0.8, afin de mettre en lumière le degré de localité des importances.

A titre d'exemples, on peut dégager parmi beaucoup d'autres le cas de l'importance de la variable **#522** pour les Tableaux C.1 et C.3 : sa corrélation pour la moyenne ou l'écart-type est nulle alors qu'au contraire, ses importances Δ_δ(**#522**) pour les trois valeurs de δ sont si élevées qu'elles en font la sixième variable la plus importante pour les trois valeurs de δ. Il convient cependant de souligner que les importances Δ_δ(**#522**) décroissent avec δ, ce qui laisse penser que le caractère local de l'indicateur **#522** n'est pas trop prononcé. Le cas des variables **521** et **#521** est différent : leur corrélation est relativement importante (en comparaison de l'ensemble des valeurs prises par tous les indicateurs) et leurs importances les classent respectivement première et troisième variable la plus importante au sens de $\Delta_\delta^{\text{ba}}$ pour les trois valeurs de δ. Finalement, l'exemple de l'indicateur **#925** vient compléter la collection : cette fois, les corrélations sont élevées alors qu'au contraire, les importances sont nulles. On peut penser dans ce cas que la variable **#925** n'est pas significative globalement pour son sens propre, mais comme révélatrice, à cette échelle seulement, d'autres effets.

Les importances CART

Les tableaux des importances calculées sur d'uniques procédures CART ont surtout pour fonction de mettre en lumière l'instabilité de Δ_δ, *i.e.* sa sensibilité aux données d'entrée (voyez

*L'importance $\Delta_\delta(k)$ de la k-ième covariable est la somme, sur tous les nœuds t de l'arbre CART pour lequel elle est calculée, du meilleur gain en contraste qu'une question $q(t,x) = \mathbb{1}\{x^{(k)} > C\}$ induirait. Si le nœud t est à la profondeur depht(t), alors ce gain est multiplié par un facteur $\delta^{\text{depth}(t)}$. L'importance $\Delta_\delta^{\text{ba}}(k)$ est la moyenne des $\Delta_\delta(k)$ sur tous les arbres CART construits lors d'une procédure Bagging.

la Section 3.3.4 du Chapitre 3). C'est le constat auquel on arrive grâce à la comparaison avec les importances agrégées $\Delta_\delta^{\mathrm{ba}}$ dont on pense en revanche qu'elles sont stabilisées, grâce au principe moteur de la procédure d'agrégation Bagging.

Cette remarque nous fournit un prétexte pour de nouveau insister sur les précautions que l'on doit prendre lorsque l'on exploite un unique arbre de régression CART. Un tel arbre seul est certes convivial (en particulier aisément interprétable), mais par trop instable (et donc peu fiable) pour être satisfaisant.

Les importances agrégées

Nous finissons par un commentaire sur les tableaux des importances agrégées $\Delta_\delta^{\mathrm{ba}}$. Ceux-ci constituent indéniablement un résultat crucial de notre méthode. Il est en effet apparu au cours des discussions que nous avons tenues avec nos correspondants de France Télécom (voir la Section 1.5.5) qu'ils souhaitaient qu'un nombre restreint d'indicateurs socio-démographiques et culturels soient dégagés, qui permettent d'expliquer à eux seuls assez correctement le trafic téléphonique. Il nous semble que les Tableaux C.3, C.6, C.9, C.12, C.15 et C.18 contribuent de façon décisive à l'élaboration d'une réponse. Aussi, nous nous proposons de les résumer dans l'intention de mettre en valeur des indicateurs les plus importants en un sens à préciser.

En premier lieu, nous insistons sur une particularité évidente du Tableau C.15 : la variable `H` indicatrice de la tranche horaire dont nous avons enrichi les données socio-démographiques et culturelles apparaît très largement prépondérante en termes d'importance agrégée (la seconde variable la plus importante est limitée à 25% d'importance). Ceci est une conséquence directe de la non-stationnarité du trafic à l'échelle d'une journée que nous avons déjà commentée dans le Chapitre 1, Section 1.2.4.

Il est satisfaisant de constater que la procédure d'estimation a bien été sensible à ce fait et qu'elle a su s'adapter. Ainsi, de façon imagée, la procédure vue comme boîte noire (selon les termes de la Section 3.4.1 du Chapitre 3) s'est scindée en plusieurs sous-boîtes dont les apprentissages se sont appuyés sur des sous-ensembles de l'échantillon complet de construction plus homogènes en termes de tranche horaire que ce dernier.

On peut aussi noter que l'indicateur `H` apparaît dans l'arbre de régression présenté dans la Figure 4.17, ce qui n'est pas automatique en dépit de l'importance prépondérante de l'indicateur. Ceci suggère d'ailleurs que lors d'ultérieures applications de la méthode élaborée dans ces pages à des données HC2, une variable mensuelle soit ajoutée aux variables INSEE.

Voici en second lieu une tentative de synthèse des tableaux cités plus haut. Nous passons en revue toutes les importances agrégées $\Delta_\delta^{\mathrm{ba}}$ pour tous les jeux de données et les paramètres δ. Nous ne retenons que, mettons, les 18 meilleurs indicateurs pour chaque configuration. Ce sont 29 indicateurs seulement (sur les 89 possibles et en incluant l'indicateur `H` ajouté pour les données Cigale à la journée) qui sont ainsi sélectionnés, pour un total de 21 si l'on omet la différenciation entre l'indicateur générique `CCC` et son indicateur compagnon `#CCC`.

Ce résultat nous paraît assez surprenant, eu égard à la variété des jeux de données impliqués, qui sont constitués pour une part de données de quantité de trafic écoulé sur dix minutes au cours d'une journée, d'autre part de quantités maximales de trafic écoulé en une heure sur une semaine. Nous donnons le détail de cette synthèse dans le Tableau 4.13.

code	matinée	mi-journée	après-midi	soirée	journée	HC2
H	0	0	0	0	100	0
PT	39	26	24	21	12	28
PT60	12	6	8	9	2	17
PT75	14	9	10	9	3	4
158	85	100	95	100	26	100
#158	63	69	79	39	12	84
521	100	57	51	41	12	61
#521	73	71	79	55	14	52
522	22	34	41	27	5	42
#522	48	88	100	26	18	43
523	54	77	72	39	13	48
#523	27	16	24	47	9	27
524	41	23	19	21	2	25
#524	26	25	23	27	5	26
525	15	38	32	32	4	29
#525	9	13	13	11	1	14
527	8	17	32	13	2	17
551	17	59	50	10	5	43
#551	10	12	14	17	5	28
553	30	23	26	20	6	33
#553	13	9	8	7	2	21
#554	10	14	15	5	2	11
633	7	14	14	13	2	19
651	5	9	8	13	5	8
721	29	8	8	9	3	15
722	7	15	27	5	8	10
741	35	19	17	27	5	14
851	12	5	4	6	5	5
930	13	5	7	30	4	16

Tableau 4.13 – Une synthèse de l'importance des variables explicatives. Dans la colonne de gauche, les codes des 29 *meilleurs indicateurs* au sens où chacun d'eux est au moins 18ième meilleur indicateur pour l'ordre décroissant d'importance agrégée pour un au moins des jeux de données (matinée, mi-journée, après-midi, soirée, journée complète, HC2) et un paramètre δ (parmi les valeurs 1, 0.9 et 0.8). Chacune des colonnes de droite correspond à un jeu de données particulier. On précise ligne par ligne les importances, échelonnées entre 0 et 100 sur les colonnes, pour les meilleurs indicateurs. L'importance de l'indicateur k est la moyenne des importances non renormalisées $\Delta_\delta^{\mathrm{ba}}(k)$ pour les trois valeurs de δ.

4.9. English summary

4.9.1. Outline

We have finally applied our method to various datasets extracted from the original Cigale and HC2 databases we introduced earlier in Chapter 1. The latter are, in order of appearence, the quantities of traffic (in Erlang) dealt with by all BTS in Paris through successive time intervals of 10 minutes between

- nine and eleven o'clock in the morning;
- eleven in the morning and one in the afternoon;
- four and six o'clock in the afternoon;
- eight and ten o'clock in the evening;
- noon and ten o'clock in the evening.

A sixth dataset is finally given as input of the method, namely the HC2 database.

We must add that we have *a priori* filtered the total covariables dataset presented in Interlude 2. Thus, we have removed obviously strongly correlated variables (*e.g.* keeping only the total population by ages numbers and deleting the latter numbers restricted to females or males). We have also reduced the number of distinct kinds of establishments' activities, merging some APET700 activities into pseudo-APET ones. This was done consistently with the APET60 classification. Besides, a new predictor is added when considering the traffic quantities between noon and ten in the evening: it indicates the hour of observation. There are finally 89 predictive variables, with 6 total population by ages numbers, 1 total number of housings, 41 numbers of establishments coded along the pseudo-APET activities (their format is `Digit Digit Digit`) and their 41 associated numbers of approximate cumulated staff (their format is `#Digit Digit Digit`). One also has to count the time of observation predictor when coping with the Cigale dataset between noon and ten in the evening.

Those applications aim at casting some light on our final procedure, focusing on

- a description of its outputs;
- some validation measurements;
- the extraction of important predictive variables for the regression model, and finally on
- the illustration of its *flexibility*.

We present some illustrative figures and tables in Section 4.2 to Section 4.7. The six sections follow the same lines, in order to ease comparisons between the various results. Thus, we shall provide for each case:

- An example of CART regression tree built on the learning set $\mathcal{L}_n$, together with a plot of the mean and standard error assigned to each leaf, and a table in which we give, for each leaf, the associated couple of mean and standard error and the proportion of BTS that belong to it.

- A graph where the averaged (on ten Bagging and Boosting procedures of 50 iterations on the basis of bootstrap uniform n-resamples $\mathcal{L}'_n$ of the learning set $\mathcal{L}_n$) validation contrasts are plotted *versus* the number of iterations. Besides, an horizontal line indicates the averaged validation contrasts achieved by single CART regression trees built on each $\mathcal{L}'_n$.
- The percentages of enhancement in terms of validation contrast yielded by Bagging or Boosting with respect to (wrt) single trees, and the corresponding standard error. See Equation (4.1) for an accurate definition.
- The percentages of enhancement in terms of ℓ^2-norm for the mean yielded by Bagging or Boosting, see Equation (4.2) for a precise definition.
- Finally, two graphes side by side where the empirical means of the traffic quantities are plotted *versus* the predicted means by Bagging (left) and Boosting (right) for each element of the validation set, and for one only among the ten aggregation procedures.

4.9.2. Some results

We emphasize that the aggregation procedures always improve (in average) the performances of a single tree. Moreover, Bagging performs better than Boosting (yet in average). This relies heavily on the choice of learning and validation sets. It may happen sometimes that aggregation yields poorer results than those given by a single tree, though seldom. On the contrary, it may happen that aggregating yields 50% enhancements. The large values of the standard errors for enhancements in averaged validation contrasts correspond to this feature.

We also observed that the choice of the threshold parameter λ involved in the selection of a good tree for each CART procedure plays an important role in terms of global performances of the aggregation algorithms. Further simulation studies should allow to cast some light on the corresponding behaviour (but we will have to implement a more efficient code than our for Matlab).

Finally, Table 4.13 presents 29 best predictors on the basis of the aggregated variables importance calculated through the whole Bagging procedures.

5

Detecting abrupt changes in random fields*

Résumé

Nous étudions dans ce chapitre certaines propriétés asymptotiques d'un M-estimateur dans un cadre de détection de ruptures dans la loi d'un champ aléatoire. Cette classe de problèmes comprend notamment la reconnaissance de formes. Nous avons recours à diverses techniques, parmi lesquelles on trouvera des arguments classiques de M-estimation, des inégalités de concentration, des inégalités maximales pour des variables aléatoires dépendantes et du ϕ-mélange. Lorsque la complexité du vrai modèle (*i.e.* son ordre au sens du Chapitre 7) n'est pas connue, nous recourons à des techniques de pénalisation du contraste. L'ensemble des résultats est valable sous des hypothèse raisonnables que nous commentons. Des exemples élémentaires viennent illustrer notre propos.

Abstract

This chapter is devoted to the study of some asymptotic properties of a M-estimator in a framework of detection of abrupt changes in random field's distribution. This class of problems includes *e.g.* recovery of sets. It involves various techniques, including M-estimation method, concentration inequalities, maximal inequalities for dependent random variables and ϕ-mixing. Penalization of the criterion function when the size of the true model (*i.e.* its order according to the terminology of Chapter 7) is unknown is performed. All the results apply under mild, discussed assumptions. Simple examples are provided.

*Paru sous forme d'article dans le sixième volume de la revue ESAIM P&S en 2002.
I would like to thank warmly Jérôme Dedecker for his clear, helpful introduction to mixing.

Au menu

5.1. Introduction **167**
5.2. The partitions and the associated parameters **170**
5.2.1. Introducing partitions and associated parameters 170
5.2.2. Pseudo-distances for partitions and parameters 171
5.3. Modelization, observations, contrast **173**
5.3.1. Observations and first assumptions 174
5.3.2. Further assumptions: on the contrast 175
5.4. Controlling random fluctuations *via* maximal inequalities **176**
5.5. The case of known cardinality of the true partition **178**
5.5.1. Definition of the estimator 178
5.5.2. Consistency 178
5.5.3. Rate of convergence 179
5.5.4. Number of misclassified observations 182
5.6. The case of unknown cardinality of the true partition **183**
5.6.1. Definition of the estimator 183
5.6.2. Consistency 183
5.7. Appendix **184**
5.7.1. Proof of Proposition 5.3.3 184
5.7.2. Exploring Assumption **A6** 186
5.7.3. Proof of Lemma 5.5.3 189

5.1. Introduction

The problem of abrupt changes detecting includes a wide range of subjects unified by a common basic framework: observation of a random process whose distribution is long-scale heterogeneous but short-scale homogeneous on some regions. Comprehensive presentations can be be found in the three books (Basseville and Nikiforov 1993; Brodsky and Darkhovsky 1993; Carlstein, Müller, and Siegmund 1994).

The mathematical methods include M-estimation as in the present chapter or (Lavielle and Ludeña 2000; Lavielle and Moulines 2000) and also nonparametric or Bayesian techniques, see *e.g.* (Antoniadis, Gijbels, and MacGibbon 2000; Lavielle and Lebarbier 2001).

Handling estimation or test in a multiple changes case with an unknown number of changes is crucial and intricate. Akaike's and Schwarz's papers (1974, 1978) are most of the time invoked as milestones, as Yao's (1988), who proved consistency of Schwarz's criterion based estimator in case of independent Gaussian observations. Penalization methods are widely used, for instance in context of the estimation of the order of a process (see Akaike 1974), of the order of a mixture (see Dacunha-Castelle and Gassiat 1997), or generally in a context of statistical learning theory (see for instance the lectures notes Lugosi 2000). Barron, Birgé, and Massart obtained in 1999 some precise bounds in a framework of regression and density estimation. Penalization in view of estimating a number of change-points is widely used, for instance among the previous citations in (Lavielle and Ludeña 2000; Lavielle and Moulines 2000).

Examples

One of the simplest models for change-points can be summarized by the following model: one observes responses $Y(X_i)$ at $X_i = i$ with $Y(X_i) = \vartheta^\star(X_i) + \varepsilon(X_i)$ for centered possibly dependent $\varepsilon(X_i)$ and some piecewise constant function $\vartheta^\star$. Here, $X_1, \ldots, X_n$ should be understood as *regular times of observation.* A first natural extension consists of observing at random points X_i on a *d-dimensional lattice.* Then to allow observation throughout some *general d-dimensional space* $\mathcal{X}$. And finally to observe some process Y indexed by $x \in \mathcal{X}$ at *randomly chosen points* X_i of $\mathcal{X}$.

Recovery of sets obviously enters in this framework, too: one observes an image $\mathcal{X}$ composed of an object $\tau_0^\star$ and a background through noisy observations (X_i, Y_i) $(i = 1, \ldots, n)$, with independent $X_i \in \mathcal{X}$ and responses $Y_i = f(X_i)\mathbb{1}\{X_i \in \tau_0^\star\} + \xi_i$, for some function f bounded away from 0 from below and random centered noise ξ_i. Here, $(X_1, \ldots, X_n)$ are supposed independent of the mutually independent n-tuple $(\xi_1, \ldots, \xi_n)$. The aim is to estimate $\tau_0^\star$, or equivalently the partition $\tau^\star = (\tau_0^\star, \mathcal{X} - \tau_0^\star)$.

Aim of this chapter

We address in this chapter the estimation of a partition $\tau^\star$ of $\mathcal{X}$ from possibly dependent random observations Y_i at independent and identically P-distributed points X_i in $\mathcal{X}$. The proofs are based on Lavielle's paper (1999). The model actually consists of a couple $(\tau^\star, \theta^\star)$: $\tau^\star = (\tau_j^\star)_{1 \le j \le K^\star}$ is a partition with $K^\star$ subsets, where $K^\star$ (the *cardinality* of $\tau^\star$) is possibly unknown, and $\theta^\star$ is a collection of $K^\star$ finite-dimensional parameters $\theta_j^\star$. We define for convenience $\vartheta^\star = \sum_{j=1}^{K^\star} \theta_j^\star \mathbb{1}\{\tau_j^\star\}$. We consider that changes affect the marginal distribution of Y_i's: conditionally on X_i, Y_i has a distribution which depends on $\vartheta^\star(X_i)$. We assume that there exists an *ad hoc* contrast J_n associated with the problem.

Indeed, we estimate $\vartheta^\star$ by minimum contrast estimation and related techniques. Suppose first that we choose *a priori* the cardinality K of the estimator. By definition, the estimator $\hat{\vartheta}_n = (\hat{\tau}_n, \hat{\theta}_n)$ bounds from below the contrast $J_n(\tau, \theta)$ computed at any model (τ, θ) of cardinality K.

Involved techniques

J_n is naturally decomposed into the sum of a first term that depends only on $X_1, \ldots, X_n$ and a term of random centered fluctuations. Fluctuations take the form $\Sigma_n(G) = \sum_{i=1}^n Z_i \mathbb{1}\{X_i \in G\}$ for $Z_i = Y_i - \mathrm{E}(Y_i | X_i)$ and any G in a set $\mathcal{G}$. Section 5.4 is devoted to the control of those fluctuations *via* maximal inequalities.

A maximal inequality consists of an upper bound of the probability for $\sup\{\|\Sigma_n(G)\|_\infty : G \in \mathcal{G}\}$ to be greater than some $\delta > 0$. In the simple case where partitions are constructed with elementary rectangles, one can derive easily such maximal inequalities from mild control of the second order moment of the fluctuations (see Móricz, Serfling, and Stout 1982; Móricz 1983). This problem is more difficult in a general framework where partitions are constructed with elements of a larger class of sets (see Dedecker 2001). In comparison with the previous simple case, control of moments of order any $p > 2$ is needed here.

Denote $\mathbb{P}_n$ the empirical measure of $(X_1, \ldots, X_n)$. Another theoretical complication arises from the need to derive lower bounds for $(P(G) - \mathbb{P}_n(G))/P(G)$ from bounds of $P(G)$ for a large class of sets G. Actually, this is possible with large probability for sets G satisfying $P(G) \ge r_n$

for some sequence $\{r_n\} \downarrow 0$ carefully chosen. We deal with this difficulty thanks to concentration inequalities (refer to Massart 2000; Talagrand 1996a), see Section 5.3.1.

Results for a priori known cardinality $K^\star$. Penalization.

We finally obtain under mild assumptions and for *a priori* known $K^\star$ that estimation is asymptotically consistent and we bound from below rates of convergences. Quite surprisingly, but according to Lavielle's former results, the rate of convergence of the estimate $\hat{\tau}_n$ of $\tau^\star$ does not seem to depend on the dependence structure of Y_i's. It is strongly related to the rate $\{r_n\}$ mentioned in the previous subsection.

One can generalize those results for known $K^\star$. We can indeed construct an estimator $\hat{\vartheta}_{n,K}$ of $\vartheta^\star$ for any *a priori* choice of the cardinality K of the estimator. The point is then to select the best estimator among them. This is roughly speaking the aim of the penalization method: replace the contrast $J_n(\tau, \theta)$ by its penalized version $J_n(\tau, \theta) + \beta_n K$, with $\beta_n > 0$. The added term $\beta_n K$ penalizes the models with large cardinality whereas those models are favoured when minimizing $J_n(\tau, \theta)$ alone.

We prove that, for sequences $\{\beta_n\} \downarrow 0$ slowly enough, penalized estimation yields a consistent estimated triplet $(\hat{K}_n, \hat{\tau}_n, \hat{\theta}_n)$. Naturally, dependence structure of Y_i's affects the maximum rate of convergence for $\{\beta_n\}$.

Comparison with previous works

We noticed earlier that the field of recovery of sets is part of the general problem of abrupt changes detection. Thus, we may wish to compare our results to classical ones in that field. Choose Mammen and Tsybakov's (1995) paper where the authors derive some optimal convergence rates. Recall the previous crude description of the recovery of sets problem. Here, the partition to estimate has cardinality 2, so the penalization procedure is not needed. The point is to estimate $\tau_0^\star$. Roughly speaking, the authors prove that the risk for the maximum likelihood estimator (which is also a M-estimator) achieves the best possible rate of convergence in the minimax approach. Nevertheless, those results rely on independence of responses Y_i. On the contrary, our results apply in a framework of M-estimation of abrupt changes from dependent observations and are satisfying in this context, see again the former citations.

Asymptotics

This chapter is concerned with asymptotic results. In the whole text, the expression "as $n, \delta \uparrow \infty$" will correspond to limits $\lim_{\delta \to \infty} \lim_{n \to \infty}$, and *idem* for "as $n, \eta \uparrow \infty$".

The practical interest of detecting abrupt changes in the general setting described above is certain, though our asymptotic results are mainly of theoretical value. They ensure confidence in a reasonable idealistic framework and encourage to find practical recipes to apply. Indeed, rigorous minimum contrast estimation is here computationally intractable and the penalization coefficient β_n would have to take a fixed value for real observed data. The choice of such a value would be justified by practical considerations as presented *e.g.* in (Bai, Rao, and Wu 1999; Lavielle). An automatic choice would require non asymptotic theory, see for example (Barron, Birgé, and Massart 1999), but this is beyond the scope of this chapter.

Notation

In the whole chapter, different positive constants might be denoted by the same letter C.

The organization of the chapter is as follows: we introduce in Section 5.2 the partitions and the associated parameters to be studied and we define a pseudo-distance between them with useful properties. Section 5.3 is dedicated to the description of both the observations and the contrast to be minimized. Further assumptions are presented in Section 5.4. They deal with some crucial maximal inequalities. We consider estimation for known cardinality $K^\star$ of $\tau^\star$ in Section 5.5 and use those results to address the unknown cardinality case in Section 5.6. The Appendix 5.7 consists of three parts: the first one devoted to the postponed proof of a proposition; the second one to an exploration of the assumptions presented in Section 5.4; the third one to a sketch of proof of a technical lemma.

5.2. The partitions and the associated parameters

5.2.1. Introducing partitions and associated parameters

Set a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ upon which random variables will be defined.

Consider some probability space $(\mathcal{X}, \mathcal{G}, P)$ where P has support $\mathcal{X}$, *i.e.* $\{x \in \mathcal{X} : x \in \mathcal{O} \Longrightarrow P(\mathcal{O}) > 0\} = \mathcal{X}$ ($\mathcal{O}$ denotes an open set). $\mathcal{X}$ is typically included in $\mathbb{R}^d$. We will define partitions of $\mathcal{X}$ in the next paragraphs.

First, choose some set $\mathcal{F}_0 \subset \mathcal{G}$ of measurable sets. Roughly speaking, a partition τ of $\mathcal{X}$ will be constructed as a collection (τ_k) satisfying $\cup_k \tau_k = \mathcal{X}$ and where any τ_k is a finite union of elements of $\mathcal{F}_0$

Then, define $\mathcal{F}$ which contains all finite unions of elements of $\mathcal{F}_0$ and pairwise intersection of such sets. Moreover, we suppose for sake of simplicity (that is to overcome measurability difficulties) that all the mathematical expressions in this chapter involving suprema over subsets of $\mathcal{F}$ are measurable (it suffices that for each of them, suprema are P-almost surely equal to suprema over some countable subsets).

Examples of $\mathcal{F}_0$ when $\mathcal{X} \subset \mathbb{R}^d$ include the set of all rectangles of the form $\Pi_{i=1}^d (a_i, b_i]$ (simply called *rectangles* in the whole chapter); the set of all the polygons whose edges have lengths bounded below by some positive constant (*polygons* for short); or more generally (including rectangles and polygons), some Vapnik-Červonenkis class whose Vapnik-Červonenkis dimension is finite (for references, see *e.g.* van der Vaart 1998; van der Vaart and Wellner 1996; Vapnik 1998). In the sequel, *VC* will stand for Vapnik-Červonenkis.

Other assumptions will concern $\mathcal{F}_0$ and $\mathcal{F}$: we will state them in Section 5.3.

Definition 5.2.1. We will consider $\mathcal{F}$-partitions (or shortly *partitions*) of $\mathcal{X}$. The set of all the partitions is denoted $\mathcal{T}$. Any $\tau \in \mathcal{T}$, $\tau = (\tau_k)_{1 \le k \le K}$, is a collection of subsets of $\mathcal{X}$. K is called *cardinality* of τ, also denoted $\mathrm{card}(\tau)$. Any τ_k can be written as an union $\cup_l \tau_k(l)$ of non intersecting elements $\tau_k(l)$ of $\mathcal{F}_0$ whose P-probabilities must be bounded from below by some fixed $\Delta^\star > 0$.

$\mathcal{T}_K$ denotes the set of partitions with cardinality K.

Remark 5.2.2. Condition of minimal P-probability for the pieces $\tau_k(l)$ of $\tau_k = \cup_l \tau_k(l)$ stands for technical reasons. We will actually suppose that we know some lower bound of $\Delta^\star$. Besides,

this condition yields that there exists a finite maximal partition cardinality $K_{\max}$ and that any τ_k is a finite union of $\tau_k(l)$.

Some parameters are to be associated with a partition in the following way: a partition $\tau \in \mathcal{T}$ with cardinality K may go with a collection θ of K Θ-valued vectors. Here, Θ is an open and precompact subset of $\mathbb{R}^p$. Thus, for $\tau = (\tau_k)_{1 \leq k \leq K}$ and $\theta = (\theta_k)_{1 \leq k \leq K}$, the parameter θ_k goes with τ_k. We will denote $\Theta_K = \Theta^K$.

5.2.2. Pseudo-distances for partitions and parameters

To start with, let us recall some notations. For two sets A and B, $A \nabla B$ denotes their asymmetrical difference and $A \,\Delta\, B$ their symmetrical difference, that is

$$A \nabla B = A \setminus A \cap B \quad \text{and} \quad A \,\Delta\, B = (A \nabla B) \cup (B \nabla A).$$

We wish to define a pseudo-distance* between two $\mathcal{F}$-partitions of the set $\mathcal{X}$ that generalizes the natural definition in the usual one-dimensional case, see (Lavielle 1999): for t and $t^\star$ two increasing vectors (respectively of length K and $K^\star$), the pseudo-distance is taken to be $\max_{1 \leq j \leq K^\star} \min_{1 \leq k \leq K} |t_k - t_j^\star|$. Thus, that pseudo-distance is the largest distance between points of $t^\star$ and their respective closest point in t. Observe that it is zero if and only if each point of $t^\star$ appears in t. These considerations lead to the following:

Definition 5.2.3. Let τ and $\tau^\star$ be two $\mathcal{F}$-partitions of the set $\mathcal{X}$. Denote K and $K^\star$ their respective cardinality. The gap $g(\tau, \tau^\star)$ between them is defined as follows

$$g(\tau, \tau^\star) = \max_{1 \leq j \leq K^\star} \min_{\mathcal{K}} P\left(\left(\bigcup_{k \in \mathcal{K}} \tau_k \right) \Delta\, \tau_j^\star \right).$$

The index $\mathcal{K}$ in the infimum ranges over all subsets of $\{1, \ldots, K\}$.

For $j = 1, \ldots, K^\star$, we denote $\mathcal{K}_j$ a smallest subset of $\{1, \ldots, K\}$ achieving the minimum in the definition for fixed j. Consequently, we have

$$g(\tau, \tau^\star) = \max_{1 \leq j \leq K^\star} P\left(\left(\bigcup_{k \in \mathcal{K}_j} \tau_k \right) \Delta\, \tau_j^\star \right).$$

Let us present a few interesting properties of the gap g.

Proposition 5.2.4. *Consider two $\mathcal{F}$-partitions $\tau^\star = (\tau_j^\star)_{1 \leq j \leq K^\star}$ and $\tau = (\tau_k)_{1 \leq k \leq K}$.*

(i) *Let j be in $\{1, \ldots, K^\star\}$ and k in $\{1, \ldots, K\}$. Observe that if $\tau_k \subset \tau_j^\star$, then $k \in \mathcal{K}_j$ whereas $\tau_k \cap \tau_j^\star = \emptyset$ implies $k \notin \mathcal{K}_j$. Not surprisingly, if $k \notin \mathcal{K}_j$, then $P(\tau_k \nabla \tau_j^\star) \geq P(\tau_k \cap \tau_j^\star)$. On the contrary, if $k \in \mathcal{K}_j$ and $\operatorname{card}(\mathcal{K}_j) > 1$, then $P(\tau_k \nabla \tau_j^\star) \leq P(\tau_k \cap \tau_j^\star)$. When $\mathcal{K}_j = \{k\}$, the former inequality holds as soon as $g(\tau, \tau^\star) \leq \Delta^\star / 2$.*

*We did not know then the notion of Caccioppoli partitions and the associated distance that makes of the set of all Caccioppoli partitions whose perimeters are uniformly bounded by some $\zeta > 0$ and composed of at most K pieces with nonzero measure, a compact metric space. For a brief introduction, see the Appendix B. Some future work here ?

(ii) *Set $j_0 \neq j_1$ and $k_0 \in \mathcal{K}_{j_0}$. Then $P(\tau_{k_0} \cap \tau^\star_{j_1}) \leq g(\tau, \tau^\star)$.*

(iii) *If $g(\tau, \tau^\star) = 0$, then for all j, there exists $\mathcal{K}_j$ such that $\tau^\star_j = \bigcup_{k \in \mathcal{K}_j} \tau_k$ (equalities hold up to P-null sets, as the following conclusions). We derive from this that $\mathcal{K}_j$'s are mutually disjoint and $K \geq K^\star$: τ is a sub-partition of $\tau^\star$. Thus, when $g(\tau, \tau^\star) = 0$ with $K = K^\star$, we do have $\tau = \tau^\star$.*

Suppose now that $g(\tau, \tau^\star) < \Delta^\star/2$. We still have mutually disjoint $\mathcal{K}_j$'s, and therefore again, $K \geq K^\star$. In particular, when $K = K^\star$, one can assume that $\mathcal{K}_j = \{j\}$ for each j. Observe that $g(\tau, \tau^\star) \geq \Delta^\star$ as soon as $K < K^\star$.

(iv) *To conclude with, note that $\bigcup_{1 \leq j \leq K^\star} \mathcal{K}_j$ does not necessarily cover $\{1, \ldots, K\}$. Nevertheless, it does when $g(\tau, \tau^\star) < \Delta^\star / K_{\max}$.*

Proof. (i). Denoting $\widetilde{\mathcal{K}} = \mathcal{K} \cup \{k\}$ with $k \notin \mathcal{K}$, the following equalities hold true:

$$\begin{aligned} P\left(\tau^\star_j \,\Delta\, \left(\bigcup_{k \in \widetilde{\mathcal{K}}} \tau_k\right)\right) &= P\left(\tau^\star_j \nabla \left(\bigcup_{k \in \widetilde{\mathcal{K}}} \tau_k\right)\right) + \sum_{k \in \widetilde{\mathcal{K}}} P\left(\tau_k \nabla \tau^\star_j\right) \\ &= P\left(\tau^\star_j \nabla \left(\bigcup_{k \in \mathcal{K}} \tau_k\right)\right) - P\left(\tau_k \cap \tau^\star_j\right) + \sum_{k \in \mathcal{K}} P\left(\tau_k \nabla \tau^\star_j\right) + P\left(\tau_k \nabla \tau^\star_j\right) \\ &= P\left(\tau^\star_j \,\Delta\, \left(\bigcup_{k \in \mathcal{K}} \tau_k\right)\right) + P\left(\tau_k \nabla \tau^\star_j\right) - P\left(\tau_k \cap \tau^\star_j\right). \end{aligned}$$

We conclude taking on the one hand $\mathcal{K} = \mathcal{K}_j$ and $\mathcal{K} = \mathcal{K}_j - \{j\}$ on the other hand. If $\mathcal{K}_j = \{k\}$ and $g(\tau, \tau^\star) \leq \Delta^\star/2$, then use $P\left(\tau_k \nabla \tau^\star_j\right) \leq \Delta^\star/2$ and $\Delta^\star \leq P(\tau_k) = P\left(\tau_k \nabla \tau^\star_j\right) + P\left(\tau_k \cap \tau^\star_j\right)$.

(ii). We have indeed

$$P\left(\tau_{k_0} \cap \tau^\star_{j_1}\right) \leq P\left(\tau_{k_0} \nabla \tau^\star_{j_0}\right) \leq P\left(\left(\bigcup_{k \in \mathcal{K}_{j_0}} \tau_k\right) \nabla \tau^\star_{j_0}\right) \leq P\left(\left(\bigcup_{k \in \mathcal{K}_{j_0}} \tau_k\right) \Delta\, \tau^\star_{j_0}\right) \leq g(\tau, \tau^\star).$$

(iii). Let $g(\tau, \tau^\star) < \Delta^\star/2$. Suppose $k \in \mathcal{K}_{j_0}$ and $k \in \mathcal{K}_{j_1}$ with $j_0 \neq j_1$, too. Then

$$P(\tau_k) = P\left(\left(\tau_k \nabla \tau^\star_{j_0}\right) \cup \left(\tau_k \nabla \tau^\star_{j_1}\right)\right) \leq P\left(\tau_k \nabla \tau^\star_{j_0}\right) + P\left(\tau_k \nabla \tau^\star_{j_1}\right) < \Delta^\star,$$

which is excluded.

(iv). To see that, suppose we can take $k_0 \notin \bigcup_{1 \leq j \leq K^\star} \mathcal{K}_j$:

$$P(\tau_{k_0}) = \sum_{j=1} K^\star P\left(\tau_{k_0} \cap \tau^\star_j\right) \leq \sum_{j=1} K^\star P\left(\tau^\star_j \nabla \left(\bigcup_{k \in \mathcal{K}_j} \tau_k\right)\right) < K^\star \cdot \Delta^\star / K_{\max},$$

and that is impossible. □

We will use two pseudo-distances between two parameters θ and $\theta^\star$ respectively associated with τ and $\tau^\star$ (as explained in 5.2.1) that are compatible with the definition of the gap $g(\tau, \tau^\star)$. Set some parameter $\theta^\star = (\theta^\star_j)_{1 \leq j \leq K^\star}$ in $\overline{\Theta}_{K^\star}$ having no equal coordinates. Let v be a nonnegative continuous function on the set $\{\theta^\star_1, \ldots, \theta^\star_{K^\star}\} \times \overline{\Theta}$, continuously differentiable with respect to its second variable, whose derivative has continuous extension on $\{\theta^\star_1, \ldots, \theta^\star_{K^\star}\} \times \overline{\Theta}$. Furthermore, v must verify the condition $v(\theta^\star_j, \theta) = 0$ $(1 \leq j \leq K^\star,\ \theta \in \overline{\Theta})$ if and only if $\theta = \theta^\star_j$.

Definition 5.2.5. Let τ and $\tau^\star$ be two $\mathcal{F}$-partitions of the set $\mathcal{X}$, K and $K^\star$ their respective cardinality. Let θ and $\theta^\star$ be two parameters, respectively taken in Θ_K and $\Theta_{K^\star}$, $\theta^\star$ having no equal coordinates. We define the two following pseudo-distances between them

$$\begin{aligned} d_2(\theta,\theta^\star) &= \max_{1\le j\le K^\star}\max_{k\in\mathcal{K}_j}\|\theta_j^\star-\theta_k\|_2, \\ d_v(\theta,\theta^\star) &= \max_{1\le j\le K^\star}\max_{k\in\mathcal{K}_j} v(\theta_j^\star,\theta_k). \end{aligned}$$

Here, indexes $\mathcal{K}_j$ arise from definition of $g(\tau,\tau^\star)$.

Remark 5.2.6.

- Naturally, thanks to properties of v, $d_v \le Cd_2$ holds (recall that in the whole chapter, different positive constants might be denoted C).
- If $d_v(\theta,\theta^\star)=0$, then $\mathcal{K}_j$'s do not intersect each other and $K\ge K^\star$. Moreover, if $g(\tau,\tau^\star)=d_v(\theta,\theta^\star)=0$ and $K=K^\star$, then $\tau=\tau^\star$ and $\theta=\theta^\star$.

5.3. Modelization, observations, contrast

Modelization

We assume that the set $\mathcal{X}$ is covered by a $\mathcal{F}$-partition $\tau^\star$ consisting of $K^\star$ subsets $\tau_j^\star$ ($j=1,\dots,K^\star$). In the whole chapter, index j will generally be devoted to description of objects related to $(\tau^\star,\theta^\star)$. Index k will correspond to other (τ,θ). A parameter $\theta^\star$ with no equal coordinates of $\Theta_{K^\star}$ is associated with $\tau^\star$ and $\vartheta^\star$ is defined by $\vartheta^\star=\sum_j \theta_j^\star \mathbb{1}\{\tau_j^\star\}$. We define respectively $\mathcal{T}_{K,\delta}$ and $\Theta_{K,\delta}$ as the sets of all partitions τ of cardinality K such that $g(\tau,\tau^\star)>\delta$ and all parameters θ of length K such that $d_2(\theta,\theta^\star)>\delta$.

We suppose the existence of a random field indexed by $x\in\mathcal{X}$ of possibly dependent random variables (rv): for any $x\in\mathcal{X}$, a rv Y_x taking its values in $\mathbb{R}^q$ is generated according to a distribution which depends on $\vartheta(x)$.

Our aim is to estimate $K^\star$, $\tau^\star$ and $\theta^\star$ from random observations under as mild conditions as possible.

Two classical examples

Detection in the mean: Here, $Y_x=\vartheta^\star(x)+Y'_x$ for some strictly stationary field of centered rv $(Y'_x)_{x\in\mathcal{X}}$. Consequently, the vector of true parameters $\theta^\star$ is understood as the vector of the true possible means. Thus, Y_x has mean $\theta_j^\star$ if and only if $x\in\tau_j^\star$.

Detection in both mean and variance: Denoting $\theta^\star=(\mu^\star,s^{2\star})$ and $\vartheta_1^\star=\sum_j \mu_j^\star\mathbb{1}\{\tau_j^\star\}$, $\vartheta_2^\star=\sum_j s_j^{2\star}\mathbb{1}\{\tau_j^\star\}$, we define $Y_x=\vartheta_1^\star+\vartheta_2^{\star 1/2}Y'_x$. Here, $(Y'_x)_{x\in\mathcal{X}}$ is a strictly stationary field of centered rv with variance 1. In this example, Y_x has mean $\mu_j^\star$ and variance $s_j^{2\star}$ if and only if $x\in\tau_j^\star$.

5.3.1. Observations and first assumptions

One observes n rv (X_i, Y_i) $(i = 1, \ldots, n)$. X_1^n denotes the vector of mutually independent variables $(X_1, \ldots, X_n)$ generated independently of $(Y_x)_{x \in \mathcal{X}}$. X_i takes its values in $\mathcal{X}$ and $Y_i := Y_{X_i}$ takes its in $\mathbb{R}^q$. P is the common distribution of $X_1, \ldots, X_n$. $\mathbb{P}_n$ is the empirical distribution.

Remark 5.3.1. All the proofs still hold up to minor changes when $X_1, \ldots, X_n$ are not identically distributed *but still independent* and P denotes the arithmetic mean of their respective distribution P_{X_i}. Furthermore, we emphasize that the rv $Y_1, \ldots, Y_n$ are possibly *dependent.*

Consider now the first three assumptions: they concern $\mathcal{F}$ and P.

A1 The random variable $\sup\{|\mathbb{P}_n(F) - P(F)| : F \in \mathcal{F}\}$ converges P-almost surely (P-as) to zero. In other words, $\mathcal{F}$ is P-Glivenko-Cantelli.

A2 There exists a sequence $\{r_n\} \downarrow 0$ such that $\liminf_n n r_n > 0$ and

$$\lim_{\eta \to \infty} \lim_{n \to \infty} \mathrm{P}\left(\sup \left\{ \frac{P(F) - \mathbb{P}_n(F)}{P(F)} \ : \ F \in \mathcal{F}, P(F) \geq \eta r_n \right\} \geq \frac{1}{2} \right) = 0.$$

Remark 5.3.2 (on Assumptions A1 and A2). Assumption **A1** is fulfilled whenever $\mathcal{F}$ is a finite VC-dimension VC class. In the sequel, the case of $\mathcal{F}$ finite VC-dimension VC class will be the more general example for $\mathcal{F}$. For a wide family of examples, see for instance (van der Vaart 1998). On the other hand, Proposition 5.3.3 below (whose proof, postponed in Appendix 5.7.1, require *independence* of X_i's) casts some light on the Assumption **A2.**

Proposition 5.3.3. *Assumption* **A2** *holds whenever $\mathcal{F}$ is a finite VC-dimension VC class and the sequence* $\left\{ \dfrac{\log r_n}{n r_n} \right\}$ *is bounded.*

Remark 5.3.4 (on Proposition 5.3.3**).** Choices of $r_n = (\log^\alpha n)^\beta / n$ with integer $\alpha \geq 1$ and positive β are obviously included (with notation $\log^{\alpha+1} = \log \circ \log^\alpha$, $\log^1 = \log$).

The last assumption of this Section concerns the control of the moment of order h of $\mathbb{P}_n(G)$ for $G \in \mathcal{G}$:

A3 For any $h \in (1, 2)$ and $G \in \mathcal{G}$, for some constant $A > 0$ depending on h only,

$$\mathrm{E}(\mathbb{P}_n(G)^h) \leq A \, (\mathrm{E}\mathbb{P}_n(G))^h = A \, P(G)^h.$$

Remark 5.3.5 (on Assumption A3). Note that Jensen's inequality yields straightforwardly to the reversed lower bound $\mathrm{P}(G)^h \leq \mathrm{E}(\mathbb{P}_n(G)^h)$. Assumption **A3** is always satisfied for independent, non necessarily identically distributed, rv: it is a simple consequence of Rosenthal's inequality, see *e.g.* (Petrov 1995). We will use this inequality to derive useful maximal inequalities in Section 5.4.

5.3.2. Further assumptions: on the contrast

The following assumption ensures the existence of a contrast J_n adapted to our model. $J_n(\tau, \theta)$ is obtained as a sum of local contrasts $W_n(\tau_k, \theta_k)$ computed at (τ_k, θ_k).

A4 Let $\varphi : \overline{\Theta} \to \mathbb{R}$ and $\psi : \overline{\Theta} \to \mathbb{R}^r$ be two continuously differentiable functions with continuous extensions of the derivatives on $\overline{\Theta}$. Let $\xi : \mathbb{R}^q \to \mathbb{R}^r$ be such that $\xi(Y_x) \in L^1(\mathrm{P})$ for any $x \in \mathcal{X}$ and $\xi(Y_X) \in L^1(\mathrm{P})$ for X P_{X_i}-distributed. Define the local contrasts for (τ_k, θ_k) $(k = 1, \dots, K)$ by

$$W_n(\tau_k, \theta_k) = n^{-1} \sum_{i=1}^{n} \Big\{ \varphi(\theta_k) + \psi(\theta_k)^T \xi(Y_i) \Big\} \mathbb{1}\{X_i \in \tau_k\}$$

and introduce the corresponding limit contrast $w : \{\theta_1^\star, \dots, \theta_{K^\star}^\star\} \times \overline{\Theta} \to \mathbb{R}$, which is supposed to satisfy:

- P-as for all i such that $X_i \in \tau_j^\star$ and any $\theta \in \overline{\Theta}$,

$$w(\theta_j^\star, \theta) = \varphi(\theta) + \psi(\theta)^T \mathrm{E}\left(\xi(Y_i) \mid X_i\right); \tag{5.1}$$

- $w(\theta_j^\star, \theta) \geq w(\theta_j^\star, \theta_j^\star)$ for any $(\theta_j^\star, \theta) \in \{\theta_1^\star, \dots, \theta_{K^\star}^\star\} \times \overline{\Theta}$, equality if and only if $\theta = \theta_j^\star$.

Denote v the centered limit contrast, that is $v(\theta_j^\star, \theta) = w(\theta_j^\star, \theta) - w(\theta_j^\star, \theta_j^\star)$, any $(\theta_j^\star, \theta)$. Then v is nonnegative, continuous on $\{\theta_1^\star, \dots, \theta_{K^\star}^\star\} \times \overline{\Theta}$, continuously differentiable on $\{\theta_1^\star, \dots, \theta_{K^\star}^\star\} \times \Theta$ with respect to its second variable. Its derivative has continuous extension on $\{\theta_1^\star, \dots, \theta_{K^\star}^\star\} \times \overline{\Theta}$. Finally, v is zero only on the diagonal. Thus, following Definition 5.2.5 in Section 5.2.2, we can define a pseudo-distance d_v from v. Furthermore, since $\{v(\theta_j^\star, \cdot),\ j = 1, \dots, K^\star\}$ are continuous, there exist $\rho^\star, v^\star > 0$ such that, for any $j_0 \neq j_1$,

$$\inf \Big\{ v(\theta_{j_0}^\star, \theta) \ : \ \|\theta - \theta_{j_1}^\star\|_2 \leq \rho^\star \Big\} - \sup \Big\{ v(\theta_{j_0}^\star, \theta) \ : \ \|\theta - \theta_{j_0}^\star\|_2 \leq \rho^\star \Big\} \geq v^\star.$$

Remark 5.3.6 (on Assumption A4).

- Condition (5.1) in the former assumption deals with the way the rv Y_i depends on X_i through $\tau^\star$. In particular, P-as for i, i' such that $X_i,\ X_{i'} \in \tau_j^\star$,

$$\mathrm{E}(\xi(Y_i) \mid X_i) - \mathrm{E}(\xi(Y_{i'}) \mid X_{i'}) \in \mathrm{Vect}(\psi(\Theta))^\perp$$

and they are equal as soon as $\mathrm{Vect}(\psi(\Theta)) = \mathbb{R}^r$, which is clearly the case for $r = 1$ and $\xi \neq 0$.

- Note that $\mathrm{E}\left(W_n(\tau_j^\star, \theta) \mid X_1^n\right) = \mathbb{P}_n(\tau_j^\star)\, w(\theta_j^\star, \theta)$, where $\mathbb{P}_n(\tau_j^\star)$ tends to $P(\tau_j^\star)$ P-as. Thus, $w(\theta_j^\star, \cdot)$ can be understood as a rescaled limit conditional expectation of the local contrast computed at $\tau_j^\star$.

The next assumption concerns v:

A5 There exist $B > 0$, $\sigma > 0$ such that (up to change of $\rho^\star$)

$$\text{if } \|\theta - \theta_j^\star\|_2 \leq \rho^\star, \text{ then } v(\theta_j^\star, \theta) \geq B \|\theta - \theta_j^\star\|_2^\sigma \ (j = 1, \dots, K^\star).$$

Back to the classical examples

Detection in the mean: We choose the following local criterion function

$$W_n(\tau_k, \theta_k) = n^{-1} \sum_{i=1}^{n} (Y_i - \theta_k)^2 \mathbb{1}\{X_i \in \tau_k\} - n^{-1} \sum_{i=1}^{n} Y_i^2 \mathbb{1}\{X_i \in \tau_k\}.$$

Here, $\varphi(\theta) = \theta^2$, $\psi(\theta) = -2\theta$ and $\xi(y) = y$. For this particular criterion, $v(\theta_j^\star, \theta) = (\theta - \theta_j^\star)^2$ and Assumption **A5** above is satisfied.

Detection in both mean and variance: This time, we choose

$$W_n(\tau_k, \theta_k) = n^{-1} \sum_{i=1}^{n} \Big\{ \frac{(Y_i - \mu_k)^2}{s_k^2} + \log s_k^2 \Big\} \mathbb{1}\{X_i \in \tau_k\}.$$

Here, $\varphi(\mu, s^2) = \mu^2/s^2 + \log s^2$, $\psi(\mu, s^2) = (-2\mu, 1)/s^2$ and $\xi(y) = (y, y^2)$. Moreover, we have

$$v(\theta_j^\star, \theta) = \frac{(\mu_j^\star - \mu)^2}{s^2} + \log \frac{s^2}{s_j^{\star 2}} + \frac{s_j^{\star 2}}{s^2} - 1.$$

Thus, $v(\theta_j^\star, \theta)$ is twice the Kullback-Leibler information $H(\mathcal{N}_{\theta_j^\star} \,|\, \mathcal{N}_\theta)$ for Gaussian rv $\mathcal{N}_{\theta_j^\star}$ (resp. $\mathcal{N}_\theta$) of mean and variance $\theta_j^\star$ (resp. θ). Besides, Assumption **A5** above does hold for $\Theta =]a, b[\times]c, d[$ with $d > c > 0$.

Note that in both cases, minimization of $\theta \mapsto W_n(\tau_j^\star, \theta)$ leads to the natural least squares estimators of the parameter $\theta_j^\star$.

5.4. Controlling random fluctuations *via* maximal inequalities

Let us define the centered random field of fluctuations $(Z_x)_{x \in \mathcal{X}}$ and the rv Z_i $(i = 1, \ldots, n)$ by

$$Z_x = \xi(Y_x) - \mathrm{E}(\xi(Y_x)) \quad \text{and} \quad Z_i = \xi(Y_i) - \mathrm{E}(\xi(Y_i) \,|\, X_i)$$

and for any $x_1^n \in \mathcal{X}^n$, define the corresponding sums over any set $G \in \mathcal{G}$

$$\Sigma_{x_1^n}(G) = \sum_{i=1}^{n} Z_{x_i} \mathbb{1}\{x_i \in G\} \quad \text{and} \quad \Sigma_{X_1^n}(G) = \sum_{i=1}^{n} Z_i \mathbb{1}\{X_i \in G\}.$$

Denote finally $S_n(G; \theta) = \psi(\theta)^T \Sigma_{X_1^n}(G)$ and, for any $\mathcal{F}$-partition τ, $n_{kj} = n\mathbb{P}_n(\tau_k \cap \tau_j^\star)$. Then

$$W_n(\tau_k, \theta_k) = n^{-1} \sum_{j=1}^{K^\star} \Big\{ n_{kj} w(\theta_j^\star, \theta_k) + S_n(\tau_k \cap \tau_j^\star; \theta_k) \Big\}.$$

We impose

A6 There exist $C_1 > 0$ and $h \in (1,2)$ such that, for any $\delta > 0$, $G \in \mathcal{G}$,

$$\mathrm{P}\left(\sup\left\{\|\Sigma_{X_1^n}(F)\|_\infty \,:\, F \in \mathcal{F}(G)\right\} \geq \delta \,\middle|\, X_1^n\right) \leq \frac{C_1}{\delta^2}\left(\sum_{i=1}^n \mathbb{1}\{X_i \in G\}\right)^h \text{ P-as.}$$

Here, $\mathcal{F}(G)$ denotes the set $\{F \cap G \,:\, F \in \mathcal{F}\}$.

Observe that Assumptions **A3** and **A6** yield (uncondition with respect to X_1^n and then bound above) the following maximal inequality

Lemma 5.4.1. *Under Assumptions* **A3** *and* **A6**, *there exists* $C_2 > 0$ *such that, for any* $\delta > 0$ *and* $G \in \mathcal{G}$,

$$\mathrm{P}\left(\sup\left\{\|\Sigma_{X_1^n}(F)\|_\infty \,:\, F \in \mathcal{F}(G)\right\} \geq \delta\right) \leq \frac{C_2\, n^h P(G)^h}{\delta^2}.$$

Remark 5.4.2 (on Assumption A6 and Lemma 5.4.1). The aim of Assumption **A6** is to ensure the result of Lemma 5.4.1. The linearity of S_n in $\Sigma_{X_1^n}$ is needed to derive uniform control of $S_n(G;\theta)$ in (G,θ) from uniform control of $\Sigma_{X_1^n}(G)$ in G. Besides, we can propose some mild alternative condition to Assumption **A6**, see Section 5.7.2.

Back to the classical examples

Applying Lemma 5.4.1 and Assumption **A1**, we get that

$$(n\mathbb{P}_n(\tau_j^\star))^{-1}\|\Sigma_{X_1^n}(\tau_j^\star)\|_\infty = o_{\mathrm{P}}(1).$$

The previous result yields that the least squares estimators obtained by minimization of the respective contrasts at $\tau_j^\star$ are consistent.

Finally, state the last assumption: it concerns $\mathcal{F}_0$.

A7 For some constant $\gamma > 0$ depending on $\mathcal{F}_0$ only, for any $G \in \mathcal{G}$ and $r > 0$, there exists $\widetilde{G} \in \mathcal{G}$ subset of G with $P(\widetilde{G}) \leq \gamma r$ such that

$$\left\{F \in \widetilde{\mathcal{F}_0}(G) \,:\, P(F) \leq r\right\} \subset \mathcal{F}(\widetilde{G}). \tag{5.2}$$

Here, $\widetilde{\mathcal{F}_0}(G)$ denotes the set $\{F \cap G \,:\, F \in \mathcal{F}_0,\, F \nabla G \neq \emptyset\}$.

Remark 5.4.3 (on Assumption A7). Simplest examples are again when $\mathcal{F}_0$ is composed of rectangles or polygons. Besides, this result still holds when $\widetilde{\mathcal{F}_0}(G)$ is replaced in (5.2) by the set $\widetilde{\mathcal{F}}(G) = \{F \cap G \,:\, F \nabla G \neq \emptyset\}$ where F is an union of at most $K_{\max}$ elements of $\mathcal{F}_0$.

We can now state a key-result that completes Lemma 5.4.1:

Lemma 5.4.4. *Under Assumptions* **A3**, **A6**, **A7**, *there exists* $C_3 > 0$ *such that, for any* $G \in \mathcal{G}$, *any* $v > 0$,

$$\mathrm{P}\left(\sup\left\{\frac{\|\Sigma_{X_1^n}(F)\|_\infty}{nP(F)} \,:\, F \in \widetilde{\mathcal{F}}(G), P(F) \geq v\right\} \geq \delta\right) \leq \frac{C_3(nv)^{h-2}}{\delta^2}.$$

Proof. Event whose P-probability we want to bound above is included in the union over $j \geq 0$ of the events

$$\sup\left\{\|\Sigma_{X_1^n}(F)\|_\infty \,:\, F \in \widetilde{\mathcal{F}}(G), P(F) \leq 2^{j+1}v\right\} \geq 2^j nv \cdot \delta.$$

Assumption **A7** yields that the former event indexed by j is itself included in the following

$$\sup\left\{\|\Sigma_{X_1^n}(F)\|_\infty \,:\, F \in \mathcal{F}(\widetilde{G}_j)\right\} \geq 2^j nv \cdot \delta$$

for some subset $\widetilde{G}_j$ of G satisfying $P(\widetilde{G}_j) \leq \gamma 2^{j+1}v$. Lemma 5.4.1 allows to conclude. □

5.5. The case of known cardinality of the true partition

5.5.1. Definition of the estimator

We address in this section the consistency of our estimator when the cardinal K of the estimator of $\tau^\star$ is *a priori* fixed.

The estimator $(\hat{\tau}_n, \hat{\theta}_n)$ of $(\tau^\star, \theta^\star)$ is constructed by minimization over $\mathcal{T}_K \times \Theta_K$ of the contrast J_n, or equivalently of the centered contrast U_n, with

$$\begin{aligned} J_n(\tau,\theta) &= \sum_{k=1}^{K} W_n(\tau_k, \theta_k), \\ U_n(\tau,\theta) &= J_n(\tau,\theta) - J_n(\tau^\star,\theta^\star) = u_n(\tau,\theta) + e_n(\tau,\theta) \end{aligned}$$

where

$$\begin{aligned} u_n(\tau,\theta) &= n^{-1}\sum_{j=1}^{K^\star}\sum_{k=1}^{K} n_{kj} v(\theta_j^\star, \theta_k) \qquad \text{and} \\ e_n(\tau,\theta) &= n^{-1}\sum_{j=1}^{K^\star}\sum_{k=1}^{K}\Big\{S_n(\tau_k \cap \tau_j^\star;\theta_k) - S_n(\tau_k \cap \tau_j^\star;\theta_j^\star)\Big\}. \end{aligned}$$

In the sequel, we will denote $\hat{\theta}_n(\tau_k) = \arg\min\{W_n(\tau_k,\theta) : \theta \in \overline{\Theta}\}$ for any $\tau \in \mathcal{T}_K$ and $1 \leq k \leq K$. Observe then that $(\hat{\theta}_n(\tau_k))_k = \arg\min\{J_n(\tau,\theta) : \theta \in \overline{\Theta_K}\}$. Moreover, we will denote $\hat{\theta}_n^\star = \hat{\theta}_n(\tau^\star)$. We will write $\hat{\theta}_{nj}^\star$ for the j^{th} coordinate of $\hat{\theta}_n^\star$ and $\hat{\theta}_{nj}$ for the j^{th} coordinate of $\hat{\theta}_n = \hat{\theta}_n(\hat{\tau}_n)$.

The next proposition casts some light on the behaviour of the total fluctuation term e_n. It is a direct consequence of Lemma 5.4.1 since $S_n(\tau_k \cap \tau_j^\star;\theta) = \psi(\theta)^T \Sigma_{X_1^n}(\tau_k \cap \tau_j^\star)$ and ψ is bounded.

Proposition 5.5.1. *Under assumptions of Lemma* 5.4.1, e_n *is uniformly* $o_{\mathrm{P}}(1)$ *over* $\mathcal{T}_K \times \Theta_K$.

5.5.2. Consistency

Consistency is of course hopeless for $K < K^\star$, since then $g(\tau,\tau^\star) \geq \Delta^\star/2$. We prove that our estimator is consistent as soon as $K \geq K^\star$:

Theorem 5.5.2. *Set $K \geq K^\star$ and let $(\hat{\tau}_n, \hat{\theta}_n)$ be the estimator defined in Section 5.5.1. Under Assumptions* **A1**, **A3**, **A4** *and* **A6**, *$(\hat{\tau}_n, \hat{\theta}_n)$ is consistent,* i.e. *that both $g(\hat{\tau}_n, \tau^\star)$ and $d_2(\hat{\theta}_n, \theta^\star)$ converge to 0 in* P*-probability.*

Proof of Theorem 5.5.2 is based on a technical lemma of great importance throughout this chapter, namely Lemma 5.5.3, and on application of Proposition 5.5.1.

Lemma 5.5.3. *Under Assumption* **A1**, *there exists $C^\star > 0$ such that, for any K and all $\delta > 0$,*

$$\lim_{n\to\infty} \mathrm{P}\left(\bigcap_{l\geq n}\left[\forall\,(\tau,\theta) \in \mathcal{T}_K \times \Theta_K,\ u_l(\tau,\theta) \geq C^\star d_v(\theta,\theta^\star)\right]\right) = 1,$$

$$\lim_{n\to\infty} \mathrm{P}\left(\bigcap_{l\geq n}\left[\forall\,(\tau,\theta) \in \mathcal{T}_{K,\delta} \times \Theta_K,\ u_l(\tau,\theta) \geq C^\star g(\tau,\tau^\star)\right]\right) = 1.$$

A sketch of proof of Lemma 5.5.3 can be found in Appendix 5.7.3. We are able now to demonstrate Theorem 5.5.2.

Proof. (Theorem 5.5.2*)* Let us prove that $g(\hat{\tau}_n, \tau^\star) = o_{\mathrm{P}}(1)$. We would prove that $d_2(\hat{\theta}_n, \theta) = o_{\mathrm{P}}(1)$ along the same lines.

If $g(\hat{\tau}_n, \tau^\star) > \delta$, then $\inf\{U_n(\tau,\theta) : (\tau,\theta) \in \mathcal{T}_{K,\delta} \times \Theta_K\}$ is non positive and consequently, thanks to Lemma 5.5.3, for some constant $c > 0$, for any $\varepsilon > 0$ and for n large enough,

$$\sup\left\{|e_n(\tau,\theta)| : (\tau,\theta) \in \mathcal{T}_K \times \Theta_K\right\} > c\delta$$

with probability at least $1 - \varepsilon$. Thus, proof is complete, since e_n is uniformly $o_{\mathrm{P}}(1)$. □

5.5.3. Rate of convergence

The following theorem gives lower bounds for rates of convergence of $\hat{\tau}_n$ and $\hat{\theta}_n$ for $K = K^\star$. Observe that the case $K > K^\star$ would be dealt with the same proof, up to minor changes.

Theorem 5.5.4. *Set $K = K^\star$ and let $(\hat{\tau}_n, \hat{\theta}_n)$ be the estimator defined in Section 5.5.1. Under Assumptions* **A1** *to* **A7**, *assuming moreover that the coefficient σ that appears in Assumption* **A5** *satisfies $\sigma \geq 2/h$, the sequences*

$$\{r_n^{-1}\, g(\hat{\tau}_n, \tau^\star)\} \quad \textit{and} \quad \{n^{(2-h)/2(\sigma-1)}\, d_2(\hat{\theta}_n, \theta^\star)\}$$

are uniformly bounded in P*-probability.*

Remark 5.5.5. This lower bound of the rate of convergence of $\hat{\tau}_n$ does not depend on the dependence structure of the sequence (Y_i) since the coefficient h of Assumption **A6** does not appear in the bound and $\{r_n\}$ depends solely on X_1^n: $g(\hat{\tau}_n, \tau^\star) = O_{\mathrm{P}}(r_n)$. On the contrary, the rate of convergence of $\hat{\theta}_n$ does depend on the dependence structure of (Y_i). Moreover, this rate is the same than when the true partition $\tau^\star$ is known.

Proof. Denote in the sequel $\zeta = 1/2(\sigma - 1)$. We shall prove that both probabilities

$$\mathrm{P}\Big(\delta_n \geq g(\hat{\tau}_n, \tau^\star) \geq \delta r_n \,,\, d_2(\hat{\theta}_n, \theta^\star) < \delta_n\Big), \tag{5.3}$$

$$\mathrm{P}\left(\delta_n \geq d_2(\hat{\theta}_n, \theta^\star) \geq \delta n^{\zeta(h-2)} \,,\, g(\hat{\tau}_n, \tau^\star) < \delta_n\right) \tag{5.4}$$

tend to zero as $n, \delta \uparrow \infty$, where the sequence $\{\delta_n\} \downarrow 0$ slowly enough, precisely with $n^{\zeta(2-h)}\delta_n \to \infty$, and is driven from consistency. Let us deal first with the first of them: we restrict ourselves to the event whose probability is written in (5.3).

By definition, $\hat{\tau}_n$ minimizes $\tau \mapsto U_n'(\tau) = \sum_{k=1}^{K^\star} W_n(\tau_k; \hat{\theta}_n(\tau_k)) - \sum_{j=1}^{K^\star} W_n(\tau_j^\star; \hat{\theta}_{nj}^\star)$. Furthermore, the definition of $\hat{\theta}_{nj}^\star$ and simple decomposition of $U_n'(\tau)$ yield the next first inequality

$$\begin{aligned}
U_n'(\tau) &= \sum_{j=1}^{K^\star} W_n(\tau_j^\star; \hat{\theta}_{nj}) - \sum_{j=1}^{K^\star} W_n(\tau_j^\star; \hat{\theta}_{nj}^\star) + \sum_{j=1}^{K^\star} \{W_n(\tau_j \nabla \tau_j^\star; \hat{\theta}_{nj}) - W_n(\tau_j^\star \nabla \tau_j; \hat{\theta}_{nj})\} \\
&\geq \sum_{j=1}^{K^\star} \{W_n(\tau_j \nabla \tau_j^\star; \hat{\theta}_{nj}) - W_n(\tau_j^\star \nabla \tau_j; \hat{\theta}_{nj})\} \\
&\geq n^{-1} \sum_{j=1}^{K^\star} \sum_{k \neq j} \Big\{ n_{kj} v^\star + S_n(\tau_k \cap \tau_j^\star; \hat{\theta}_{nk}) - S_n(\tau_k \cap \tau_j^\star; \hat{\theta}_{nj}) \Big\},
\end{aligned}$$

where the previous one holds as soon as $\hat{\theta}_n$ is close enough to $\theta^\star$, that is for n large.

The point is now to separate the fluctuations in terms of X_i and Y_i.

Since $g(\tau, \tau^\star) \geq \delta r_n$, there exist $k_0 \neq j_0$ such that $P\left(\tau_{k_0} \cap \tau_{j_0}^\star\right) \geq \delta r_n$, up to substitution of δ. Furthermore, Assumption **A2** yields that, for any $\varepsilon > 0$ and for n, δ large enough, for any $F \in \mathcal{F}$ such that $P(F) \geq \delta r_n$, we have

$$\mathbb{P}_n(F) \geq P(F)/2 \geq \delta r_n/2, \tag{5.5}$$

with probability $1 - \varepsilon$. Forwardly, up to other substitutions on δ and $v^\star$, applying Cauchy-Schwarz's inequality and invoking boundedness of ψ, we get the next first inequality

$$\begin{aligned}
U_n'(\tau) &\geq n^{-1} \sum_{j=1}^{K^\star} \sum_{k \neq j} \Big\{ (n_{kj} \vee \delta n r_n)\, v^\star + S_n(\tau_k \cap \tau_j^\star; \hat{\theta}_{nk}) - S_n(\tau_k \cap \tau_j^\star; \hat{\theta}_{nj}) \Big\} \\
&\geq C n^{-1} \sum_{j=1}^{K^\star} \sum_{k \neq j} \left\{ \left(nP\left(\tau_k \cap \tau_j^\star\right) \vee \delta n r_n\right) - c \left\| \Sigma_{X_1^n}(\tau_k \cap \tau_j^\star) \right\|_\infty \right\}
\end{aligned}$$

with probability at least $1 - \varepsilon$ (c is a positive constant, independent of τ).

The delicate point in the previous display takes place in the second inequality. We have to verify that $n_{kj} \vee \delta n r_n \geq CnP\left(\tau_k \cap \tau_j^\star\right) \vee \delta n r_n$. Carefully considering cases where, on the one hand, n_{kj} is greater than $\delta n r_n$, and on the other hand, where it is less than $\delta n r_n$ (with subcases $P(\tau_k \cap \tau_j^\star)$ greater or less than δr_n) yields the expected result. Note that this inequality does not hold anymore if we replace $\{r_n\}$ by some sequence $\{r'_n\}$ that decreases faster to 0.

Hence, the proof will be complete if we show that the convergences to 0 as $n, \delta \uparrow \infty$ of the probabilities of the following events hold for any $c > 0$ and $j_0 \neq j_1$ (where F denotes any set of the form $\tau \cap \tau_{j_1}^\star$ with $\tau \in \mathcal{F}$ such that $P\left(\tau \,\Delta\, \tau_{j_0}^\star\right) \leq \delta_n$):

$$\sup\left\{\|\Sigma_{X_1^n}(F)\|_\infty \,:\, P(F) \leq \delta r_n\right\} \geq \delta n r_n,$$
$$\sup\left\{\frac{\|\Sigma_{X_1^n}(F)\|_\infty}{nP(F)} \,:\, P(F) \geq \delta r_n\right\} \geq c.$$

This is a direct consequence of Lemmas 5.4.1 and 5.4.4.

Let us show now that the expression in (5.4) goes to 0 when $n, \delta \uparrow \infty$, too. Observe that for n large enough and on the events whose probabilities are given by (5.4), we have (with a view to application of Lemma 5.5.3) the lower bounding $d_v(\hat{\theta}_n, \theta^\star) \geq B\delta^\sigma n^{\sigma\zeta(h-2)}$. Moreover, Assumption **A1** together with $g(\tau, \tau^\star) < \min_j P\left(\tau_j^\star\right)/2$ imply that, for any $\varepsilon > 0$ and for n large enough, $\mathbb{P}_n(\tau_j \cap \tau_j^\star) \geq P\left(\tau_j^\star\right)/4$ with probability at least $1 - \varepsilon$. Thus, for any (τ, θ) of $\mathcal{T}_{K^\star} \times \Theta_{K^\star}$ satisfying the same conditions than $(\hat{\tau}_n, \hat{\theta}_n)$ on the events whose probabilities are written in (5.4), we have with probability at least $1 - \varepsilon$, for n large enough and any $1 \leq j \leq K^\star$,

$$u_n(\tau, \theta) \geq C^\star \delta^\sigma n^{\sigma\zeta(h-2)} \vee a_1 \|\theta_j - \theta_j^\star\|_2^\sigma \tag{5.6}$$

for some $a_1 > 0$ independent of (τ, θ). Note that the first term in the maximum comes from Lemma 5.5.3. To conclude this first step, remark that the preceding inequality together with the following one

$$x/y^\sigma \vee y^{\sigma/(\sigma-1)} z \geq x^{1/\sigma} z^{(\sigma-1)/\sigma} \;(x, y, z > 0),$$

yield (for some constant $a_2 > 0$ independent of (τ, θ))

$$u_n(\tau, \theta) \geq a_2 \|\theta_j - \theta_j^\star\|_2 \delta^{\sigma-1} n^{h/2-1}. \tag{5.7}$$

The second step consists of the same kind of arguments than in the first part of the proof: we will bound from below $(\tau, \theta) \mapsto U_n(\tau, \theta) = \sum_{k=1}^{K^\star} W_n(\tau_k; \theta_k) - \sum_{j=1}^{K^\star} W_n(\tau_j^\star; \theta_j^\star)$ taking care of separating cases $k = j$ and $k \neq j$ and distributing weight we know we can count on. Precisely, consider (τ, θ) such as above: on events of probability at least $1 - \varepsilon$, for n large enough,

$$U_n(\tau, \theta) \geq n^{-1} \sum_{j=1}^{K^\star} \left\{a_3\, n u_n(\tau, \theta) + S_n(\tau_j \cap \tau_j^\star; \theta_j) - S_n(\tau_j \cap \tau_j^\star; \theta_j^\star)\right\}$$
$$+ n^{-1} \sum_{j=1}^{K^\star} \sum_{k \neq j} \left\{(n_{kj} \vee n u_n(\tau, \theta)) a_4 + S_n(\tau_k \cap \tau_j^\star; \theta_k) - S_n(\tau_k \cap \tau_j^\star; \theta_j^\star)\right\},$$

for some constants $a_3, a_4 > 0$. Hence, thanks to Assumption **A2** (as before, see (5.5)), Taylor-Lagrange's inequality and (5.6), (5.7), with probability $1 - 2\varepsilon$ for n, δ large enough,

$$U_n(\tau, \theta) \geq C n^{-1} \sum_{j=1}^{K^\star} \left\{\|\theta_j - \theta_j^\star\|_2 \delta^{\sigma-1} n^{h/2} - a \|\Sigma_{X_1^n}(\tau_j \cap \tau_j^\star)\|_\infty \|\theta_j - \theta_j^\star\|_2\right\}$$
$$+ C n^{-1} \sum_{j=1}^{K^\star} \sum_{k \neq j} \left\{\left(nP\left(\tau_k \cap \tau_j^\star\right) \vee \delta^\sigma n^{\sigma\zeta(h-2)+1}\right) b - c \|\Sigma_{X_1^n}(\tau_j \cap \tau_j^\star)\|_\infty\right\},$$

for constants $a, b, c > 0$.

Consequently, the proof will be complete if we show that the convergence to 0 as $n, \delta \uparrow \infty$ of the probabilities of the following events hold for any $c > 0$ and $j_0 \neq j_1$ (where F denotes any set of the form $\tau \cap \tau_{j_1}^\star$ with $\tau \in \mathcal{F}$ such that $P\left(\tau \triangle \tau_{j_0}^\star\right) \leq \delta_n$):

$$\sup\left\{\|\Sigma_{X_1^n}(F)\|_\infty : F \in \mathcal{F}\right\} \geq \delta n^{h/2},$$
$$\sup\left\{\|\Sigma_{X_1^n}(F)\|_\infty : P(F) \leq \delta n^{\sigma\zeta(h-2)}\right\} \geq \delta n^{\sigma\zeta(h-2)+1},$$
$$\sup\left\{\frac{\|\Sigma_{X_1^n}(F)\|_\infty}{nP(F)} : P(F) \geq \delta n^{\sigma\zeta(h-2)}\right\} \geq c.$$

This is a direct consequence of Lemmas 5.4.1 and 5.4.4 since $\sigma \geq 2/h$. □

5.5.4. Number of misclassified observations

The standard scheme of proof applied to show that the number of misclassified observations is $O_{\mathrm{P}}(1)$ requires to bound below (up to a multiplicative constant) a generic term of the form $\mathbb{P}_n(F)$ by $P(F)$ for $P(F)$ possibly less than δr_n. Thus, Assumption **A2** is useless and we can not conclude. This difficulty is overcome independently of the dimension d when proving that the number of misclassified observations is $O_{\mathrm{P}}(nr_n)$. We can obtain the boundedness in probability in the 1-dimensional case. Proof relies then on the natural ordering over $\mathbb{R}$.

Proposition 5.5.6. *Let $N_n(\hat{\tau}_n) = \sum_{j=1}^{K^\star} \sum_{k \neq j} n_{kj}$ denote the number of misclassified observations for $\hat{\tau}_n$ with respect to $\tau^\star$. Under assumptions of Theorem* 5.5.4, *$N_n(\hat{\tau}_n) = O_{\mathrm{P}}(1)$ for $d = 1$ and $N_n(\hat{\tau}_n) = O_{\mathrm{P}}(nr_n)$ for higher dimensions.*

Proof. (Sketch of) Let $\eta \in \{0, 1\}$. If $N_n(\hat{\tau}_n) \geq \delta(nr_n)^\eta$ and $g(\hat{\tau}_n, \tau^\star) \vee d_2(\hat{\theta}_n, \theta^\star) \leq \delta r_n$ (study of that configuration suffices thanks to Theorem 5.5.4), then $\hat{\tau}_n$ minimizes U_n' which is lower bounded (up to the usual substitutions and to some multiplicative constant) for large enough n, δ by

$$n^{-1} \sum_{j=1}^{K^\star} \sum_{k \neq j} \left\{(n_{kj} \vee \delta(nr_n)^\eta) - c\|\Sigma_{X_1^n}(\tau_k \cap \tau_j^\star)\|_\infty\right\}.$$

Set $d = 1$ and $\eta = 0$. We can follow the strategy of proof in (Lavielle 1999). Application of triangle's inequality shifts the problem to the control of the P-probabilities of the following events (and their left-symmetric)

$$\sup\left\{\|\sum_{l=0}^{t} Z_{(s^\star+t)}\|_\infty : t^\star, 0 \leq t \leq \delta\right\} \geq \delta,$$
$$\sup\left\{\frac{\|\sum_{l=0}^{t} Z_{(s^\star+t)}\|_\infty}{t} : t^\star, t \geq \delta\right\} \geq c.$$

Here, $Z_{(s)} = Z_i$ for $X_{(s)} = X_i$ ($\{X_{(s)}\}_s$ denotes the increasing ordered vector X_1^n). Index $t^\star$ in the supremum ranges over all right extremities of intervals constituting subsets of $\tau^\star$. For some $t^\star$, $X_{(s^\star)}$ corresponds to the nearest X_i greater than $t^\star$. Such probabilities do go to zero, as

provided by the simplest Móricz's one dimensional inequalities that we apply here in place of Lemmas 5.4.1 and 5.4.4, and proof is complete for this case. Heuristically, the conclusion holds because the union of all intervals containing at most δ points X_i of observation contains itself a $O(\delta)$ number of such points.

For $d \geq 2$, we can not proceed as above. Actually, the union of all subsets $\tau_k \cap \tau_j^\star$ that contain at most δ points X_i of observation may contain much more than $O(\delta)$ points itself, *i.e.* generally a $O(n)$. Thus, we must conclude as in proof of Theorem 5.5.4, *i.e.* choose $\eta = 1$ and conclude that $N_n(\hat{\tau}_n) = O_{\mathrm{P}}(nr_n)$. □

Remark 5.5.7. Note that in a very specific multidimensional case usually called *pixel case*, we get $N_n(\hat{\tau}_n) = O_{\mathrm{P}}(1)$. Indeed, suppose that $\mathcal{F}_0$ is composed of rectangles and that $[0,1]^d$ is decomposed into the union of n^d mutually disjoint rectangular boxes, the *pixels*. Suppose that $X_1, \ldots, X_{n^d}$ are chosen uniformly in each box. Then, $N_n(\hat{\tau}_n) = O_{\mathrm{P}}(1)$. The scheme of proof applied in the one dimensional case above applies here, because the union of all subsets $\tau_k \cap \tau_j^\star$ that contain at most δ points X_i of observation contains a $O(\delta(\log\delta)^{d-1})$ points.

5.6. The case of unknown cardinality of the true partition

5.6.1. Definition of the estimator

We address in this Section the case of an unknown cardinality $K^\star$ of $\tau^\star$. According to the former Section, we can construct an estimator $(\hat{\tau}_{n,K}, \hat{\theta}_{n,K})$ of any cardinality K, *i.e.* for any *a priori* fixed cardinality of the estimator. The question is to select the best estimator among the family $(\hat{\tau}_{n,K}, \hat{\theta}_{n,K})_K$. Naturally, models with large cardinality K are favoured, hence the idea of penalizing the contrast J_n by adding a penalization term $\beta_n K$. Its role is to balance out this effect.

Thus, estimation of the triplet $(K^\star, \tau^\star, \theta^\star)$ is performed by minimizing a penalized contrast constructed with J_n as defined in Section 5.5. The estimator of $(K^\star, \tau^\star, \theta^\star)$ is taken to achieve the minimization of the penalized contrast $\widetilde{J}_n$ given by

$$\widetilde{J}_n(K, \tau, \theta) = J_n(\tau, \theta) + \beta_n K$$

for $K \in \{1, \ldots, K_{\mathbf{max}}\}$ and (τ, θ) ranging over $\mathcal{T}_K \times \Theta_K$ (recall that $K^\star$ is bounded above by $K_{\mathbf{max}}$, as a consequence of the definition of a $\mathcal{F}$-partition).

The sequence $\{\beta_n\}$ is positive and tends to zero. The difficulty relies on the choice of its rate of convergence: since large β_n (precisely slow rate of convergence) favours simple models (that is models with small cardinality K) and *vice-versa*, the point is to calibrate its rate of convergence. Indeed, β_n appears as a trade-off between fitting the observations and avoiding too big models to be selected. Actually, the calibration will follow from the rate of convergence of the estimator $(\hat{\tau}_n, \hat{\theta}_n)$ studied in the previous Section, for *a priori* known $K^\star$.

5.6.2. Consistency

Theorem 5.6.1. *Let $\{\beta_n\}$ be a sequence of positive numbers satisfying both*

$$\beta_n \to 0 \qquad \text{and} \qquad n^{(2-h)/2(\sigma-1)}\beta_n \to \infty.$$

Under the assumptions of Theorem 5.5.4, $\widehat{K}_n = K^\star$ *with* P*-probability tending to one. Hence, the consistency of $(\hat{\tau}_n, \hat{\theta}_n)$ as defined in Theorem* 5.5.2 *still holds.*

Proof. Proof that $\mathrm{P}(\widehat{K}_n < K^\star)$ tends to zero is straightforward with Lemma 5.5.3. Indeed, if $\widehat{K}_n < K^\star$, then $g(\hat{\tau}_n, \tau^\star) \geq \Delta^\star$ (see Property (iii)), hence $u_n(\hat{\tau}_n, \hat{\theta}_n) \geq C^\star \Delta^\star$ with probability tending to one. Moreover, $\widehat{K}_n < K^\star$ yields $\beta_n(K^\star - \widehat{K}_n) \geq u_n(\hat{\tau}_n, \hat{\theta}_n) + e_n(\hat{\tau}_n, \hat{\theta}_n)$. Thus (since $\beta_n \to 0$) we can conclude if we control for n large enough the sum over K from 1 to $K^\star - 1$ of the probabilities that $\sup\{|e_n(\tau, \theta)|\} > c$ for some $c > 0$ (where (τ, θ) in the supremum ranges over $\mathcal{T}_K \times \Theta_K$). This part is then complete, because e_n is uniformly $o_{\mathrm{P}}(1)$.

Let us consider now $\mathrm{P}(\widehat{K}_n > K^\star)$. It is bounded above by the sum over K from $K^\star + 1$ up to $K_{\max}$ of probabilities that

$$\inf\left\{U_n(\tau, \theta) \,:\, (\tau, \theta) \in \mathcal{T}_K \times \Theta_K\right\} + \beta_n \leq 0.$$

For (τ, θ) ranging over $\mathcal{T}_{K,\beta_n} \times \Theta_K$, we can replace the previous events by

$$\inf\left\{U_n'(\tau) \,:\, \tau \in \mathcal{T}_{K,\beta_n}\right\} \leq 0$$

and proceed as in the first part of proof of Theorem 5.5.4; and when it ranges over $\mathcal{T}_K \times \Theta_{K,\beta_n}$, proof is similar to its second part (that is why we impose $n^{(2-h)/2(\sigma-1)}\beta_n \to \infty$). Thus, we have to focus on the P-probabilities of those events for (τ, θ) in $\mathcal{T}_K \times \Theta_K$ satisfying $g(\tau, \tau^\star) \vee d_2(\theta, \theta^\star) \leq \beta_n$.

For (τ, θ) as described above, we get, applying Taylor-Lagrange's inequality (b, c are some positive constants):

$$U_n(\tau, \theta) + \beta_n \geq e_n(\tau, \theta) + \beta_n \geq \pi n^{-1} \sum_{j=1}^{K^\star} \sum_{k \in K_j} \left\{n\beta_n - b\beta_n \left\|\Sigma_{X_1^n}(\tau_k \cap \tau_j^\star)\right\|_\infty\right\}$$
$$+ \pi n^{-1} \sum_{j=1}^{K^\star} \sum_{k \notin K_j} \left\{n\beta_n - c\left\|\Sigma_{X_1^n}(\tau_k \cap \tau_j^\star)\right\|_\infty\right\}.$$

Finally, the proof will be complete if we show that the convergence to 0 as $n \uparrow \infty$ of the probabilities of the following events hold for any $c > 0$ and $j_0 \neq j_1$ (where F denotes any set of the form $\tau \cap \tau_{j_1}^\star$ with $\tau \in \mathcal{F}$ such that $P(\tau \,\Delta\, \tau_{j_0}^\star) \leq \delta_n$, for $\{\delta_n\} \downarrow 0$ driven from consistency):

$$\sup\left\{\left\|\Sigma_{X_1^n}(F)\right\| \,:\, F \in \mathcal{F}\right\} \geq cn,$$
$$\sup\left\{\left\|\Sigma_{X_1^n}(F)\right\| \,:\, P(F) \leq \beta_n\right\} \geq cn\beta_n.$$

Once again, this is a direct consequence of Lemmas 5.4.1 and 5.4.4. □

5.7. Appendix

5.7.1. Proof of Proposition 5.3.3

Proof. Obviously, it suffices to prove that

$$\lim_{\eta \to \infty} \lim_{n \to \infty} \mathrm{P}\left(\sup\left\{\left|\frac{P(F) - \mathbb{P}_n(F)}{P(F)}\right| \,:\, F \in \mathcal{F}, P(F) \geq \eta r_n\right\} \geq \frac{1}{2}\right) = 0.$$

Thanks to Talagrand's concentration inequalities for the supremum of empirical processes (see Massart 2000, Theorem 2.4, p266), basic analysis yields an upper bound $\exp\{-f(n,\eta)\}$ where $f(n,\eta) > 0$ tends to infinity as $n, \eta \uparrow \infty$, as soon as the expectation of the supremum in the former equation goes to zero.

Let us first study the following expectation for fixed integers n, p and some $\eta > 0$:

$$\mathrm{E}\left(\sup\left\{\left|\frac{P(F)-\mathbb{P}_n(F)}{P(F)}\right| : F \in \mathcal{F}_n^p\right\}\right),$$

for $\mathcal{F}_n^p = \{F : F \in \mathcal{F},\ 2^p\eta r_n \leq P(F) < 2^{p+1}\eta r_n\}$. It is bounded from above by

$$2^{-p}(\eta r_n)^{-1}\mathrm{E}\Big(\sup\{|P(F)-\mathbb{P}_n(F)| : F \in \mathcal{F}_n^p\}\Big).$$

Symmetrization arguments (refer to Massart 2000) yield that the previous expression is bounded above by

$$2^{-p+1}(\eta n r_n)^{-1}\,\mathrm{E}\left(\sup\left\{\left|\sum_{i=1}^n \varepsilon_i \mathbb{1}\{X_i \in F\}\right| : F \in \mathcal{F}_n^p\right\}\right) \tag{5.8}$$

for independent identically distributed Rademacher rv ε_i (*i.e.* $\mathrm{P}(\varepsilon_i = 1) = \mathrm{P}(\varepsilon_i = -1) = 1/2$) independent of X_1^n. Furthermore, for any $\mathcal{C} \subset \mathcal{F}$, the next result holds (as a consequence of Hoeffding's inequality, see Problem 2.14.8 of van der Vaart and Wellner 1996): for $a = \sup_{F\in\mathcal{C}} P(F)$, V the VC-dimension of $\mathcal{F}$ and some constant $A(\mathcal{F})$ depending on $\mathcal{F}$ only,

$$\mathrm{E}\left(\sup\left\{\left|\sum_{i=1}^n \varepsilon_i \mathbb{1}\{X_i \in F\}\right| : F \in \mathcal{C}\right\}\right) \leq Cn^{1/2}\left[\left(a + \frac{V}{n}\log\frac{V}{a}\right)\log\frac{A(\mathcal{F})}{a}\right]^{1/2}.$$

We apply this result together with the inequality $(x+y)^{1/2} \leq x^{1/2} + y^{1/2}$ $(x, y > 0)$ to (5.8) and get that, for $A' = A(\mathcal{F}) \vee V$,

$$\mathrm{E}\left(\sup\left\{\left|\frac{P(F)-\mathbb{P}_n(F)}{P(F)}\right| : F \in \mathcal{F}_n^p\right\}\right) \leq C\sqrt{2}^{-p}\left(\frac{\log(A'(\eta r_n)^{-1})}{\eta n r_n}\right)^{1/2} + C2^{-p}(2A')^{1/2}\frac{\log(A'(\eta r_n)^{-1})}{\eta n r_n}.$$

Finally, the expectation of the supremum of interest over $\mathcal{F}_n$ is bounded above by the next sum that can be controlled as shown

$$\sum_{p\geq 0}\mathrm{E}\left(\sup\left\{\left|\frac{P(F)-\mathbb{P}_n(F)}{P(F)}\right| : F \in \mathcal{F}_n^p\right\}\right) \leq C\left(\frac{\log(A'(\eta r_n)^{-1})}{\eta n r_n}\right)^{1/2} + CA'^{1/2}\frac{\log(A'(\eta r_n)^{-1})}{\eta n r_n}.$$

The result follows immediately. □

5.7.2. Exploring Assumption A6

Two alternative assumptions

We propose in this Section to study Assumption **A6**. In order to make things clearer, we will suppose through this Section that $r = 1$, that is that ξ is real valued. Thus, $\|\cdot\|_\infty$ is systematically replaced by absolute values. The results are easy to adapt to the general case. Let us start with a very special case where slight control of the conditional second order moments suffices to ensure Assumption **A6**. State

A6a There exist $C_0 > 0$ and $h \in (1,2)$ such that, for any $G \in \mathcal{G}$,

$$\mathrm{E}\left(\Sigma_{X_1^n}(G)^2 \,\middle|\, X_1^n\right) \leq C_0 \left(\sum_{i=1}^n \mathbb{1}\{X_i \in G\}\right)^h \text{ P-as.}$$

Note that **A6a** would obviously be satisfied for $h = 1$ if the rv Z_i were independent. Moreover, Assumption **A6a** implies Assumption **A6** in the basic but fundamental case of rectangles:

Proposition 5.7.1. *Assumption* **A6** *is satisfied as soon as* **A6a** *holds when* $\mathcal{F}_0$ *is composed of rectangles.*

Proof of Proposition 5.7.1 relies on an adaptation of the method proposed in (Móricz, Serfling, and Stout 1982; Móricz 1983) to show such a result on the real line and its extension to the multidimensional case. It can be done by induction and uses then thoroughly basic properties of decomposition of rectangles in union of rectangles.

The sequel draws its inspiration from the theory of dependent variables and random fields, see (Doukhan 1994; Rio 2000) and especially (Dedecker 2001). Actually, a natural loosened conditional Marcinkiewicz-Zygmund inequality yields Assumption **A6** for VC class $\mathcal{F}$. Indeed, let Assumption **A6b** consists of the following:

A6b There exist $C_0 > 0$ and $h \in [1,2)$ such that, for any $p > 2$ and $G \in \mathcal{G}$,

$$\mathrm{E}\left(|\Sigma_{X_1^n}(G)|^p \,\middle|\, X_1^n\right) \leq C_0{}^p p^{\frac{p}{2}} \left(\sum_{i=1}^n \mathbb{1}\{X_i \in G\}\right)^{h\frac{p}{2}} \text{ P-as.}$$

Remark 5.7.2 (on Assumption A6b). The previous inequality is said to be "loosened" because of the power h in the right hand term, where $h = 1$ is usually expected. It is sharp an inequality thanks to the particular form of the factor $C_0{}^p p^{\frac{p}{2}}$: it allows some efficient optimization in p producing the expected final result *via* Pisier's method for some rich class $\mathcal{F}$ – namely finite VC-dimension VC class. Precise statement is given by Proposition 5.7.3. It underlines how exceptional seems Proposition 5.7.1, where only control of second order moments is needed, to compare with Assumption **A6b** and control of moments of order any $p > 2$. It is known that such a simple condition of control of moments can not be sufficient in the simple case of polygons.

Proposition 5.7.3. *Assumption* **A6** *is satisfied as soon as Assumption* **A6b** *holds when* $\mathcal{F}$ *is a finite VC-dimension VC class.*

Proof. All the inequalities to come hold P-as. We set $\Sigma_n(G)$ for $\Sigma_{X_1^n}(G)$ and E_n for $\mathrm{E}(\cdot \mid X_1^n)$. $\|\cdot\|_p$ denotes the L^p norm with respect to the conditional probability $\mathrm{P}(\cdot \mid X_1^n)$.

Pisier's method (see Dedecker 2001) consists of writing

$$\begin{aligned}\mathrm{E}_n\Big(\sup_{F\in\mathcal{F}(G)}|\Sigma_n(F)|^2\Big) &\le \Big\|\sup_{F\in\mathcal{F}(G)}|\Sigma_n(F)|\Big\|_p^2 \le \left(\sum_{F\in\Gamma_n}\mathrm{E}_n(|\Sigma_n(F)|^p)\right)^{\frac{2}{p}} \\ &\le N_n^{\frac{2}{p}}\max_{F\in\Gamma_n}\|\Sigma_n(F)\|_p^2 \le N_n^{\frac{2}{p}} p\, {C_0}^2\left(\sum_{i=1}^n \mathbb{1}\{X_i\in G\}\right)^h,\end{aligned}$$

where Γ_n denotes a family of sets of minimal cardinality N_n such that any separation of $\{X_1,\dots,X_n\}\cap G$ by elements of $\mathcal{F}(G)$ can be achieved with an element of Γ_n. Since $\mathcal{F}$ has finite VC-dimension, so does $\mathcal{F}(G)$ and N_n is finite. Set $H_n = \log N_n$: VC theory (see for instance Vapnik 1998) ensures that H_n is bounded above by $V(1+\log\sum \mathbb{1}\{X_i\in G\})$ (V denotes the VC-dimension of $\mathcal{F}$).

Optimization in p yields

$$\begin{aligned}\mathrm{E}_n\Big(\sup_{F\in\mathcal{F}(G)}|\Sigma_n(F)|^2\Big) &\le 2{C_0}^2 H_n\left(\sum_{i=1}^n \mathbb{1}\{X_i\in G\}\right)^h \\ &\le 2{C_0}^2 V\left(\sum_{i=1}^n \mathbb{1}\{X_i\in G\}\right)^h\left(1+\log\sum_{i=1}^n \mathbb{1}\{X_i\in G\}\right).\end{aligned}$$

Set $\varepsilon>0$ such that $h+\varepsilon<2$: there exists C_1 depending on ε but neither on G nor on n such that

$$\mathrm{E}_n\Big(\sup_{F\in\mathcal{F}(G)}|\Sigma_n(F)|^2\Big) \le C_1\left(\sum_{i=1}^n \mathbb{1}\{X_i\in G\}\right)^{h+\varepsilon}$$

and Markov's inequality concludes the proof. □

Assumptions A6a and A6b for regular lattices

We propose here to illustrate both Assumption **A6a** and Assumption **A6b** in a simple but nevertheless natural and interesting case where $\mathcal{X}=\mathbb{Z}^d$ is a regular lattice and the field $(Z_x)_{x\in\mathcal{X}}$ is bounded and strictly stationary. The sequel is widely inspired from (Dedecker 2001) again. Our aim is to determine some ultimate assumptions that imply both **A6a** and **A6b**.

First, let us recall the definition of the ϕ-mixing coefficient $\phi(\mathcal{U},\mathcal{V})$ of two σ-algebras $\mathcal{U}$ and $\mathcal{V}$ of $\mathcal{A}$: it is given by

$$\phi(\mathcal{U},\mathcal{V}) = \sup\{\|P(V|\mathcal{U})-P(V)\|_\infty \,:\, V\in\mathcal{V}\}.$$

Note that $\phi(\mathcal{U},\mathcal{V})\in[0,1]$, with value 0 for independent σ-fields only. Let us introduce the nonincreasing sequence $\{\phi(t)\}_{t\ge1}$ with definition

$$\phi(t) = \sup\Big\{\phi\Big(\sigma(Z_y,\ y\in\mathcal{Y}),\sigma(Z_x)\Big)\,:\, x\in\mathcal{X},\ \mathcal{Y}\subset\mathcal{X}\ ,\ d(x,\mathcal{Y})\ge t\Big\},$$

where $\mathcal{Y}$ denotes any finite subset of $\mathcal{X}$ and $d(x,\mathcal{Y})$ is the infimum for y ranging throughout $\mathcal{Y}$ of $d(x,y) = \min_{1\le i\le d}|x_i - y_i|$. Its role is to resume the whole system of dependency of the field $(Z_x)_{x\in\mathcal{X}}$.

We can state now the final result of this section and infer from it a sufficient rate of convergence to 0 for $\{\phi(t)\}$ to ensure that Assumptions **A6a** and **A6b** hold true:

Proposition 5.7.4. *Suppose that $\mathcal{X} = \mathbb{Z}^d$ and that $(Z_x)_{x\in\mathcal{X}}$ is bounded and strictly stationary. Assumptions* **A6a** *and* **A6b** *are satisfied as soon as, for some $1 \le h < 2$, $C_0 > 0$ depending on h and for any $n \ge 1$,*

$$\sum_{t=1}^{n} t^{d-1}\phi(t) \le C_0\, n^{h-1}. \tag{5.9}$$

The previous inequality is satisfied e.g. *when $\phi(t) = O(t^{-(d+1-h)})$.*

Proposition 5.7.4 is a corollary of Proposition 5.7.6 below. The strategy of the proof relies on a Burkholder-like inequality as shown by Dedecker in 2001, namely in Proposition 1 (a) of this paper. Following his method, we get that, for any $G \in \mathcal{G}$ and $\mathcal{Y}_n \subset \mathcal{X}$ of cardinality n,

$$\begin{aligned}\mathrm{E}(|\Sigma_{\mathcal{Y}_n}(G)|^p) &\le \left(2p\sum_{x\in\mathcal{Y}_n}\left\{\|{Z_x}^2\|_{p/2} + \sum_{x'\in\mathcal{Y}_n}\|Z_{x'}\mathrm{E}_{d(x,x')}(Z_x)\|_{p/2}\mathbb{1}\{x'\in G\}\right\}\mathbb{1}\{x\in G\}\right)^{p/2}\\ &\le (2p)^{p/2}\left(\|Z_0\|_\infty + {\|Z_0\|_\infty}^2\right)\\ &\qquad\left(\sum_{x\in\mathcal{Y}_n}\mathbb{1}\{x\in G\} + \sum_{x,x'\in\mathcal{Y}_n}\|\mathrm{E}_{d(x,x')}(Z_x)\|_{p/2}\mathbb{1}\{x,x'\in G\}\right)^{p/2}.\end{aligned} \tag{5.10}$$

In the previous display, $\mathrm{E}_{d(x,x')}(Z_x)$ denotes the conditional expectation of Z_x with respect to the σ-field $\sigma(Z_y : y\in\mathcal{Y}_n,\ d(x,y)\ge d(x,x'))$.

We can derive from (5.10) some first ultimate condition on the field $(Z_x)_{x\in\mathcal{X}}$:

Proposition 5.7.5. *Suppose that $\mathcal{X} = \mathbb{Z}^d$ and that $(Z_x)_{x\in\mathcal{X}}$ is centered, bounded and strictly stationary. Then Assumptions* **A6a** *and* **A6b** *hold as soon as, for some $1 \le h < 2$, $C_0 > 0$ depending on h and for any $p \ge 2$, $G\in\mathcal{G}$, $n\ge 1$ and $\mathcal{Y}_n \subset \mathcal{X}$ of cardinality n, for any $x\in\mathcal{Y}_n$,*

$$\sum_{x'\in\mathcal{Y}_n}\|\mathrm{E}_{d(x,x')}(Z_x)\|_{p/2} \le C_0 n^{h-1}.$$

*The condition above holds with $h = 1$ for m-conditionally centered fields (*i.e. *fields such that $\mathrm{E}_{d(x,x')}(Z_x) = 0$ for $d(x,x')\ge m$). This includes m-dependence and consequently, independence.*

Furthermore, combining the next upper bound for the right hand term of (5.10)

$$(2p)^{p/2}\left(\|Z_0\|_\infty + {\|Z_0\|_\infty}^2\right)\left(\sum_{x\in\mathcal{Y}_n}\mathbb{1}\{x\in G\} + \sum_{x,x'\in\mathcal{Y}_n}\|\mathrm{E}_{d(x,x')}(Z_x)\|_\infty\mathbb{1}\{x,x'\in G\}\right)^{p/2}$$

with Serfling's inequality (see Serfling 1968)

$$\|\mathrm{E}_{d(x,x')}(Z_x)\|_\infty \leq 2\phi(d(x,x'))\|Z_x\|_\infty,$$

yields:

Proposition 5.7.6. *Suppose that $\mathcal{X} = \mathbb{Z}^d$ and that $(Z_x)_{x\in\mathcal{X}}$ is centered, bounded and strictly stationary. Then Assumptions* **A6a** *and* **A6b** *hold as soon as, for some $1 \leq h < 2$, $C_0 > 0$ depending on h and for any $G \in \mathcal{G}$, any $n \geq 1$ and any $\mathcal{Y}_n \subset \mathcal{X}$ of cardinality n, for any $x \in \mathcal{Y}_n$,*

$$\sum_{x'\in\mathcal{Y}_n} \|\mathrm{E}_{d(x,x')}(Z_x)\|_\infty \leq C_0 n^{h-1}.$$

Condition (5.9) in Proposition 5.7.4 is sufficient.

5.7.3. Proof of Lemma 5.5.3

Proof. To begin with, we will establish a smart and simple lower bounding of expressions of the type $f_{ij}(\theta^\star,\alpha) = \inf\{\alpha v(\theta_j^\star,\theta) + (1-\alpha)v(\theta_i^\star,\theta) \,:\, \theta\in\Theta\}$ for $1 \leq i \neq j \leq K^\star$ and $0 < \alpha < 1$. Let $A_{ij}(\theta^\star) = 2f_{ij}(\theta^\star, 1/2)$ and $\underline{A}(\theta^\star) = \min A_{ij}(\theta^\star)$, denoted $\underline{A}$ in the following. With those definitions, $\underline{A} > 0$ and for all $i \neq j$ and $0 < \alpha < 1$,

$$f_{ij}(\theta^\star,\alpha) \geq \underline{A}\min(\alpha, 1-\alpha).$$

Set $\varepsilon > 0$ and define the events $\Omega_n(\delta)$ ($\delta > 0$, $n \geq 1$) as

$$\Omega_n(\delta) = \bigcap_{l\geq n}\Big[\sup\Big\{|P_l(F) - P(F)| \,:\, F\in\mathcal{F}\Big\} \leq \delta\Big]. \tag{5.11}$$

Since $\mathrm{P}(\Omega_n(\delta)) \uparrow 1$ as $n \uparrow \infty$, there exists $n_1 \geq 1$ such that $\mathrm{P}(\Omega_{n_1}(\delta_1)) \geq 1-\varepsilon$ with $\delta_1 = \Delta^\star/16K_{\max}^2$ (choice of this particular value $\alpha\Delta^\star$ will be justified at the very end of the proof). Let us restrict ourselves to the event $\Omega_{n_1}(\delta_1)$ and consider (τ,θ) in $\mathcal{T}_{K,\Delta^\star/4K_{\max}} \times \Theta_K$, $n \geq n_1$. We can prove that $u_n(\tau,\theta)$ is bounded from below by a positive constant independent of n and (τ,θ). Indeed, it is sufficient to prove the existence of k, j_0, j_1 such that n_{kj_0}/n and n_{kj_1}/n are both bounded from below by such a constant c, since then, for $\alpha = n_{kj_0}/(n_{kj_0}+n_{kj_1})$,

$$\begin{aligned} u_n(\tau,\theta) &\geq \frac{n_{kj_0}+n_{kj_1}}{n}\Big(\alpha v(\theta_{j_0}^\star,\theta_k) + (1-\alpha)v(\theta_{j_1},\theta_k)\Big) \\ &\geq c\,\underline{A}. \end{aligned}$$

Furthermore, since $g(\tau,\tau^\star) > \Delta^\star/4\overline{K}$, if $g(\tau,\tau^\star) = P(\cup_{\mathcal{K}_{j_0}}\tau_k \,\Delta\, \tau_{j_0}^\star)$ (where the union is taken over $k\in\mathcal{K}_{j_0}$), we have either $A_1 > \Delta^\star/8\overline{K}$ or $A_2 > \Delta^\star/8\overline{K} \geq A_1$ for

$$A_1 = P\left(\tau_{j_0}^\star \nabla \bigcup_{k\in\mathcal{K}_{j_0}} \tau_k\right) \quad\text{and}\quad A_2 = P\left(\bigcup_{k\in\mathcal{K}_{j_0}} \tau_k \nabla \tau_{j_0}^\star\right).$$

Consider the first case: there necessarily exists $k \notin \mathcal{K}_{j_0}$ such that $P(\tau_k \cap \tau_{j_0}^\star) \geq \Delta^\star/8\overline{K}^2$. We also know that $k \notin \mathcal{K}_{j_0}$ yields $P(\tau_k\nabla\tau_{j_0}^\star) \geq P(\tau_k \cap \tau_{j_0}^\star)$ (see Proposition 5.2.4 (i)), hence

there exists $j_1 \neq j_0$ such that $P(\tau_k \cap \tau_{j_1}^\star) \geq \Delta^\star/8\overline{K}^3$. Consequently, both n_{kj_0}/n and n_{kj_1}/n are bounded from below by some adequate $c > 0$. Consider now the second case. There necessarily exists $k \in \mathcal{K}_{j_0}$ such that $P(\tau_k \nabla \tau_{j_0}^\star) \geq \Delta^\star/8\overline{K}^2$ and therefore $j_1 \neq j_0$ such that $P(\tau_k \cap \tau_{j_1}^\star) \geq \Delta^\star/8\overline{K}^3$. If $\mathrm{card}(\mathcal{K}_{j_0}) > 1$, then $P(\tau_k \cap \tau_{j_0}^\star) \geq P(\tau_k \nabla \tau_{j_0}^\star)$ (see Proposition 5.2.4 (i) again), whereas if $\mathcal{K}_{j_0} = \{j_0\}$, then $P(\tau_{j_0}^\star \nabla \tau_k) = A_1 \leq \Delta^\star/8K_{\max}$ implies $P(\tau_k \cap \tau_{j_0}^\star)$ roughly bounded from below by $\Delta^\star/2$ and we conclude as above.

At last, since $g(\tau, \tau^\star) \leq 1$ and v is bounded from above by its supremum over the compact set $\overline{\Theta} \times \overline{\Theta}$, the study of the case $g(\tau, \tau^\star) > \Delta^\star/4\overline{K}$ is completed.

Let us deal now with the lower bounding in $d_v(\theta, \theta^\star)$ for the models (τ, θ) ranging over $(\mathcal{T}_K - \mathcal{T}_{K,\Delta^\star/4K_{\max}}) \times \Theta_K$. Since

$$u_n(\tau, \theta) \geq \max_{1 \leq j \leq K^\star} \max_{k \in \mathcal{K}_j} \frac{n_{kj}}{n} v(\theta_j^\star, \theta_k),$$

it suffices to bound by below all $P(\tau_k \cap \tau_j^\star)$'s by some positive constant independent of n and (τ, θ) and greater than δ_1. So, let k be in $\mathcal{K}_j$ for some j. If $\tau_k \subset \tau_j^\star$, then $P(\tau_k \cap \tau_j^\star) \geq \Delta^\star > \delta_1$. Suppose now that $\tau_k \not\subset \tau_j$. First, if $\mathcal{K}_j = \{k\}$, then $P(\tau_k \nabla \tau_j^\star) \leq g(\tau, \tau^\star) \leq \Delta^\star/4\overline{K}$ yields $P(\tau_k \cap \tau_j^\star) \geq \Delta^\star/2 > \delta_1$. Secondly, for $\mathrm{card}(\mathcal{K}_j) > 1$, $k \in \mathcal{K}_j$ ensures that $P(\tau_k \nabla \tau_j^\star) \leq P(\tau_k \cap \tau_j^\star)$ (see Proposition 5.2.4 (i) once again), hence $2P(\tau_k \cap \tau_j^\star) \geq \Delta^\star$ and that concludes this part of the proof.

Set $\delta > 0$. We still have to show that $u_n(\tau, \theta) \geq C^\star g(\tau, \tau^\star)$ for τ verifying $\delta < g(\tau, \tau^\star) \leq \Delta^\star/4\overline{K}$ and $\theta \in \Theta_K$, with probability larger than $1 - \varepsilon$. First, as above, there exists $n_0 \geq n_1$ such that $\mathrm{P}(\Omega_{n_0}(\delta_0)) \geq 1 - \varepsilon$, for $\delta_0 = \min(\delta_1, \delta/2)$. Let us restrict ourselves to the event $\Omega_{n_0}(\delta_0)$ from now. Set $(\tau, \theta) \in \mathcal{T}_{K,\delta} \times \Theta_K$ with $g(\tau, \tau^\star) \leq \Delta^\star/4\overline{K}$.

Suppose that $g(\tau, \tau^\star)$ is achieved for j_0 and $\mathcal{K}_{j_0}$. Since then

$$\sum_{k \notin \mathcal{K}_{j_0}} \frac{n_{kj_0}}{n} + \sum_{j \neq j_0} \sum_{k \in \mathcal{K}_{j_0}} \frac{n_{kj}}{n} \geq g(\tau, \tau^\star)/2,$$

it is sufficient to prove that $u_n(\tau, \theta)$ is greater than the left hand term of the preceding inequality, up to some positive multiplicative constant independent of both n and (τ, θ).

Let k be in $\mathcal{K}_{j_0}$. Remember we obtained among other things that $P(\tau_k \cap \tau_{j_0}^\star) \geq \Delta^\star/2$, hence $n_{kj_0}/n \geq (1 - 1/2\overline{K})\,\Delta^\star/2$. Furthermore, if $\tau_k \cap \tau_j^\star \neq \emptyset$ for some $j \neq j_0$, then necessarily $P(\tau_k \cap \tau_j^\star) \leq g(\tau, \tau^\star) \leq \Delta^\star/4\overline{K}$ (see Proposition 5.2.4 (ii)) and consequently, $n_{kj}/n \leq \Delta^\star/2\overline{K}$. Thus, $n_{kj_0}/n \geq n_{kj}/n$ and

$$\begin{aligned} u_n(\tau, \theta) &\geq \frac{n_{kj_0} + n_{kj}}{n} \Big(\alpha v(\theta_{j_0}^\star, \theta_k) + (1 - \alpha) v(\theta_j, \theta_k) \Big) \\ &\geq \alpha \frac{n_{kj_0} + n_{kj}}{n} \underline{A} \\ &= \frac{n_{kj}}{n} \underline{A}, \end{aligned}$$

for $\alpha = n_{kj_0}/(n_{kj_0} + n_{kj})$ and forwardly

$$\overline{K}^2\, u_n(\tau, \theta) \geq \underline{A} \sum_{j \neq j_0} \sum_{k \in \mathcal{K}_{j_0}} \frac{n_{kj}}{n}.$$

Moreover, if $k \notin \mathcal{K}_{j_0}$, $P(\tau_k \cap \tau^\star_{j_0}) \leq g(\tau, \tau^\star)$, hence $n_{kj_0}/n \leq \Delta^\star/2\overline{K}$. There also exists $j \neq j_0$ such that $P(\tau_k \cap \tau^\star_j) \geq \Delta^\star/\overline{K}$ and then, $n_{kj}/n \geq \Delta^\star/2\overline{K}$ (we choose $\alpha\Delta^\star$ so that this expected inequality holds). We derive as above that

$$\overline{K}\, u_n(\tau, \theta) \geq \underline{A} \sum_{k \notin \mathcal{K}_{j_0}} \frac{n_{kj_0}}{n},$$

which concludes the proof. □

6

Interlude : A motivated introduction to Orlicz spaces and some Large Deviations Principles*

Résumé

Cet interlude est dédié à la présentation d'un cadre de travail, d'outils et de résultats que nous exploiterons à des fins d'étude de propriétés statistiques asymptotiques de deux estimateurs de l'ordre d'un modèle dans le prochain chapitre. Il comprend quatre parties complémentaires. Nous introduisons dans la première d'entre elles la notion d'espace de Orlicz, considérant tout particulièrement les formes linéaires qui agissent sur de tels espaces. La seconde partie est consacrée à un espace de Orlicz particulier qui nous intéressera au premier chef, ainsi qu'à l'énoncé (commenté) d'un résultat de Grandes Déviations de type Sanov pour la mesure empirique dû à Léonard et Najim. Nous nous penchons dans la troisième partie sur deux résultats de Moyennes Déviations pour la mesure empirique recentrée, l'un nouveau, l'autre dû à Wu. Ce dernier seulement nous permettra de mener à terme l'une des preuves cruciales du prochain chapitre. La dernière partie du présent interlude consiste en la preuve du nouveau principe de Moyennes Déviations avec une approche orientée vers les espaces de Orlicz.

Abstract

We present in the following interlude a framework, a few tools and results that the forthcoming study of some statistical asymptotic properties of two estimators of the order of a model will require in the next chapter. It is divided into four complementary parts. We introduce in the first one the setting of Orlicz spaces, with a special interest for linear forms on such sets. Then, the second part is devoted to a Sanov Large Deviations Principle for the empirical measure in a framework of Orlicz spaces due to Léonard and Najim. The third part is concerned with two Moderate Deviations Principles for the centered empirical measure, one of them new, the other one proved by Wu. The latter theorem will allow to complete an important proof in the next chapter. The fourth part is finally dedicated to the proof of the new Moderate Deviations Principle, again with an Orlicz spaces approach.

*I would like to thank Stéphane Boucheron for the reference (Léonard and Najim 2000) as well as Christian Léonard and Jamal Najim for stimulating discussions on their Sanov theorem.

Au menu

6.1. Orlicz spaces **195**
6.1.1. Introducing general Orlicz spaces 195
6.1.2. Continuous and singular parts 196
6.2. A Sanov's LDP with a view to statistical application **197**
6.2.1. Introducing a particular Orlicz space 197
6.2.2. The functions $H(\cdot\,|P^\star)$ and $I(\cdot\,|P^\star)$ 197
6.2.3. Statement of the Sanov's LDP 199
6.3. Two MDP with a view to statistical application **200**
6.3.1. Introduction 200
6.3.2. Statements 201
6.3.3. Comparison 203
6.4. Appendix **204**

6.1. Orlicz spaces

This section introduces the Orlicz spaces setting we shall need hereafter. It is inspired from a preliminary section of (Léonard and Najim 2000). The first subsection is devoted to the very basis of the theory, while the second one is concerned with some useful refinements. For a comprehensive reference, see (Rao and Ren 1991).

In the sequel, we shall use linear functional notations for expectations and integrals, *i.e.* write μf for the integral $\int f \mathrm{d}\mu$ of a function f with respect to a measure μ.

6.1.1. Introducing general Orlicz spaces

Let τ denote an even convex nonnegative function on $\mathbb{R}_+$, that admits at least a finite value and such that $\tau(s)$ tends to infinity as s tends to infinity (τ is then a *Young function*). Let $P^\star$ be a probability measure on the measurable space $(\mathcal{Y}, \mathcal{F})$. One can define the four vector spaces

$$\begin{aligned} \mathcal{L}_\tau &= \mathcal{L}_\tau(P^\star) = \left\{ f \in \mathbb{R}^{\mathcal{Y}} : \exists\, a > 0,\ P^\star \tau(f/a) < \infty \right\}, \\ \mathcal{M}_\tau &= \mathcal{M}_\tau(P^\star) = \left\{ f \in \mathbb{R}^{\mathcal{Y}} : \forall\, a > 0,\ P^\star \tau(f/a) < \infty \right\} \subset \mathcal{L}_\tau \end{aligned}$$

and L_τ, M_τ which correspond respectively to $\mathcal{L}_\tau$, $\mathcal{M}_\tau$ with identification of $P^\star$-almost everywhere ($P^\star$-ae) equal functions. Observe that M_τ contains the set of all the bounded measurable functions. The following formula

$$\|f\|_\tau = \inf \left\{ a > 0 : P^\star \tau(f/a) \leq 1 \right\},\ f \in L_\tau$$

defines a norm on L_τ such that $(L_\tau, \|\cdot\|_\tau)$ is a Banach space, called *Orlicz space* associated with τ.

Let τ^* be the convex conjugate of τ: since τ^* is also a Young function, one can consider the Orlicz space L_{τ^*}. Any $g \in L_{\tau^*}$ defines a continuous linear form on L_τ for the duality bracket $\langle f, g \rangle = P^\star fg$. Indeed, for any $f \in L_\tau, g \in L_{\tau^*}$, $fg \in L^1(P^\star)$ thanks to the inequality

$$P^\star |fg| \leq 2 \|f\|_\tau \|g\|_{\tau^*}. \tag{6.1}$$

Thus, L_{τ^*} can be identified with a subspace of the topological dual L'_τ of $(L_\tau, \|\cdot\|_\tau)$. Indeed, one can prove that $(L_{\tau^*}, \|\cdot\|_{\tau^*})$ is isomorphic to the topological dual M'_τ.

Introduce now L^*_τ and $\mathcal{L}^*_\tau$, which are respectively the algebraic dual of L_τ and $\mathcal{L}_\tau$. In the sequel, the following crucial inclusions will prevail:

Proposition 6.1.1.

$$L_{\tau^*} \subset L'_\tau \subset L^*_\tau \subset \mathcal{L}^*_\tau. \tag{6.2}$$

6.1.2. Continuous and singular parts

We will actually use a comprehensive description of L'_τ (in view of the identification of L_{τ^*} as a subspace of L'_τ) stated in Theorem 6.1.3 (for a proof of this result, see Léonard 2000). First, let us define the $P^\star$-singular elements of L'_τ. Since L_τ is a Riesz space (one can define $f_1 \vee f_2$ and $f_1 \wedge f_2$ for any $f_1, f_2 \in L_\tau$, hence $|f| = f \vee 0 - f \wedge 0$) and since any $Q \in L'_\tau$ is relatively bounded ($\{Q(f) : |f| \leq g\}$ is bounded), then L'_τ is also a Riesz space (and forwardly, an element $|Q| = Q \vee 0 - Q \wedge 0$ of L'_τ is associated with $Q \in L'_\tau$).

Definition 6.1.2. A linear form $Q \in L'_\tau$ is $P^\star$-singular if there exists a sequence $\{A_p\}$ of measurable sets such that $|Q|\mathbb{1}\{A_p^c\} = 0$ for all $p \geq 1$ and moreover $\lim_p P^\star(A_p) = 0$.

We denote L^s_τ the set of all $P^\star$-singular elements of L'_τ. Then

Theorem 6.1.3. *Suppose that τ is finite. The topological dual L'_τ of $(L_\tau, \|\cdot\|_\tau)$ is isomorphic to the direct sum $L_{\tau^*} \oplus L^s_\tau$.*

Thus, any continuous linear form $Q \in L'_\tau$ on L_τ is uniquely decomposed into the sum $Q = Q^a + Q^s$, where Q^a and Q^s are continuous, Q^s is $P^\star$-singular and

$$Q^a f = P^\star \frac{\mathrm{d}Q^a}{\mathrm{d}P^\star} f \text{ (any } f \in L_\tau),$$

where $\frac{\mathrm{d}Q^a}{\mathrm{d}P^\star} \in L_{\tau^*}$ is the sole element of L_{τ^*} such that the previous relation does hold. Here, Q^a and Q^s are respectively the *continuous part* and the *singular part* of Q.

The next results illustrate their respective contribution to Q: the first one is based on the definition of the $P^\star$-singularity and the dominated convergence theorem and the second one is a consequence of Proposition 6.1.4 and the dominated convergence theorem again.

Proposition 6.1.4. *Suppose that τ is finite and take $Q^s \in L^s_\tau$. Then $Q^s(M_\tau) = \{0\}$.*

Lemma 6.1.5. *Let $Q \in L'_\tau$ and Q^a, Q^s be its continuous and singular parts, respectively. Choose $f \in L_\tau$ and a sequence $\{f_n\}$ of bounded functions that converges pointwise to f with $\sup_n |f_n| = O(f)$. Then*

$$Qf_n \underset{n\to\infty}{\longrightarrow} Q^a f \quad \text{and} \quad Qf\mathbb{1}\{|f| > n\} \underset{n\to\infty}{\longrightarrow} Q^s f.$$

6.2. A Sanov's Large Deviations Principle with a view to statistical application

This section is devoted to the presentation of some useful tools, including a Sanov's Large Deviations Principle (*i.e.* a Large Deviations Principle for the empirical measure). Those tools are required further by our statistical study of the rate of underestimation of two estimators of the order of a model, whose properties (consistency, underestimation and overestimation) are investigated in Chapter 7.

We shall first introduce a particular Orlicz space, namely the Orlicz space of functions that admit some exponential moments wrt a fixed measure. Then, we shall concentrate on two intimately related tools, one of them being the Kullback-Leibler divergence. And finally, a Sanov's Large Deviations Principle (LDP) due to Léonard and Najim (2000) will be stated, and commented.

6.2.1. Introducing a particular Orlicz space

In the sequel, the Young function τ is set to $\tau(s) = \exp(|s|) - |s| - 1$ $(s \in \mathbb{R})$. Then, $\tau^*(t) = (1+|t|)\log(1+|t|) - |t|$ $(t \in \mathbb{R})$. Hence,

$$\begin{aligned} \mathcal{L}_\tau &= \left\{ f \in \mathbb{R}^{\mathcal{Y}} : \exists\, a > 0,\ P^\star \exp(a|f|) < \infty \right\}, \\ \mathcal{M}_\tau &= \left\{ f \in \mathbb{R}^{\mathcal{Y}} : \forall\, a > 0,\ P^\star \exp(a|f|) < \infty \right\}. \end{aligned}$$

If $f \in \mathcal{L}_\tau$, we say that f admits *some* exponential moments and if $f \in \mathcal{M}_\tau$, we say that f admits *all* exponential moments.

For any $f \in \mathcal{L}_\tau$, $Q \mapsto Qf$ defines an element of $\mathcal{L}_\tau^*$. We denote $\sigma(\mathcal{L}_\tau^*, \mathcal{L}_\tau)$ the coarsest topology on $\mathcal{L}_\tau^*$ that makes those linear forms continuous (note that it is not metrizable), and $\varsigma(\mathcal{L}_\tau^*, \mathcal{L}_\tau)$ the smallest σ-field that makes them measurable. In view of a future proof, let us emphasize that the topology $\sigma(\mathcal{L}_\tau^*, \mathcal{L}_\tau)$ is generated by the collection of open sets

$$\mathcal{O}(f, x, \varepsilon) = \{Q \in \mathcal{L}_\tau^* : |Qf - x| < \varepsilon\},$$

for any $f \in \mathcal{L}_\tau$, $x \in \mathbb{R}$, $\varepsilon > 0$.

We introduce the subspace $\mathcal{Q}$ of $\mathcal{L}_\tau^*$

$$\mathcal{Q} = \{Q \in \mathcal{L}_\tau^* : Q \geq 0,\ Q1 = 1\}$$

which is endowed with the σ-field $\mathcal{S}$ induced by $\varsigma(\mathcal{L}_\tau^*, \mathcal{L}_\tau)$ and the topology $\mathcal{T}$ induced by $\sigma(\mathcal{L}_\tau^*, \mathcal{L}_\tau)$.

6.2.2. The functions $H(\cdot\,|P^\star)$ and $I(\cdot\,|P^\star)$

Let us recall for sake of completeness the definition of the Kullback-Leibler divergence, first introduced by Kullback and Leibler in the early 1950's. Its role is very important in Probability, Statistics and Information theory. Denote $M_1(\mathcal{Y})$ the set of all probability measures on $(\mathcal{Y}, \mathcal{F})$. The Kullback-Leibler divergence of the probability measures Q wrt P is defined by

$$H(Q\,|P) = P\frac{\mathrm{d}Q}{\mathrm{d}P}\log\frac{\mathrm{d}Q}{\mathrm{d}P} = Q\,\log\frac{\mathrm{d}Q}{\mathrm{d}P}$$

when $Q << P$, $H(Q\,|P) = \infty$ otherwise. One often mentions then that, although it is sometimes called Kullback-Leibler distance, it has not the mathematical properties of a distance. Nevertheless, comparison with the Hellinger (mathematical) distance affords to think in some situations of the Kullback-Leibler divergence *almost* as a distance, but this is beyond our scope. We give (without proof) for future use a glimpse of the properties that the Kullback-Leibler divergence satisfies in the proposition below (see for instance Dupuis and Ellis 1997):

Proposition 6.2.1 (some useful elementary properties of the KL divergence). *$H(\cdot\,|\cdot)$ is a convex, lower semicontinuous function on $M_1(\mathcal{Y}) \times M_1(\mathcal{Y})$. In particular, $H(\cdot\,|\cdot)$ is a convex, lower semicontinuous function of each variable, separately. Furthermore, for any $P \in M_1(\mathcal{Y})$, $H(\cdot\,|P)$ has compact level sets for the weak topology on $M_1(\mathcal{Y})$,* i.e. *that for any $\alpha \in \mathbb{R}$, the level sets*

$$\{Q \in M_1(\mathcal{Y}) : H(Q\,|P) \leq \alpha\}$$

are compact wrt the coarsest topology on $M_1(\mathcal{Y})$ that renders the functions $Q \mapsto Q\,f$ continuous for any f continuous and bounded. In other words, $H(\cdot\,|P)$ is a convex good rate function, according to the typical terminology of the Large Deviations theory.

Let us define now $I(\cdot|P^\star)$ on $\mathcal{Q}$ as

$$I(Q\,|P^\star) = Q^a \log\left(\frac{\mathrm{d}Q^a}{\mathrm{d}P^\star}\right) + \sup\{Q^s f : f \in \mathcal{L}_\tau,\ P^\star \exp(f) < \infty\}$$

for any $Q \in \mathcal{Q} \cap L'_\tau$ and $I(Q\,|P^\star) = \infty$ otherwise.

Note here that the inclusions (6.2) of Proposition 6.1.1 give sense to $\mathcal{Q} \cap L'_\tau$. The remainder of this subsection will aim at giving some arguments that allow to think of $I(\cdot\,|P^\star)$ as a generalization of $H(\cdot\,|P^\star)$.

Remark 6.2.2. Consider $Q \in \mathcal{Q}$ such that $I(Q\,|P^\star) < \infty$: then $Q = Q^a + Q^s$ with $Q, Q^a, Q^s \in L'_\tau$. We shall show that the continuous part Q^a of Q is identifiable with a probability measure and that Q^s is a nonnegative linear form.

First, Lemma 6.1.5 readily yields that $Q^s \geq 0$ since

$$Q^s f = \lim_{n\to\infty} Qf\mathbb{1}\{f > n\} \geq 0 \quad (\text{any } f \geq 0,\ f \in L_\tau)$$

as a limit of nonnegative numbers. Furthermore, Proposition 6.1.4 implies $Q^a 1 = 1$ and $Q^a \geq 0$ because

$$Q\mathbb{1}\left\{\frac{\mathrm{d}Q^a}{\mathrm{d}P^\star} < 0\right\} = Q^a\mathbb{1}\left\{\frac{\mathrm{d}Q^a}{\mathrm{d}P^\star} < 0\right\} = P^\star\frac{\mathrm{d}Q^a}{\mathrm{d}P^\star}\mathbb{1}\left\{\frac{\mathrm{d}Q^a}{\mathrm{d}P^\star} < 0\right\} \geq 0.$$

Hence $\dfrac{\mathrm{d}Q^a}{\mathrm{d}P^\star}$ is a probability density wrt $P^\star$. Besides, for any $F \in \mathcal{F}$,

$$Q^a\mathbb{1}\{F\} = P^\star\frac{\mathrm{d}Q^a}{\mathrm{d}P^\star}\mathbb{1}\{F\},$$

so Q^a is identifiable with a probability measure dominated by $P^\star$ with corresponding density $\frac{\mathrm{d}Q^a}{\mathrm{d}P^\star}$. Consequently, $I(Q^a\,|P^\star) = H(Q^a\,|P^\star)$, the Kullback-Leibler divergence of the probabilities Q^a and $P^\star$.

The next proposition casts some more light on the relationship between $I(\cdot\,|P^\star)$ and $H(\cdot\,|P^\star)$.

Proposition 6.2.3.
• *Set $Q \in \mathcal{Q} \cap M_1(\mathcal{Y})$,* i.e. *$Q \in \mathcal{Q}$ such that $Q(F) = Q\mathbb{1}\{F\}$ (any $F \in \mathcal{F}$) defines a probability measure on $\mathcal{Y}$. Then $I(Q\,|P^\star) < \infty$ yields $I(Q\,|P^\star) = H(Q\,|P^\star)$.*
• *Set $Q \in M_1(\mathcal{Y})$ such that $H(Q\,|P^\star) < \infty$. Then one has $\dfrac{\mathrm{d}Q}{\mathrm{d}P^\star} \in M_{\tau^*}$, hence $Q \in \mathcal{Q} \cap L'_\tau$, hence $I(Q\,|P^\star) = H(Q\,|P^\star)$.*
• *In particular, if $Q \in \mathcal{Q} \cap M_1(\mathcal{Y})$, then $H(Q\,|P^\star) = I(Q\,|P^\star)$.*

Proof. • Set $Q \in \mathcal{Q} \cap M_1(\mathcal{Y})$ and assume that $I(Q\,|P^\star) < \infty$. Thus, $Q = Q^a + Q^s$ in virtue of Theorem 6.1.3 with the usual notations. Remark 6.2.2 ensures that Q^a is identifiable with a probability measure dominated by $P^\star$.

Consequently, $Q^s = Q - Q^a$ is a signed measure with $Q^s(\mathcal{Y}) = 0$. For such a measure, the total variation $|Q^s|$ satisfies $|Q^s|(\mathcal{Y}) = 2\sup\{Q^s(F) : F \in \mathcal{F}\}$, which is zero here in virtue of Proposition 6.1.4. We therefore conclude that $Q = Q^a$ and $H(Q\,|P^\star) = I(Q\,|P^\star)$ in virtue of Remark 6.2.2 above.

• Set $Q \in M_1(\mathcal{Y})$ and suppose that $H(Q\,|P^\star) < \infty$. Then, $Q << P^\star$ with a probability density f wrt $P^\star$ which satisfies $P^\star f \log f < \infty$. Let us prove that $f \in M_{\tau^*}$, hence that $Q \in L'_\tau$ with a view to the inclusions (6.2). The conclusion then stems from Remark 6.2.2.

Set $a > 0$. The point is to show that $P^\star\tau^*(af) < \infty$, *i.e.* $P^\star(1 + af)\log(1 + af) < \infty$. This expression is bounded above by $2\log 2 + 2a\,P^\star\varphi_a(f)$ where $\varphi_a(t) = t\log(1 + at) > 0$. Now, $\varphi_a = O(\varphi)$ for $\varphi(t) = t\log t$. Hence, $\varphi_a = O(\varphi \vee 0)$ since $\varphi \wedge 0$ is bounded. We then conclude because $H(Q\,|P^\star) < \infty$ yields $P^\star(\varphi \vee 0)(f) < \infty$.

• The last point is readily derived from the two first ones. □

6.2.3. Statement of the Sanov's Large Deviations Principle

The Sanov's LDP stated below is one of the keys of our proof when we evaluate the rate of underestimation of two order estimators in Chapter 7. This proof involves the upper bound of the LDP and the evaluation of the infimum of a nonnegative function over the closure of a particular set. The latter evaluation can not be performed without drastic assumptions for closure wrt the weak topology or the τ-topology. Unfortunately, classical Sanov's theorem involve those topologies. On the contrary, one can perform the evaluation in very general settings for the stronger topology used in Theorem 6.2.4.

Let us state the extended Sanov's LDP of interest (proved by Léonard and Najim with a projective limit method). We set here a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ upon which all the random variables (rv) will be defined in the sequel of this chapter. We denote E the expectation wrt P.

Theorem 6.2.4 (Extended Sanov's theorem, Léonard and Najim). *Let $Y_1, \ldots, Y_n$ be a n-tuple of independent $P^\star$-identically distributed (i.i.d.) rv on $\mathcal{Y}$. We denote $\mathbb{P}_n$ the empirical measure, which is a rv on the set $\mathcal{Q}$. Then the sequence $\{\mathcal{L}(\mathbb{P}_n)\}$ of the distributions of $\mathbb{P}_n$ satisfies a LDP on $\mathcal{Q}$ equipped with the topology $\mathcal{T}$ and the σ-field $\mathcal{S}$ with the convex good rate function $I(\cdot\,|P^\star)$,* i.e. *for all $S \in \mathcal{S}$,*

$$-I(\mathrm{int}(S)\,|P^\star) \le \liminf_{n\to\infty} n^{-1}\log \mathrm{P}(\mathbb{P}_n \in S) \le \limsup_{n\to\infty} n^{-1}\log \mathrm{P}(\mathbb{P}_n \in S) \le -I(\mathrm{cl}(S)\,|P^\star),$$

and $I(\cdot\,|P^\star) : \mathcal{Q} \to [0,\infty]$ *is a convex, lower semicontinuous mapping such that* $\{Q \in \mathcal{Q} : I(Q\,|P^\star) \leq \alpha\}$ *is compact for any* $\alpha > 0$. *In other words,* $I(\cdot\,|P^\star)$ *is a convex, good rate function wrt the* $\mathcal{T}$*-topology.*

In the theorem, $\mathrm{int}(S)$ and $\mathrm{cl}(S)$ respectively denote the interior and the closure of S wrt the topology $\mathcal{T}$. Moreover, $I(\Pi\,|P^\star)$ denotes $\inf\{I(Q\,|P^\star) : Q \in \Pi\}$ (any $\Pi \subset \mathcal{Q}$), as for the whole sequel of this thesis.

Remark 6.2.5 (the sophisticated setting of Theorem 6.2.4 is not superfluous).

- **A Sanov theorem on $M_1(\mathcal{Y})$ is insufficient for our purpose:** as announced earlier, Theorem 6.2.4 will be the key of the proofs of Theorems 7.5.3 and 7.5.6. But the latter proofs rely on the continuity of the linear forms

$$Q \mapsto Q\,f$$

  on $\mathcal{Q}$ for some functions f that possibly admit only some exponential moments (*i.e.* $f \in \mathcal{L}_\tau \setminus \mathcal{M}_\tau$). Now, it is not possible to replace $\mathcal{Q}$ by $M_1(\mathcal{Y})$: indeed, Schied has proved in his 1998's paper that the extension of a Sanov's theorem on $M_1(\mathcal{Y})$ to a topology on $M_1(\mathcal{Y})$ that renders continuous the above mapping for some fixed f is possible if and only f admits all exponential moments (*i.e.* $f \in \mathcal{M}_\tau$), which is the classical Cramér condition.

- Provided that we are interested in the continuity of the linear forms above for some functions f that admit only some exponential moments, the $\sigma(\mathcal{L}_\tau^*, \mathcal{L}_\tau)$-topology on $\mathcal{L}_\tau^*$ is natural. Besides, it is a convenient framework for proofs of LDP based on a projective limit approach (see Theorem 6.4.1).

- **The Kullback-Leibler divergence would not be a convenient rate function:** (Schied 1998) also yields that one can not expect a Sanov's theorem on $\mathcal{Q}$ equipped with the $\sigma(\mathcal{Q}, \mathcal{L}_\tau)$-topology and with *good* rate function

$$I'(Q\,|P^\star) = I(Q^a\,|P^\star)$$

  when $Q \in \mathcal{Q} \cap L'_\tau$, infinity otherwise. This stems from the fact that $H(\cdot\,|P^\star)$ does not have compact level sets on $M_1(\mathcal{Y})$ equipped with the so-called τ^f-topology when $f \in \mathcal{L}_\tau \setminus \mathcal{M}_\tau$. The latter topology is the coarsest topology on $M_1(\mathcal{Y})$ that renders continuous the linear forms on $M_1(\mathcal{Y})$: $Q \mapsto Q\,g$, for any measurable g dominated by f. For further comments, refer to (Léonard and Najim 2000).

6.3. Two Moderate Deviations Principles with a view to statistical application

6.3.1. Introduction

This section is dedicated to the statement and comparison of two Moderate Deviations Principles (MDP), one of them new, with a view to statistical application. Indeed, we shall need such a MDP when trying to evaluate the rate of overestimation of two order estimators whose properties (consistency, rate of underestimation and overestimation) are carefully studied in Chapter 7.

Precisely, the scheme of proof that provides an evaluation of the rate of overestimation relies on the application of the upper bound of a MDP for centered empirical measures. One then has to evaluate the infimum of a function on the closure of some sets. Classical results of MDP already include such principles for the centered empirical measures in the weak topology, or the τ-topology, see for instance (de Acosta 1994). They are nonetheless insufficient for our purpose because, once again (see the preliminary comment on Theorem 6.2.4), the respective closures wrt those topologies are not easily handled.

We therefore proved a new MDP for the centered empirical measure, whose statement is to find in Theorem 6.3.1. It is in the spirit of the extended Sanov's LDP presented in Section 6.2, involving particularly the $\sigma(\mathcal{L}_\tau^*, \mathcal{L}_\tau)$-topology. Its proof, postponed in section Appendix 6.4, is quite alike the one of the extended Sanov's LDP proposed by Léonard and Najim.

Though this MDP may deal with functional spaces of unbounded functions (namely with spaces of functions that admit some exponential moments) and, from that point of view is more general than Wu's MDP for centered empirical measure (1994) (whose theorem typically applies to spaces of functions having an envelope function that admits some exponential moments), we only managed to conclude thanks to Wu's, because the associated underlying topology of his MDP is the uniform topology, which is stronger than ours. For more details, refer to Section 7.6 of Chapter 7.

6.3.2. Statements

We shall consider positive sequences $\{b_n\}$ which satisfy both

$$n^{1/2}\, b_n^{-1} = o(1) \quad \text{and} \quad b_n = o(n). \tag{6.3}$$

Besides, let us define the nonnegative convex map $J(\cdot\,|P^\star)$ on $M(\mathcal{Y})$ by

$$J(Q\,|P^\star) = P^\star \frac{1}{2}\left(\frac{\mathrm{d}Q}{\mathrm{d}P^\star}\right)^2$$

when Q is dominated by $P^\star$ and $Q\,\mathbb{1} = 0$, $J(Q\,|P^\star) = \infty$ otherwise.

Theorem 6.3.1. *Let $Y_1, \ldots, Y_n$ be a n-tuple of independent $P^\star$-identically distributed rv on $\mathcal{Y}$. Let $\mathbb{P}_n$ be the empirical measure and $(\mathbb{P}_n - P^\star)$ be the centered empirical measure, which are rv's on the set $M(\mathcal{Y})$ equipped with the σ-field $\mathcal{S}'$ induced by $\varsigma(\mathcal{L}_\tau^*, \mathcal{L}_\tau)$. Then, for any sequence $\{b_n\}$ such that (6.3) holds true, the sequence $\{\mathcal{L}(n\, b_n^{-1}(\mathbb{P}_n - P^\star))\}$ of the distributions of $n\, b_n^{-1}(\mathbb{P}_n - P^\star)$ satisfies a MDP on $M(\mathcal{Y})$ equipped with the topology $\mathcal{T}'$ induced by $\sigma(\mathcal{L}_\tau^*, \mathcal{L}_\tau)$ and the σ-field $\mathcal{S}'$ with good rate function $J(\cdot\,|P^\star)$ and normalizing sequence $\{n^{-1}\, {b_n}^2\}$, i.e. for all $S \in \mathcal{S}$,*

$$-J(\mathrm{int}(S)\,|P^\star) \le \liminf_{n\to\infty} n\, b_n^{-2} \log \mathrm{P}(n\, b_n^{-1}(\mathbb{P}_n - P^\star) \in S) \le$$
$$\limsup_{n\to\infty} n\, b_n^{-2} \log \mathrm{P}(n\, b_n^{-1}(\mathbb{P}_n - P^\star) \in S) \le -J(\mathrm{cl}(S)\,|P^\star),$$

and $J(\cdot\,|P^\star)$ is a lower semicontinuous mapping such that $\{Q \in M(\mathcal{Y}) : J(Q\,|P^\star) \le \alpha\}$ is compact for any $\alpha > 0$.

In the theorem, $\mathrm{int}(S)$ and $\mathrm{cl}(S)$ respectively denote the interior and the closure of S wrt $\mathcal{T}'$. Besides, $J(\Pi\,|P^\star)$ is by definition the infimum of $J(Q\,|P^\star)$ for Q ranging over Π (any $\Pi \subset M(\mathcal{Y})$).

Accordingly to what we announced earlier, let us now state Wu's MDP (refer to Wu 1994). We shall consider some sequences $\{v_n\}$ of positive numbers that satisfy (6.3) and such that the two more conditions below hold true:

$$\{v_n\} \text{ increasing} \quad \text{and} \quad v_{nk} \leq A\, k^{1-\delta}\, v_n \tag{6.4}$$

for some $A \geq 1$, $0 < \delta < 1$, any $n, k \geq 1$ (or in other words, $\{v_n\}$ is not too close to $\{n\}$). Furthermore, let $\mathcal{G}$ be a class of functions in $L^2(P^\star)$*. We denote $\ell^\infty(\mathcal{G})$ the set of all uniformly bounded, real functions on $\mathcal{G}$ (*i.e.* all functions $B \in \mathbb{R}^{\mathcal{G}}$ such that $\|B\|_{\mathcal{G}} = \sup_{g\in\mathcal{G}} |B\,g| < \infty$), equipped with the topology and the σ-field induced by the uniform norm $\|\cdot\|_{\mathcal{G}}$.

Any $Q \in M(\mathcal{Y})$ is associated an element of $\ell^\infty(\mathcal{G})$ denoted Q^∞, defined by

$$Q^\infty\, g = Q\, g \quad (\text{any } g \in \mathcal{G}).$$

We finally define, for any $B \in \ell^\infty$,

$$J^\infty(B\,|P^\star) = \inf\{J(Q\,|P^\star) : Q \in M(\mathcal{Y}),\ Q^\infty = B\}.$$

Theorem 6.3.2 (Wu). *Let $Y_1, \ldots, Y_n$ be a n-tuple of independent $P^\star$-identically distributed rv on $\mathcal{Y}$. Let $\mathbb{P}_n$ be the empirical measure and $(\mathbb{P}_n - P^\star)$ be the centered empirical measure. Then $(\mathbb{P}_n - P^\star)^\infty$ is a rv on the set $\ell^\infty(\mathcal{G})$ equipped with the uniform topology.*

Then, for any sequence $\{v_n\}$ such that (6.3) and (6.4) hold true, the sequence

$$\{\mathcal{L}(n\, v_n^{-1}(\mathbb{P}_n - P^\star)^\infty)\}$$

of the distributions of $n\, v_n^{-1}(\mathbb{P}_n - P^\star)^\infty$ satisfies a MDP on $\ell^\infty(\mathcal{Y})$ with rate function $J^\infty(\cdot\,|P^\star)$ and normalizing sequence $\{n^{-1}\, {v_n}^2\}$ if and only if the three conditions below hold true:

(i) $(\mathcal{G}, \|\cdot\|_2)$ *is totally bounded;*

(ii) $n\, v_n^{-1}(\mathbb{P}_n - P^\star)^\infty = o_{\mathrm{P}}(1)$;

(iii) *there exists $M > 0$ such that, for all $u > 0$,*

$$\limsup_{n\to\infty} n\, v_n^{-2} \log\left(n\, \mathrm{P}\left(\sup_{g\in\mathcal{G}} |g(Y_1)| > u\, v_n\right)\right) \leq -u^2/M.$$

Remark 6.3.3. This elegant theorem crucially relies on an earlier result due to Ledoux (1992). In the latter, the author derives some necessary and sufficient conditions (very similar to the conditions above in Theorem 6.3.2) in order to ensure that the majorization part of a MDP holds true. The setting is concerned with rv with values in a Banach space and the method of proof relies on isoperimetric techniques.

*We assume that either $\mathcal{G}$ is countable or that all the expressions to come involving suprema are measurable – it suffices *e.g.* that for each of them, suprema are P-almost surely equal to suprema over some countable subsets.

6.3.3. Comparison

Theorem 6.3.2 is certainly deeper than Theorem 6.3.1. It gives necessary and sufficient conditions so that the upper bound of the MDP hold true. These conditions give an insight on the phenomenon of MDP for the empirical measure wrt the strong topology of the uniform norm on $\ell^\infty(\mathcal{G})$. A typical case where all the conditions are satisfied is the following:

Lemma 6.3.4. *The conditions* (i), (ii), (iii) *of Theorem* 6.3.2 *hold true* e.g. *when*

(a) *one replaces* $n^{1/2}\, {v_n}^{-1} = o(1)$ *by* $(n \log n)^{1/2}\, {v_n}^{-1} = o(1)$;

(b) $\mathcal{G}$ *is* $P^\star$*-Donsker*[*];

(c) $\mathcal{G}$ *admits an envelope function* G *which have some exponential moments,* i.e. $G \in \mathcal{L}_\tau$.

Proof. Indeed, the two first conditions (i) and (ii) are satisfied thanks to Assumption (a) and (b). Besides, if a nonnegative envelope function G belongs to $\mathcal{L}_\tau$, let us denote $a > 0$ such that $P^\star \exp(a\, G) = C < \infty$; then one has

$$n\, \mathrm{P}\left(\sup_{g\in\mathcal{G}} |g(Y_1)| > u\, v_n\right) \leq n\, \mathrm{P}\left(G(Y_1) > u\, v_n\right) =$$

$$n\, \mathrm{P}\left(\exp\left\{a(G(Y_1) - u\, v_n)\right\} > 1\right) \leq n\, \exp(-a\, u\, v_n)\, P^\star \exp(a\, G),$$

hence

$$n\, {v_n}^{-2} \log\left(n\, \mathrm{P}\left(\sup_{g\in\mathcal{G}} |g(Y_1)| > u\, v_n\right)\right) \leq n\, {v_n}^{-2}(\log n + C) - u(a\, n\, {v_n}^{-1}) = o(1) - u\,[o(1)]^{-1}$$

and condition (iii) is also satisfied. □

We have to emphasize that Theorem 6.3.1 holds true for the $\sigma(\mathcal{L}_\tau^\star, \mathcal{L}_\tau)$-topology, which is coarser than the uniform topology. Nevertheless, our result applies in a framework where the elements of $\mathcal{G}$ admit some exponential moments, *i.e.* when the class $\mathcal{G}$ is included in $\mathcal{L}_\tau$, but does not require any assumption on an envelope function.

Finally, Theorem 6.3.1 is not a corollary of Theorem 6.3.2 and we hope that we will manage to take advantage of its lack of condition on an envelope function for future work.

[*]Let us recall briefly the definition of a $P^\star$-Donsker class (for more details, refer to van der Vaart 1998). Consider a class $\mathcal{H}$ of real-valued measurable functions on $\mathcal{Y}$ such that $\sup\{|h(y) - P^\star h| : h \in \mathcal{H}\}$ is finite for every $y \in \mathcal{Y}$. Then $\mathcal{H}$ is $P^\star$-Donsker if the sequence of processes $\{n^{1/2}(\mathbb{P}_n - P^\star)\, h : h \in \mathcal{H}\}$ converges in distribution to a tight limit process in the space $\ell^\infty(\mathcal{H})$ (informally speaking, the latter means that the behaviour of the process can be described, within a small error margin, by the behaviour of the marginal vectors). Then, the limit process $\mathbb{G}^\star$ is a Gaussian process with zero mean and covariance function

$$\mathrm{E}\,\mathbb{G}^\star h\, \mathbb{G}^\star h' = P^\star hh' - P^\star h\, P^\star h'.$$

6.4. Appendix

Let us recall the version of the projective limit theorem we apply below.

Theorem 6.4.1 (de Acosta). *Consider the projective system $\{\mathbb{R}^F, p_F^G\}$ where $F \subset G$ are finite subsets of $\mathcal{L}_\tau$, p_F^G is the restriction map from $\mathbb{R}^G$ to $\mathbb{R}^F$. Let p_F be the projection mapping $\mathcal{L}_\tau^*$ on $\mathbb{R}^F$ according to $p_F Q = \{Qf : f \in F\}$ (any $Q \in \mathcal{L}_\tau^*$). $\mathcal{L}_\tau^*$ is endowed with the initial topology induced by the maps p_F, i.e. the $\sigma(\mathcal{L}_\tau^*, \mathcal{L}_\tau)$-topology and with the σ- field that makes the projections p_F measurable.*

Let $\{r_n\}$ be a sequence of positive numbers that tends to infinity. Consider finally a sequence of probability measures $\{\mu_n\}$ such that, for any F finite subset of $\mathcal{L}_\tau$, the sequence $\{\mu_n \circ p_F^{-1}\}$ satisfies a Large Deviations Principle (LDP) with normalizing sequence $\{r_n\}$ and good rate function J_F.

Then $\{\mu_n\}$ also satisfies a LDP with normalizing sequence $\{r_n\}$ and good rate function J defined by

$$J(Q) = \sup\{J_F(p_F Q) : F \subset \mathcal{L}_\tau,\ F \text{ finite}\}. \tag{6.5}$$

Proof. (of Theorem 6.3.1*)* Let us consider the probability measures $\mu_n = \mathcal{L}(n\, b_n^{-1}(\mathbb{P}_n - P^\star))$, the normalizing sequence of term $r_n = n^{-1} {b_n}^2$ and choose a finite subset $F = \{f_1, \ldots, f_d\}$ of $\mathcal{L}_\tau$. Observe that

$$\mu_n \circ p_F^{-1} = \mathcal{L}\left(n\, b_n^{-1}\, p_F\left(\mathbb{P}_n - P^\star\right)\right) = \mathcal{L}\left(b_n^{-1}\, S_n\right), \text{ with } S_n = \sum_{i=1}^{n} \mathbf{f}(Y_i) - P^\star \mathbf{f},$$

where $\mathbf{f}$ is the column vector $(f_1, \ldots, f_d)^T$. In this finite dimensional setting, Theorem 3.7.1 of the reference book (Dembo and Zeitouni 1998) (for historical references, see Feller 1971; Petrov 1975) ensures that the sequence of the distributions $\mathcal{L}\left(b_n^{-1}\, S_n\right)$ satisfies a LDP with normalizing sequence $\{r_n\}$ and good rate function

$$J_F(z) = \sup\left\{x^T z - \frac{1}{2}\,\mathrm{E}\left(x^T Z\right)^2 : x \in \mathbb{R}^d\right\},$$

for $Z = \mathbf{f}(Y_1) - P^\star \mathbf{f}$. Since F above is arbitrary, $\{\mu_n\}$ does satisfy a LDP with normalizing sequence $\{r_n\}$ and good rate function J given by (6.5) in virtue of Theorem 6.4.1.

Besides, for any $x \in \mathbb{R}^d$, $g = x^T \mathbf{f} \in \mathcal{L}_\tau$ and $z = p_F Q$, we have both $x^T Z = g(Y_1) - P^\star g$ and $x^T z = Q\, g$, hence

$$J(Q) = \sup\{Q\, f - P^\star(f - P^\star f)^2/2 : f \in \mathcal{L}_\tau\}.$$

Particularly, this new expression clearly yields that $J(Q) = \infty$ whenever $Q\, \mathbb{1} \neq 0$: indeed, $J(Q)$ is bounded below by the supremum for $x \in \mathbb{R}$ of $x\, Q\, \mathbb{1}$.

Let us identify more precisely the rate function J. We shall prove that

- $J(Q) < \infty \Rightarrow Q \in L'_\tau$.

- If $J(Q) < \infty$ then $Q^s = 0$.

- Finally, $J(Q^a) = J(Q^a|P^\star)$.

• Take $Q \in \mathcal{L}_\tau^*$, $f \in \mathcal{L}_\tau$ such that $f = 0$ $P^\star$-ae and let x be any real number:

$$Q\, xf \leq x^2/2\, P^\star(f - P^\star f)^2 + J(Q) = J(Q),$$

hence $Q\, f = 0$ whenever $J(Q) < \infty$. In other words, if $J(Q)$ is finite then $Q \in L_\tau^*$. Now, define

$$\gamma(s) = \exp(s) - s - 1 \quad (s \in \mathbb{R})$$

and choose any $Q \in \mathcal{L}_\tau^*$ and $f \in \mathcal{L}_\tau$. The trivial inequality $\gamma(s) \geq s^2/2$ yields

$$Q\, f - P^\star\gamma(f) \leq Q\, f - P^\star f^2/2 \leq Q\, f - P^\star(f - P^\star f)^2/2.$$

Now, optimization in $f \in \mathcal{L}_\tau$ and $\gamma \leq \tau$ ensure that, for any $h \in L_\tau$ with $\|h\|_\tau = 1$,

$$|Q\, h| \leq J(Q) + P^\star\gamma(h) \leq J(Q) + P^\star\tau(h) = (J(Q) + 1)\, \|h\|_\tau$$

which is precisely the condition of continuity of Q as soon as $J(Q)$ is finite.

• Assume that $J(Q)$ is finite. Then $Q \in L_\tau'$ and $Q = Q^a + Q^s$ where Q^a and Q^s are respectively the absolute and the singular parts of Q. Let u, v be two elements of L_τ and $n \geq 1$ be an integer. We define $w_n^a = (-n \vee u \wedge n)\mathbb{1}\{|v| \leq n\} \in L_\tau$, $w_n^s = v\mathbb{1}\{|v| > n\} \in L_\tau$, $w_n = w_n^a + w_n^s$. The definition of $J(Q)$ and the boundedness of w_n^a yield

$$J(Q) \geq Q\, w_n - P^\star(w_n - P^\star w_n)^2/2 \geq Q\, w_n - P^\star {w_n}^2/2 = Q^a w_n - P^\star {w_n}^2/2 + Q\, w_n^s.$$

On the one hand, since w_n pointwisely increases to u, the monotone convergence theorem ensures that $Q^a w_n$ and $P^\star {w_n}^2$ respectively increase to $Q^a u$ and $P^\star u^2$. On the other hand, $Q\, w_n^s$ tends to $Q^s v$. Thus, letting n go to infinity, we get from the previous inequalities that

$$J(Q) \geq Q^a u - P^\star u^2/2 + Q^s v \text{ (any } u, v \in L_\tau).$$

We conclude the proof of the second point by emphasizing that the supremum of $Q^s v$ for v ranging over L_τ is finite if and only if $Q^s = 0$ (indeed, this is obviously sufficient and whenever $Q^s \neq 0$, there exists v such that $Q^s v > 0$ and $Q^s nv = n\, Q^s v$ increases to infinity as n also does).

• To conclude with, let $g \in L_{\tau^*}$ denote $\frac{\mathrm{d}Q^a}{\mathrm{d}P^\star}$ for some Q satisfying $J(Q) < \infty$. Remember that necessarily, $Q^a\, \mathbb{1} = P^\star g = 0$. Thus, for any $f \in L_\tau$, $Q^a f = P^\star fg = P^\star(f - P^\star f)g$. Now, the inequality $2\,(f - P^\star f)g \leq (f - P^\star f)^2 + g^2$ yields that $J(Q)$ is bounded above by $P^\star g^2/2$. Define finally $g_n = g\mathbb{1}\{|g| \leq n\} \in L_\tau$: $J(Q)$ is bounded below by $P^\star g^2 \mathbb{1}\{|g| \leq n\}/2$ which increases to $P^\star g^2/2$ as desired by monotone convergence: the last assertion is proved.

Lemma 4.1.5 of reduction of LPD in (Dembo and Zeitouni 1998) concludes the proof. □

7

Estimating the order of a model*

Résumé

Nous étudions dans ce chapitre certaines propriétés asymptotiques de deux estimateurs de l'ordre d'un modèle. L'ordre d'un modèle parmi une famille de modèles emboîtés est un nombre qui rend compte du degré de complexité du modèle relativement à la collection entière. C'est aussi l'ordre de toute loi qui appartient à ce modèle mais pas au plus grand sous-modèle qu'il contient. Nous présentons deux estimateurs pénalisés complémentaires. Leurs propriétés de consistance, leurs vitesses de sous- et sur-estimation sont l'objet de tous nos soins (ce sont les vitesses de décroissance des probabilités de sous- ou sur-estimer l'ordre, respectivement). Les résultats découlent d'une approche fonctionnelle : les preuves reposent sur l'application de résultats de Principes de Grandes et Moyennes Déviations pour la mesure empirique, sur une Loi du Logaritme Itéré ainsi que sur le lemme de Stein. Un exemple de tour à la Huber permet de raffiner l'ensemble des résultats en termes de contraintes sur la fonction de pénalité. Quelques exemples sont enfin étudiés soigneusement.

Abstract

We study in this chapter some asymptotic properties of an estimator of the order of a model. Given a collection of nested models and a distribution in their union, the order of the latter is the order of the smallest model the distribution belongs to. It is a number that quantifies the sophistication of the model regarding the whole collection. We define two complementary penalized estimators. We investigate their behaviours in terms of consistency and rates of under- and overestimation (*i.e.* when the estimated order is lower or larger than the true one, respectively). We have a linear functional approach of the problem. The proofs involve Large and Moderate Deviations Principles for the empirical measure, a bounded Law of the Iterated Logarithm and Stein's lemma. An example of à la Huber trick allows to enhance the results in terms of range of the penalty function. Some examples are carefully addressed.

*I would like to thank Pascal Massart who kindly drew our attention on the à la Huber trick.

Au menu

7.1. Introduction **209**
7.2. Presentation of the model and three examples **213**
7.2.1. The framework 213
7.2.2. Mixtures, abrupt changes and regressions 214
7.2.3. Earlier references for future comparisons 217
7.3. Two penalized maximum likelihood estimators **218**
7.4. Consistency **219**
7.4.1. Statement of the results 220
7.4.2. Proofs and more comments 223
7.5. Underestimation **228**
7.5.1. Stein's lemma for a lower bound on the rate 228
7.5.2. Upper bounds on the rate 230
7.5.3. Proofs and more comments, to be continued 233
7.5.4. Proofs, continued 237
7.5.5. Proofs, end 239
7.6. Overestimation **243**
7.6.1. Stein's lemma for a lower bound on the rate 243
7.6.2. Upper bounds on the rate 244
7.6.3. Proof 246
7.7. Back to the three examples **248**
7.7.1. The mixture of distributions 248
7.7.2. Abrupt changes and various regressions 253
7.8. Appendix **256**

7.1. Introduction

Problem at stake

Let $\{\Pi_K\}_{K\geq 1}$ be an increasing family of models, *i.e.* of collections of probability distributions. We assume that the models Π_K are parametric, *i.e.* that there exists an increasing family of parameter sets $\{\Theta_K\}_{K\geq 1}$ such that $\Pi_K = \{P_\theta : \theta \in \Theta_K\}$. K is the order of the model Π_K. It is also the order of any P_θ that belongs to $\Pi_K \setminus \Pi_{K-1}$. Informally, K quantifies the complexity of Π_K among the whole collection.

Let us mention some important examples which fit into this framework. They will be carefully addressed in this paper.

Mixture of distributions (MD): $\mathcal{D}$ is a class of densities γ_u indexed by a parameter $u \in \mathcal{U}$. F is a distribution on $\mathcal{U}$. The model Π_K is constituted of the probabilities P_θ whose densities with respect to (wrt) μ are written

$$p_\theta(y) = \int_{\mathcal{U}} \gamma_u(y) \mathrm{d}F_\theta(u),$$

where F_θ has finite support $\{u_1, \ldots, u_K\}$, $F_\theta(u_k) = \pi_k$ and $\theta = (\pi_1, \ldots, \pi_K) \times (u_1, \ldots, u_K)$.

Such models are used for clustering. There exist two families of methods that cope with the order estimation in this framework: informal graphical techniques and hypothesis testing procedures. Nonetheless, the latter are very difficult to handle since the mixing proportions of a given model lie on the boundary of the parameters set for a larger model. See below for further comments on that feature. For monographes on mixtures, see (Everitt and Hand 1981; Titterington et al. 1985; McLachlan and Basford 1988) and the paper (Lindsay and Lesperance 1995).

Abrupt changes in the mean (AC): $\mathcal{T}_K$ is a collection of partitions of the space $\mathcal{X} \subset \mathbb{R}^q$, $\mathcal{M}$ is a compact set of $\mathbb{R}$ and γ_0 is the Gaussian density on $\mathbb{R}$ with mean 0 and variance σ^2. The model Π_K is constituted of the probabilities P_θ whose densities wrt the Lebesgue measure $\mu^{\otimes q+1}$ are written (for any $(x, y) \in \mathcal{X} \times \mathbb{R}$)

$$p_\theta(x, y) = \gamma_0 \left(y - \sum_{k=1}^{K} m_k \mathbb{1}\{x \in \tau_k\} \right),$$

where $\tau = (\tau_k)_{1 \leq k \leq K} \in \mathcal{T}$, $m_k \in \mathcal{M}$ and $\theta = (\tau_1, \ldots, \tau_K) \times (m_1, \ldots, m_K)$.

Various regressions (VR): $\{t_k\}_{k \geq 1}$ is a free family of functions on $\mathcal{X} = [0, 1]$, $\mathcal{M}$ is a compact set of $\mathcal{Y} = \mathbb{R}$ and γ_0 is defined as in the previous example. The model Π_K is constituted of the probabilities P_θ whose densities wrt $\mu^{\otimes 2}$ are written (for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$)

$$p_\theta(x, y) = \gamma_0 \left(y - \sum_{k=1}^{K} m_k t_k(x) \right),$$

where $\theta = (m_1, \ldots, m_K)$.

The statistical problem we shall tackle is the following: given the families $\{\Pi_K\}_{K \geq 1}$ and $\{\Theta_K\}_{K \geq 1}$ and some observations $Y_1, \ldots, Y_n$ drawn along the unknown distribution $P^\star \in \Pi_{K^\star} \setminus \Pi_{K^\star - 1}$ (with convention $\Pi_0 = \emptyset$), we want to estimate $K^\star$, *i.e.* the order of $P^\star$.

Penalized empirical criteria

Our procedure of estimation involves the minimization of an empirical criterion $U_n(\theta)$ which is an abbreviation for $U_n(\theta\,; Y_1, \ldots, Y_n)$ where θ ranges over Θ_K and $K \geq 1$. It is well known since Mallows (1973) and Akaike (1974) published their seminal papers that such a procedure requires a penalization device. Indeed, large models are favoured wrt smaller ones since obviously, $\inf_{\theta \in \Theta_{K+1}} U_n(\theta)$ is always lower than $\inf_{\theta \in \Theta_K} U_n(\theta)$*. Regarding the empirical criterion as a bias term, this reflects that crude minimization of the criterion is equivalent to minimization of the bias, regardless of the subsequent increase of the variance. This is a typical feature of overfitting. The addition of a positive penalty term $\mathrm{pen}(n, K)$ to the criterion, where pen is an increasing function of K, heuristically simulates a variance term, so that the minimization takes into account both bias and variance. Now, informally, a correctly balanced penalization term yields good statistical properties for the corresponding estimator. The literature about penalization is abundant and it would be illusory to try to present here a comprehensive overview

*We assume that all the expressions involving extrema and empirical processes are measurable. It suffices for instance that each of them equals an extremum over a countable set, almost surely.

on that subject. We nonetheless refer the reader to (Barron et al. 1999; Massart 2000) for an insight on model selection *via* penalization with a view to risk bounds, and also to the collection of papers concerned with estimation of an order that we shall cite hereafter: indeed, all of them use the penalization device.

Some known results and methods

Let us briefly survey the history of order estimation. The selection is of course far from being comprehensive. We tried to pick up some important papers regarding the encompassed results and enhancements, methods and bibliographies. The interested reader should refer to the papers themselves for more precise overviews for each of them.

Henna (1985) considered the problem of estimation of the number of components in a mixture. His estimator was based on the minimization of empirical least squares. Note that standard maximum likelihood estimation was not tractable because the asymptotic behaviour of the maximal likelihood statistics in mixture models was not known yet. Ghosh and Sen (1985) and Self and Liang (1987) gave first answers under restrictive assumptions in order to avoid the difficulties due to lack of identifiability. The case of *overestimation* (*i.e.* when the estimated order K is larger than the true order $K^\star$) is in particular excluded. Senoussi (1990) studied a penalized quasi-likelihood estimator of the order in convex models. He proved the almost sure consistency *with a prior bound* on the true order $K^\star$. A comprehensive answer with a view to testing in the framework of mixtures was proposed as recently as 1999 by Dacunha-Castelle and Gassiat thanks to the so-called locally conic parameterization. However, Leroux (1992) showed that for its penalized maximum likelihood estimator of the order of a mixture, *underestimation* (*i.e.* when the estimated order K is lower than the true order $K^\star$) asymptotically almost never happened. Later results for maximum likelihood estimation were obtained by Keribin (2000) for mixtures and by (Haughton and Keribin 2001) for general regular families (extending earlier results of Haughton 1989 for exponential models). The two latter papers rely on the use of the locally conic parameterization of Dacunha-Castelle and Gassiat and on standard methods of expansion of the likelihood ratio. Both papers contain results of consistency as well as some rates of convergence of the probabilities of underestimation and overestimation *with a prior bound on* $K^\star$. Finally, Gassiat (2002) recently proved two general, simple yet powerful inequalities on likelihood ratios which allow *e.g.* to derive the consistency of a penalized maximum likelihood estimator of the number of populations (*i.e.* the order) of a mixture with Markov regime.

Some alternative methods have naturally been explored. For instance, in a framework of mixture models, Antoniadis and Berruyer (1986) have constructed a consistent estimator of the order of a mixture of one-dimensional exponential distributions. It relies on the theory of differential equations and involves Hankel matrices (more on this below). The authors proved the consistency of their estimator *without any prior bound on* $K^\star$. Later on, Dacunha-Castelle and Gassiat (1997) have investigated some statistical properties of a peculiar estimator of the order also based on Hankel matrices. They followed an original idea of Lindsay (1989), which differs sensibly from the earlier approach of Antoniadis and Berruyer. This estimator is derived from the minimization of a penalized empirical criterion $U_n(\theta)$ $(\theta \in \Theta_K)$ which estimates the determinant of the Hankel matrice of the K first moments of the unknown mixing distribution. Indeed, the true order is the first K that makes the determinant of the latter Hankel matrice zero. The authors proved the almost sure consistency *without any prior bound on* $K^\star$ as well as a nonasymptotic control of the probability of misestimating $K^\star$. Those results hold true under

mild assumptions. It is nevertheless necessary that there exist some consistent estimators of the moments of the mixing distribution. Finally, the results can even apply in other models and some examples are provided. Another example of alternative approach is given by (Guyon and Yao 1999) in a framework where dependence of the observations is allowed. The authors established nonasymptotic evaluations of the probabilities of underestimating and overestimating *with a prior bound on* $K^\star$. Their assumptions are satisfied in various models (including regression models with least squares estimation, time series and random fields) as soon as the empirical criterion $U_n(\theta)$ is *factorized*, *i.e.* whenever there exist a deterministic function U and a statistic T_n of the observations $Y_1, \ldots, Y_n$ such that $U_n(\theta) = U(\theta, T_n)$. We emphasize that this condition is not fulfilled for mixture models. Later again, James, Priebe, and Marchette (2001) used kernel density estimation together with an iterative scheme in order to estimate the order of a mixture model. They proved the consistency of their estimator *without any prior bound on* $K^\star$ and also explored some computational issues. Finally, we have studied in Chapter 5 a problem of detection of abrupt changes in random fields that involves the estimation of the order of a model. We proved that our estimator is weakly consistent.

There is also another competitive field of research concerned with order estimation into which our framework does not fit. It is the field of estimation of the order of a Hidden Markov Model. We refer the reader to (Gassiat and Boucheron 2001) for an overview. Besides, the approach in the latter paper and in the current one have much in common. From our point of view, this illustrates the generality of our approach. A description follows.

Forthcoming method and results

Let us present now our method and results. The main feature of our approach is that we identify the empirical measure $\mathbb{P}_n$ with a linear form on some set of functions that contains the log-likelihood of our models, *i.e.* the functions $\ell_\theta = \log p_\theta$ ($\theta \in \Theta_K$, $K \geq 1$). In the sequel, we shall use linear functional notations for expectations and integrals, *i.e.* write μf for the integral $\int f \mathrm{d}\mu$ of a function f with respect to a measure μ. Another important feature is that we do not use any property of the maximum likelihood estimators of the parameter $\theta^\star$ associated with the true distribution $P^\star$, though our estimators of the order are maximum likelihood estimators. Indeed, we investigate some *asymptotic* statistical properties of the two following estimators:

$$\widehat{K}_n^{\mathrm{L}} = \inf\left\{ K \geq 1 : \sup_{\theta\in\Theta_K} \mathbb{P}_n \ell_\theta - n^{-1}\mathrm{pen}(n,K) \geq \sup_{\theta\in\Theta_{K+1}} \mathbb{P}_n \ell_\theta - n^{-1}\mathrm{pen}(n,K+1) \right\},$$

$$\widehat{K}_n^{\mathrm{G}} = \arg\sup_{K\geq 1} \left\{ \sup_{\theta\in\Theta_K} \mathbb{P}_n \ell_\theta - n^{-1}\mathrm{pen}(n,K) \right\},$$

where pen is a positive penalty function. $\widehat{K}_n^{\mathrm{L}}$ is the first local maximizer of the criterion (hence the L in its name) and $\widehat{K}_n^{\mathrm{G}}$ is a global maximizer (hence the G). We study both $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$ because their respective qualities and deficiencies cast some light on their intrinsic nature.

We first consider some consistency properties: $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$ are consistent under mild assumptions, respectively *without and with a prior bound on* $K^\star$. When dealing with $\widehat{K}_n^{\mathrm{L}}$, a peculiar assumption is relevant. Informally, the latter strengthens the condition of inclusion of Π_K into Π_{K+1} by requiring that the infimum of the Kullback-Leibler divergence $H(P|P^\star)$ for P ranging over Π_{K+1} is *stricly* lower than the infimum of the same quantity taken over Π_K. This condition is useless for $\widehat{K}_n^{\mathrm{G}}$. The other important assumption required by the two of them ensures that

one can apply a bounded law of the iterated logarithm (see Dudley and Philipp 1983), which is the key of the proof of the almost never overestimation property.

Then, we focus on the underestimation probabilities. An application of Stein's lemma (see Bahadur et al. 1980) yields an optimal lower bound for any estimator of the order that does not almost surely overestimate the order. We derive upper bounds from a Sanov large deviations principle for the empirical measure $\mathbb{P}_n$. We actually use the generalized Sanov theorem by Léonard and Najim (2000), see Interlude 6, regarding $\mathbb{P}_n$ as a linear functional on the set of functions f which admit some exponential moments wrt $P^\star$, *i.e.* equivalently such that there exists $a > 0$ ensuring $P^\star \exp(a|f|) < \infty$.

Finally, we consider the overestimation probabilities. Stein's lemma is used again and implies that the rate of convergence is slower than exponential in n. We derive a rate of overestimation from a moderate deviations principle due to Wu (1994) (see Interlude 6).

All the results are improved in terms of the allowed range for the penalty functions pen thanks to an example of *à la Huber trick* (which is named after the author of Huber 1967). Heuristically, an à la Huber trick consists of rescaling in order to enhance performances of an empirical process. There exist numerous examples of applications of such a trick in Statistics and we shall give a few of them. The enhancement naturally requires more restrictive assumptions that we shall comment.

Organization of the chapter

Section 7.2 is dedicated to the presentation of the model. We also formally introduce the three examples we presented above. We define our estimators $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$ of the order in Section 7.3 and comment their definitions. Section 7.4 is devoted to the results of consistency, Section 7.5 to the results of underestimation and Section 7.6 to the results of overestimation. The study of the three examples is addressed carefully in Section 7.7. Finally, Section 7.8 encloses a postponed technical proof.

7.2. Presentation of the model and three examples

7.2.1. The framework

This chapter addresses the problem of the estimation of the order of the true model in an increasing family of models. Precisely, let $\{(\Theta_K, d_K)\}$ be an increasing sequence of metric compact sets equipped with a distance denoted d_K and indexed by their *order* $K \geq 1$. Since there is usually no confusion, we shall abbreviate in the sequel d_K to d. For any $K \geq 1$,

$$\Pi_K = \{P_\theta : \theta \in \Theta_K\} \subset \Pi_{K+1}$$

is the set of all the possible distributions with respect to (wrt) the model of order K on $(\mathcal{Y}, \mathcal{F})$. On the basis of some random observations, we aim at estimating the order of the model that contains the distribution they are drawn along. We shall also denote for convenience

$$\Theta_\infty = \bigcup_{K \geq 1} \Theta_K \quad \text{and} \quad \Pi_\infty = \bigcup_{K \geq 1} \Pi_K.$$

We suppose that some measure μ on $(\mathcal{Y}, \mathcal{F})$ dominates all the probability measures P_θ for any $\theta \in \Theta_\infty$. We denote p_θ the density of P_θ wrt μ and $\ell_\theta = \log p_\theta$ (convention $\log 0 = -\infty$).

Those functions will sometimes be considered as μ-almost everywhere (μ-ae) defined functions and sometimes as everywhere defined functions: we set their values equal to one and zero respectively on the μ-null sets of undefinition.

From now on, we assume that $\mathcal{Y}$ is Polish. Moreover, we require the following assumption (**Comp** stands for Compactness):

Comp The sets Π_K are compact for the weak topology on $M_1(\mathcal{Y})$.

Recall that the weak topology on the set $M_1(\mathcal{Y})$ of all probability measures on $\mathcal{Y}$ is the coarsest topology that renders continuous the functions $P \mapsto Pf$ for any $f \in \mathbb{R}^{\mathcal{Y}}$ continuous and bounded. This topology is metrizable. Here is a simple case:

Proposition 7.2.1. *Take $\mathcal{Y} = \mathbb{R}^q$. If the parameterization $\theta \mapsto p_\theta(y)$ is continuous for μ-almost all $y \in \mathcal{Y}$ and $(p_\theta - p_{\theta_0})$ is dominated in $L^1(\mu)$ uniformly in θ in a neighbourhood of θ_0 for any θ_0, then Assumption* **Comp** *holds.*

Proof. Since the weak topology is metrizable, it suffices to prove that any sequence $\{P_{\theta_q}\}$ in Π_K converges along some subsequence to some element of Π_K. One can suppose that $\theta_q \to \theta_0$ because Θ_K is a compact metric set. We conclude thanks to Lévy's continuity theorem and dominated convergence. □

7.2.2. Mixtures, abrupt changes and regressions

Let us introduce briefly three important examples. They will be carefully addressed further in Section 7.7. We emphasize that in examples **AC** and **VR**, $\mathcal{X} \times \mathcal{Y}$ is substituted to $\mathcal{Y}$.

Mixture of distributions example (MD) Let $\mathcal{Y} = \mathbb{R}$ and $\mathcal{D} = \{\gamma_u : u \in \mathcal{U}\}$ be a set of real densities γ_u (wrt the Lebesgue measure μ) with parameter $u \in \mathcal{U}$ compact set of $\mathbb{R}^q$. For some mixing distribution F on $\mathcal{U}$, one can define the distribution with density p wrt μ

$$p(y) = \int_{\mathcal{U}} \gamma_u(y) \mathrm{d}F(u) \quad (\text{any } y \in \mathcal{Y}).$$

Such a distribution is a mixture of distributions. A finite mixture is a mixture whose mixing distribution has finite support.

We assume that the set $\mathcal{D}$ satisfies the condition of identifiability of the mixture

$$\forall\, y \in \mathcal{Y},\ \int_{\mathcal{U}} \gamma_u(y) \mathrm{d}F_1(u) = \int_{\mathcal{U}} \gamma_u(y) \mathrm{d}F_2(u) \Longrightarrow F_1 = F_2.$$

We shall focus on three examples of class $\mathcal{D}$ for which the latter condition holds true, namely

- **Mixture of Gaussian distributions in the mean (MGM):** the class $\mathcal{D}$ of all real Gaussian densities with mean $m \in \mathcal{M} = \mathcal{U}$ and same variance σ^2;
- **Mixture of Gaussian distributions in mean and variance (RKO):** the class $\mathcal{D}$ of all real Gaussian densities with mean m and variance σ^2, $(m, \sigma^2) \in \mathcal{M} \times \mathcal{S} = \mathcal{U}$ ($\mathcal{S} \subset \mathbb{R}_+^*$);

- **Mixture of exponential distributions*** **(ME):** the class $\mathcal{D}$ of all exponential densities with mean $m \in \mathcal{M}$.

We define $\Pi_1 = \mathcal{D}$ and for any $K \geq 2$, the set Π_K of finite mixtures of order K

$$\Pi_K = \left\{ p_\theta \, \mathrm{d}\mu = \sum_{k=1}^{K-1} \pi_k \gamma_{u_k} + \left(1 - \sum_{k=1}^{K-1} \pi_k \right) \gamma_{u_K} \, \mathrm{d}\mu : \theta = (\boldsymbol{\pi}, \boldsymbol{u}) \in \Theta_K \right\},$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{K-1})$ is a vector of nonnegative numbers satisfying $\sum_k \pi_k \leq 1$ and $\boldsymbol{u}$ is a vector of parameters with coordinates in $\mathcal{U}$.

Here, the order K of the model Π_K is the maximum number of elementary densities melted in any $p_\theta \in \Pi_K$.

Note that for the examples **MGM**, **RKO** and **ME**, Assumption **Comp** holds true thanks to Proposition 7.2.1. Besides, those settings correspond to the following mixture of populations models, where the hidden random variable (rv) X takes its values in an unknown finite subset $\{u_1^\star, \ldots, u_{K^\star}^\star\}$ of $\mathcal{U}$ (we do not know $K^\star$ either), with unknown distribution $\mathrm{P}(X = u_j^\star) = \pi_j^\star$ $(j = 1, \ldots, K^\star)$:

MGM

$$Y = X + e.$$

Here, e is drawn along a centered Gaussian distribution of variance σ^2 independently of X.

RKO

$$Y = X_1 + X_2 \, e.$$

Here, e is drawn along a centered Gaussian distribution of variance 1 independently of $X = (X_1, X_2)$.

ME

$$Y = X \, e.$$

Here, e is drawn along an exponential distribution with mean 1, independently of X.

Indeed, Y has then density $p_{\theta^\star}$ for $\theta^\star = (\boldsymbol{\pi}^\star, \mathbf{u}^\star)$, with $\boldsymbol{\pi}^\star = (\pi_1^\star, \ldots, \pi_{K^\star-1}^\star)$ and $\mathbf{u}^\star = (u_1^\star, \ldots, u_{K^\star}^\star)$.

Abrupt changes in the mean example (AC) Let $(\mathcal{X}, \mathcal{B}, P)$ be an open subset of $\mathbb{R}^q$ equipped with the trace of the Borel σ-field and a probability measure P dominated by the Lebesgue measure $\mu^{\otimes q}$, whose density wrt $\mu^{\otimes q}$ is denoted p. Let also $\mathcal{Y}$ be a subset of $\mathbb{R}$ and ζ be a positive constant. Let $\mathcal{T}$ be the set of all finite Caccioppoli partitions of $\mathcal{X}$ whose perimeters are uniformly bounded by ζ.

*For an example with real data, refer to Chapter 1, Section 1.2: we adjust there a mixture of exponential distributions for duration of mobile phone calls.

Remark 7.2.2 (Caccioppoli partitions). For the definition and the properties of those partitions, refer to Appendix B. We introduce this sophisticated setting because it allows the construction of a metric on $\mathcal{T}$ for which $\mathcal{T}$ is a compact metric set.

A finite partition $\tau \in \mathcal{T}$ is *roughly* a finite collection $(\tau_k)_{1\leq k\leq K}$ of subsets τ_k of $\mathcal{X}$ such that $P(\mathcal{X} \setminus \cup_k \tau_k) = 0$, $P(\tau_k) > 0$ and $P(\tau_k \cap \tau_{k'}) = 0$ for any $k, k' \neq k$. The *cardinality* of $\tau = (\tau_k)_{1\leq k\leq K}$ equals K. We denote $\mathcal{T}_K$ the set of all partitions of cardinality K and

$$\mathcal{T}_{\leq K} = \bigcup_{k=1}^{K} \mathcal{T}_k.$$

As in Chapter 5, two partitions $\tau = (\tau_k)_{1\leq k\leq K}$ and $(\tau'_k)_{1\leq k\leq K'}$ are equal as soon as $K = K'$ and $\tau_k = \tau'_k$ for any $1 \leq k \leq K$ up to a reindexing of, say, τ'. One can finally define a metric d_P on $\mathcal{T}$ such that

- $(\mathcal{T}, d_P)$ is a compact metric set, and
- $\mathcal{T}_{\leq K}$ is closed in $\mathcal{T}$ equipped with the topology induced by d_P.

Choose now $\mathcal{M} \subset \mathbb{R}$ a compact set and define, for any $K \geq 1$, any $\tau \in \mathcal{T}_K$, any $\boldsymbol{m} = (m_1, \ldots, m_K) \in \mathcal{M}^K$,

$$\begin{aligned}
&\theta = (\tau, \boldsymbol{m}),\\
&f_\theta(x) = \sum_{k=1}^{K} m_k \mathbb{1}\{x \in \tau_k\} \quad (\text{any } x \in \mathcal{X}) \quad \text{and}\\
&p_\theta(x, y) = \gamma(y\,; f_\theta(x))\, p(x) \quad (\text{any } x \in \mathcal{X},\ y \in \mathcal{Y})
\end{aligned}$$

where $\gamma(\cdot\,; m)$ is a real density wrt μ with mean m. Then, for any $K \geq 1$,

$$\begin{aligned}
\Pi_1 &= \left\{p_\theta \,\mathrm{d}\mu^{\otimes q+1} : \theta \in \Theta_1\right\} \quad \text{with} \quad \Theta_1 = \{\mathcal{X}\} \times \mathcal{M}\\
\Pi_{K+1} &= \Pi_K \cup \left\{p_\theta \,\mathrm{d}\mu^{\otimes q+1} : \theta \in \Theta_{K+1}\right\} \quad \text{with} \quad \Theta_{K+1} = \Theta_K \cup \left(\mathcal{T}_{K+1} \times \mathcal{M}^{K+1}\right).
\end{aligned}$$

Here, the order K is the largest cardinality of the involved partitions in Π_K. Besides, Assumption **Comp** is satisfied in virtue of the compactness of $\mathcal{T}_{\leq K}$ and $\mathcal{M}$ as soon as γ allows *e.g.* application of Proposition 7.2.1, for instance γ continuous such that $\gamma(\cdot\,; m)$ is bounded above by an integrable function independently of $m \in \mathcal{M}$. The Gaussian kernel $\gamma(y\,; m) = \gamma_m(y)$ (with the notations introduced in the former **MGM** example) provides such a case, and we choose it from now on.

This setting corresponds to the abrupt changes in the mean model, where both X and Y are observed:

$$Y = f^\star(X) + e, \tag{7.1}$$

for some function $f^\star(x) = \sum_j m_j^\star \mathbb{1}\{x \in \tau_j^\star\}$ whose underlying partition $\tau^\star$ belongs to $\mathcal{T}_{K^\star}$ with unknown $K^\star$; for X drawn along $p\,\mathrm{d}\mu^{\otimes q}$; and for e independent of X with distribution $\gamma(\cdot\,; 0)\,\mathrm{d}\mu$, *i.e.* centered Gaussian distribution with variance σ^2. Hence, (X, Y) has density $p_{\theta^\star}$ for $\theta^\star = (\tau^\star, \mathbf{m}^\star)$ (where $\mathbf{m}^\star = (m_1^\star, \ldots, m_{K^\star}^\star)$).

Various regression examples (VR) Let $\mathcal{X} = [0,1]$ and $\mathcal{Y} = \mathbb{R}$ be equipped with the Lebesgue measure μ on Borel sets. Let $\{t_K\}_{K\geq 1}$ be a free system of continuous functions on $\mathcal{X}$ and $\mathcal{U}$ be a compact set of $\mathbb{R}$ that contains 0. We assume that the family $\{t_K(\mathcal{X})\}_{K\geq 1}$ is bounded. We define $\Theta_K = \mathcal{U}^K$ and for any $\theta \in \Theta_K$, $(x,y) \in \mathcal{X} \times \mathcal{Y}$,

$$f_\theta(x) = \sum_{k=1}^{K} \theta_k t_k(x) \quad \text{and}$$
$$p_\theta(x,y) = \gamma(y; f_\theta(x)),$$

where $\gamma(\cdot\,; m)$ is the Gaussian density with mean m and variance σ^2 wrt μ. Here, K is the maximal number of base functions t_k involved in the definition of f_θ ($\theta \in \Theta_K$). Observe that Assumption **Comp** is satisfied thanks to Proposition 7.2.1.

This setting corresponds to the regression model

$$Y = f^\star(X) + e \tag{7.2}$$

where X is uniformly distributed on $[0,1]$ independently of e which is drawn along a centered, variance σ^2 Gaussian distribution, and $f^\star = \sum_{k=1}^{K^\star} \theta_k^\star t_k$ for some unknown $K^\star$, with the additional condition $\theta_{K^\star} \neq 0$. Indeed, (X,Y) then has $p_{\theta^\star}$ for density wrt μ.

7.2.3. Earlier references for future comparisons

We shall compare our forthcoming results with earlier ones already introduced in Section 7.1. The same remark naturally still holds true: our selection is far from being exhaustive. The papers for comparison are chosen regarding the encompassed results and enhancements, methods and bibliographies. The interested reader should refer to the papers themselves for more precise overviews for each of them.

The estimation of the order of a mixture has attracted a lot of attention from the mid-eighties. The problem is known to be difficult because estimating the parameters of a mixture is quite hard, even when the true order is known. Henna (1985) first considered this problem with an empirical least squares approach. Later, Leroux (1992) used a log-likelihood procedure. In their 1997 paper, Dacunha-Castelle and Gassiat used thoroughly the fact that it is possible to know the cardinality of the support of a finite mixing distribution using Hankel moment matrices. Recently, Keribin (2000) improved the results of Leroux in the same framework of mixture models with maximum likelihood approach, and James et al. (2001) explored a semiparametric method based on kernel density estimation and Kullback-Leibler distance.

The results of (Haughton 1988) in a model selection *via* maximum likelihood procedure from an exponential family apply to the order estimation problem. Haughton and Keribin (2001) generalize those results to regular families.

Guyon and Yao (1999) study a general class of order selection criteria under a peculiar assumption of factorization to be discussed later. Dependence of the observation is allowed and models such as regression models, time series and random fields are included, but not mixtures. Their results particularly apply to the order estimation problem.

Finally, we shall also evoke (Gassiat and Boucheron 2001), although our results do not apply in their particular framework of estimation of the order of a Hidden Markov Model. Indeed, it is a competitive problem of order estimation. Besides, a few resemblances are worth some words of comparison. For earlier references on Markov order estimation, see the bibliography in the paper.

7.3. Two penalized maximum likelihood estimators

We set here as usual a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ upon which all the rv will be defined in the sequel of this chapter. We denote E the expectation wrt P.

We observe n rv $Y_1, \ldots, Y_n$ taking their values in $\mathcal{Y}$. We assume that $Y_1, \ldots, Y_n$ are *independent and identically distributed.* Their common distribution will be taken in Π_∞ (except temporarily in Section 7.4) and will generally be the one defined below:

$$P_{\theta^\star} = P^\star \in \Pi_{K^\star}. \tag{7.3}$$

We require that $\Pi_{K^\star}$ is here the smallest of the Π_K's containing $P^\star$, or equivalently that $P^\star \in \Pi_{K^\star} \setminus \Pi_{K^\star-1}$. Whenever there might be confusion (particularly for change of probability argument), we will denote P_{P_θ} (respectively E_{P_θ}) for P (respectively E) to emphasize that the rv Y_i's are P_θ-distributed (*i.e.* $\mathrm{P}_{P_\theta}(Y_1 \in F) = P_\theta(F)$ and $\mathrm{P}(Y_1 \in F) = \mathrm{P}_{P^\star}(Y_1 \in F) = P^\star(F)$, any $F \in \mathcal{F}$). We focus our attention on the estimation of the order of the true model, *i.e.* on $K^\star$ when the true distribution is $P^\star$, which is by convention $K^\star = \infty$ when $P^\star \notin \Pi_\infty$. $\mathbb{P}_n$ denotes the empirical mean of $(Y_1, \ldots, Y_n)$ and

$$\ell_n(\theta) = \sum_{i=1}^{n} \log p_\theta(Y_i) = n\mathbb{P}_n \ell_\theta$$

is the log-likelihood of the observations.

In the sequel, we will study some properties (consistency, rates of underestimation and overestimation) of two *penalized maximum likelihood estimators* of $K^\star$ (when the true distribution is $P^\star$), respectively defined by:

penalized local maximum likelihood: the first local maximizer of the penalized log-likelihood of the observations, *i.e.*

$$\widehat{K}_n^{\mathrm{L}} = \inf\left\{ K \geq 1 : \sup_{\theta \in \Theta_K} \ell_n(\theta) - \mathrm{pen}(n,K) \geq \sup_{\theta \in \Theta_{K+1}} \ell_n(\theta) - \mathrm{pen}(n,K+1) \right\};$$

penalized global maximum likelihood: the global maximiser of the penalized log-likelihood of the observations, *i.e.*

$$\widehat{K}_n^{\mathrm{G}} = \arg\sup_{K \geq 1} \left\{ \sup_{\theta \in \Theta_K} \ell_n(\theta) - \mathrm{pen}(n,K) \right\},$$

which is always bounded below by $\widehat{K}_n^{\mathrm{L}}$.

For the latter estimator, most of the forthcoming results will require a prior bound $K_{\max}$ on $K^\star$.

The positive penalty function pen must satisfy (**P** stands for Penalty)

P0 On the one hand, $\mathrm{pen}(n,\cdot)$ is an increasing function for any $n \geq 1$. On the other hand, $\mathrm{pen}(n,K) \to \infty$ as $n \to \infty$ and $\mathrm{pen}(n,K) = o(n)$ for any $K \geq 1$.

The typical form of the function pen is $\mathrm{pen}(n,K) = v_n D(K)$, with $D \in \mathbb{R}^{\mathbb{N}}$ increasing and $\{v_n\}$ such that both v_n increases to ∞ and $v_n = o(n)$.

Remark 7.3.1 (penalization). The penalization device has become widely popular since Mallows (1973) and Akaike (1974) first introduced it for purpose of model selection. Its aim is to penalize the models with large complexity (*i.e.* in the present case, with large order K) by adding a positive term to the empirical contrast criterion to minimize (here, minus the log-likelihood) in M-estimation procedures. Indeed, from a naive point of view, large models tend to be favoured in terms of M-estimation wrt smaller ones, as in the present situation, since one always has $\sup\{\ell_n(\theta) : \theta \in \Theta_K\} \leq \sup\{\ell_n(\theta) : \theta \in \Theta_{K+1}\}$. In other words, as naive as the former, if one sees the empirical contrast criterion term as a *bias term*, then large models allow better minimization of the bias, regardless of the subsequent increase of the variance (this formulation is justified *e.g.* by some simple considerations on regression models, see Chapter 3, Section 3.2.1). Now, the penalization device can be seen as a *simulated variance term*. The heuristic is that, when correctly balanced, this added term yields better results, *i.e.* that simultaneous minimization of both bias and variance performs better than minimization of the sole bias. For a general work on model selection *via* penalization, see (Barron et al. 1999).

All the papers we introduced earlier with a view to future comparisons in Section 7.2.3 use the penalization device.

Remark 7.3.2 (local *vs* global maximizations).
Local rather than global maximization yields both interesting algorithmic and technical simplifications we also take advantage of hereafter (see Sections 7.4, 7.6).

Informally, the technical simplification relies on the fact that the overestimation event is simply included in a single event which only involves the models of order $K^\star$ and $K^\star + 1$ (see *e.g.* the proof of Proposition 7.4.13 for more details). We particularly show that $\widehat{K}_n^{\mathrm{L}}$ is consistent without any prior bound on $K^\star$, whereas such an assumption is needed when considering $\widehat{K}_n^{\mathrm{G}}$. Nevertheless, another restrictive assumption is required for the latter result of consistency of $\widehat{K}_n^{\mathrm{L}}$, which strenghtenes the original assumption of inclusion $\Pi_K \subset \Pi_{K+1}$. Besides, the obtained rates of underestimation are poorer for $\widehat{K}_n^{\mathrm{L}}$ than for $\widehat{K}_n^{\mathrm{G}}$.

Among the references presented in Section 7.2.3, all but (Henna 1985; James et al. 2001) have estimators whose definitions rely on global (rather than local) maximization of an empirical criterion.

7.4. Consistency

We state and prove in this section three results about our two estimators $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$. Their combination casts good light on the behaviour of the estimators in terms of consistency. We show, under appropriate assumptions, that: if the distribution $P^\star$ does not belong to Π_∞, then $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$ both explode to infinity, almost surely — see Theorem 7.4.1; if $P^\star$ belongs to $\Pi_{K^\star}$, then $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$ both equal $K^\star$ for a number n of observations large enough, almost surely — see Theorem 7.4.3 and its enhancement Theorem 7.4.5. The latter improvement relies on slightly heavier assumptions and mostly on a technical subtlety (an example of *à la Huber trick*) we shall address shortly.

7.4.1. Statement of the results

Two results of consistency

Let us first state some assumptions (**C** stands for Consistency). We shall temporarily denote $p^\star$ and $\ell^\star$ the density and the log-density of $P^\star << \mu$ (forgetting the index $\theta^\star$) in order to encompass the case of $P^\star \notin \Pi_\infty$. Finally, in the whole sequel, $H(P|\Pi)$ (respectively $H(\Pi|Q)$) will denote the infimum $\inf\{H(P|Q) : Q \in \Pi\}$ (respectively $\inf\{H(P|Q) : P \in \Pi\}$) of the Kullback-Leibler divergence $H(P|Q)$ of P wrt Q, where Q ranges over Π (respectively P ranges over Π).

C1 $H(P^\star|\Pi_1)$ is finite.

C2 If $P^\star \notin \Pi_K$, then $H(P^\star|\Pi_{K+1}) < H(P^\star|\Pi_K)$.

C3 There exist $l, u \in \mathbb{R}^{\mathcal{Y}}$ such that $(u-l) \in L^1(P^\star)$ with

$$l \leq \ell^\star \leq u \quad \text{and} \quad l \leq \ell_\theta \leq u \quad (\text{any } \theta \in \Theta_\infty).$$

C4 The parameterization $\theta \mapsto \ell_\theta(y)$ from Θ_K to $\mathbb{R}$ is continuous for any $y \in \mathcal{Y}$.

Theorem 7.4.1 (explosion). *Assume that Assumptions* **Com**, **P0** *and* **C2**–**C4** *hold true. Suppose moreover that $P^\star$ does not belong to Π_∞. Then*

$$\mathrm{P}\left(\liminf_{n\to\infty} \widehat{K}_n^{\mathrm{L}} = \infty\right) = \mathrm{P}\left(\liminf_{n\to\infty} \widehat{K}_n^{\mathrm{G}} = \infty\right) = 1.$$

Remark 7.4.2.

- The assumptions of the theorem are fulfilled for the **MGM**, **RKO**, **ME** and **AC** examples, see Section 7.7.
- The theorem applies as soon as $P^\star$ does not belong to Π_∞. A typical example where $P^\star \notin \Pi_\infty$ is when the family of models described by the collection $\{\Pi_K\}_{K\geq 1}$ is not sophisticated enough, though possibly *almost*. More precisely, when considering each example introduced in Section 7.2.2:

MD the support of the mixing distribution $F^\star$ on $\mathcal{M}$ is not finite.

AC the function $f^\star$ in (7.1) is not piecewise constant, though the density $p^\star$ of $P^\star$ wrt $\mu^{\otimes q+1}$ is still written

$$p^\star(x,y) = \gamma(y; f^\star(x))p(x) \quad (\text{any } (x,y) \in \mathcal{X}\times\mathcal{Y}).$$

VR the regression function $f^\star$ in (7.2) satisfies

$$f^\star(x) = \sum_{k\geq 0} \theta_k^\star t_k \quad (\text{any } x \in \mathcal{X})$$

for some sequence $\{\theta_k^\star\}$ with an infinite number of nonzero terms.

Thus, Theorem 7.4.1 establishes that in such cases, the estimators $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$ behave as we expect, *i.e.* almost surely, $\widehat{K}_n^{\mathrm{L}} = \infty$ and $\widehat{K}_n^{\mathrm{G}} = \infty$ for a number n of observations large enough.

Let us introduce now some notations in order to ease the statement of the next assumption. From now on, Log will denote the truncated log, *i.e.* $\mathrm{Log}(x) = \log(x \vee e)$ $(x \in \mathbb{R})$. The function φ will also be defined by $\varphi(x) = x^2/\mathrm{Log}\,\mathrm{Log}(x)$ $(x \in \mathbb{R})$.

C5a Let $K > K^\star$ be an integer to be chosen later. The class of functions

$$\mathcal{G} = \left\{ g_\theta = (\ell_\theta - \ell^\star) : \theta \in \Theta_K \right\}$$

is $P^\star$-Donsker and $\varphi(u - l) \in L^1(P^\star)$.

P1a The penalty function satisfies, for any $K \geq 1$,

$$\liminf_{n\to\infty} \frac{\mathrm{pen}(n, K+1)}{\mathrm{pen}(n, K)} > 1 \quad \text{and} \quad \limsup_{n\to\infty} \frac{(n \log\log n)^{1/2}}{\mathrm{pen}(n, K)} = 0.$$

Theorem 7.4.3 (consistency). *Assume that Assumptions* **Com**, **P0–P1a**, **C1** *and* **C3–C4** *hold true. Suppose moreover that $P^\star$ does belong to Π_∞, precisely $P^\star \in \Pi_{K^\star} \setminus \Pi_{K^\star-1}$.*
• *If Assumption* **C5a** *is fulfilled for $K = K^\star + 1$,* if Assumption **C2** is satisfied, *then* without any prior bound on $K^\star$,

$$\mathrm{P}\left(\lim_{n\to\infty} \widehat{K}_n^{\mathrm{L}} = K^\star \right) = 1.$$

• *Furthermore,* if we know a prior bound $K_{\max}$ on $K^\star$, *if moreover Assumption* **C5a** *holds true for $K = K_{\max}$, then* without invoking Assumption **C2**,

$$\mathrm{P}\left(\lim_{n\to\infty} \widehat{K}_n^{\mathrm{G}} = K^\star \right) = 1,$$

Remark 7.4.4.
• The assumptions of the theorem are fulfilled in the **MGM** example for both $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$ and they are satisfied in the **VR** example for $\widehat{K}_n^{\mathrm{G}}$, see Section 7.7.
• This theorem states the consistency of $\widehat{K}_n^{\mathrm{G}}$ and $\widehat{K}_n^{\mathrm{L}}$ when the true model belongs to the family of models $\{\Pi_K\}_{K\geq 1}$. The consistency of $\widehat{K}_n^{\mathrm{L}}$ holds true without any prior bound on $K^\star$, but with an additional condition which requires that $H(P^\star|\Pi_K)$ decreases as K increases. On the contrary, the consistency of $\widehat{K}_n^{\mathrm{G}}$ holds true without any condition on the behaviour of $H(P^\star|\Pi_K)$ when K increases, but with an additional condition of known prior bound on $K^\star$. From that point of view, the two results are complementary: one informally prefers to use $\widehat{K}_n^{\mathrm{L}}$ for its algorithmic simplicity and lack of condition on $K^\star$, but one can always rather use $\widehat{K}_n^{\mathrm{G}}$ when the additional condition on $H(P^\star|\Pi_K)$ is not fulfilled. See Section 7.7 for concrete examples.

The proof of Theorem 7.4.3 relies on the application of a functional Law of the iterated logarithm (bounded LIL) whose validity is guaranteed by Assumption **C5a**. Furthermore, the other assumptions are very weak, including Assumption **C1**, which could have equivalently been

There exists (a minimal) $K_{\min} \geq 1$ such that $H(P^\star|\Pi_{K_{\min}})$ is finite;

we would have then studied the whole problem for $K \geq K_{\min}$, or $K' \geq 1$ up to a reindexing.
• Let us now review the results of consistency included in the references of Section 7.2.3. Henna (1985) proves the almost sure consistency of his estimator *without any prior bound on the true order.* His proof uses thoroughly the local nature of his estimator, according to the terminology of

Remark 7.3.2. Leroux (1992) proves the almost sure non-underestimation of his penalized global maximum likelihood estimator. Keribin (2000) proves that almost sure non-overestimation also holds true for the latter, *with a prior bound on* $K^\star$. Besides, her scheme of proof is alike ours in the $\widehat{K}_n^{\mathrm{G}}$ case.

Dacunha-Castelle and Gassiat (1997) prove that their order estimator is almost surely consistent *without any prior bound for* $K^\star$, though their method is global according to the terminology of Remark 7.3.2. This is true even when the family of melted distributions $\mathcal{D}$ is not dominated by some measure (a case where maximum likelihood estimation is impossible). We nevertheless emphasize that this result relies on some consistent estimation of the moments of the mixing distribution. Furthermore, their method has numerical interest even though global maximization requires calculus for any order: indeed, no computation of the mixing parameters is needed, but estimation of moments of the mixing distribution. However, the latter numerical interest is tempered by the difficult algorithmic task of computing determinants of large matrices. A solution might consist of combining those results with the earlier work of Antoniadis and Berruyer (1986), who carefully addressed the algorithmic issues. James et al. (2001) show that their local estimator is almost surely consistent *without any prior bound* either.

Finally, almost sure consistency *with a prior bound* is obtained in (Guyon and Yao 1999; Haughton and Keribin 2001) for their global estimators, whereas almost sure consistency is proved *without any* for the global estimators of (Gassiat and Boucheron 2001).

Refined consistency

In the whole sequel, we shall denote $H(\theta) = H(P^\star | P_\theta)$ (any $\theta \in \Theta_\infty$).

C5b Let $K > K^\star$ be an integer to be chosen later. The class of functions

$$\mathcal{G} = \left\{ g_\theta = \frac{\ell_\theta - \ell^\star}{H(\theta)^{1/2}} : \theta \in \Theta_K,\ H(\theta) > 0 \right\}$$

is $P^\star$-Donsker. Moreover, there exists an envelope function G for the class $\mathcal{G}$ such that $\varphi(G) \in L^1(P^\star)$ (we introduced earlier the function $\varphi(x) = x^2/\mathrm{LogLog}(x)$, $x \in \mathbb{R}$).

P1b The penalty function also satisfies, for any $K \geq 1$,

$$\liminf_{n\to\infty} \frac{\mathrm{pen}(n, K+1)}{\mathrm{pen}(n, K)} > 1 \quad \text{and} \quad \limsup_{n\to\infty} \frac{\log\log n}{\mathrm{pen}(n, K)} = 0.$$

Theorem 7.4.5 (refined consistency). *Suppose that Assumptions* **Com**, **P0** *and* **P1b**, **C1** *and* **C3–C4** *hold true. Suppose moreover that* $P^\star$ *belongs to* Π_∞, *precisely* $P^\star \in \Pi_{K^\star} \setminus \Pi_{K^\star-1}$.
• *If Assumption* **C5b** *is fulfilled for* $K = K^\star + 1$, if Assumption **C2** is satisfied, *then* without any prior bound on $K^\star$,

$$\mathrm{P}\left(\lim_{n\to\infty} \widehat{K}_n^{\mathrm{L}} = K^\star \right) = 1.$$

• *Besides,* if we know a prior bound $K_{\max}$ on $K^\star$, *if moreover Assumption* **C5b** *holds true for* $K = K_{\max}$, *then* without invoking Assumption **C2**,

$$\mathrm{P}\left(\lim_{n\to\infty} \widehat{K}_n^{\mathrm{G}} = K^\star \right) = 1.$$

Remark 7.4.6 (on Theorem 7.4.5).
• The assumptions of the theorem hold true in the **MGM** example, see Section 7.7.
• We have improved here the previous consistency result of Theorem 7.4.3 in terms of range of the admissible penalty functions, which is dramatically extended here, compare Assumption **P1a** with Assumption **P1b**. This extension naturally has its cost, as the comparison of Assumptions **C5a** and **C5b** reveals. It is nevertheless worth noting that this impressive enhancement is mostly due to the application of a so-called *à la Huber trick* we shall present hereafter in Section 7.4.2.

7.4.2. Proofs and more comments

The two estimators almost surely do not underestimate

We shall begin with a lemma:

Lemma 7.4.7. *Whenever Assumptions* **Com** *and* **C3**–**C4** *hold true, then almost surely, for any* $K \geq 1$,

$$\sup_{\theta \in \Theta_K} \mathbb{P}_n(\ell_\theta - \ell^\star) \underset{n\to\infty}{\longrightarrow} -H(P^\star|\Pi_K).$$

Proof. Since $H(P^\star|\cdot)$ is lower semicontinuous wrt the weak topology on $M_1(\mathcal{Y})$ and Π_K is compact for the same topology, there exists $\theta_0 \in \Pi_K$ such that $H(P^\star|\Pi_K) = H(P^\star|P_{\theta_0}) = P^\star(\ell^\star - \ell_{\theta_0})$.

Now, Assumption **C3** yields that $\ell_{\theta_0} - \ell^\star \in L^1(P^\star)$ and forwardly, the Strong law of large numbers ensures that almost surely,

$$\liminf_{n\to\infty} \sup_{\theta\in\Theta_K} \mathbb{P}_n(\ell_\theta - \ell^\star) \geq \liminf_{n\to\infty} \mathbb{P}_n(\ell_{\theta_0} - \ell^\star) \geq P^\star(\ell_{\theta_0} - \ell^\star).$$

Besides, introducing the modulus of continuity $\omega_h(y) = \sup\{|(\ell_\theta - \ell_{\theta'})(y)| : d(\theta,\theta') \leq h\}$ (any $y \in \mathcal{Y}$), one observes that ω_h is uniformly (in h) dominated in $L^1(P^\star)$ by $(u-l)$ in virtue of Assumption **C3**. Set $h > 0$ and invoke the Borel-Lebesgue property and Assumption **Com** in order to cover the compact set Θ_K by $\cup_{j=1}^r B_d(\theta_j, h)$. Then, for any $1 \leq i \leq n$, one has for some j: $\ell_\theta(X_i) \leq \ell_{\theta_j}(X_i) + \omega_h(X_i)$, hence applying $\mathbb{P}_n$ on both sides of the inequality (it acts as a nonnegative linear form), taking the supremum in j first, then in θ and finally adding $-\mathbb{P}_n\ell^\star$ on both sides, we get that

$$\sup_{\theta\in\Theta_K} \mathbb{P}_n(\ell_\theta - \ell^\star) \leq \max_{1\leq j\leq r} \mathbb{P}_n(\ell_{\theta_j} - \ell^\star) + \mathbb{P}_n\omega_h.$$

Then the Strong law of large numbers yields, almost surely

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta_K} \mathbb{P}_n(\ell_\theta - \ell^\star) \leq \max_{1\leq j\leq r} P^\star(\ell_{\theta_j} - \ell^\star) + P^\star\omega_h \leq P^\star(\ell_{\theta_0} - \ell^\star) + P^\star\omega_h,$$

and this completes the proof, since Assumption **C4** and the dominated convergence theorem ensure that $P^\star\omega_h$ tends to zero as h does, so that almost surely

$$\sup_{\theta\in\Theta_K} \mathbb{P}_n(\ell_\theta - \ell^\star) \underset{n\to\infty}{\longrightarrow} -H(P^\star|\Pi_K),$$

and because the latter is true for any integer K (which are in countable number). □

This lemma allows to prove Theorem 7.4.1:

Proof. (of Theorem 7.4.1) It suffices to prove that $\widehat{K}_n^{\mathrm{L}}$ almost surely tends to infinity because $\widehat{K}_n^{\mathrm{L}}$ bounds below $\widehat{K}_n^{\mathrm{G}}$.

Let us work on the event of probability 1 of Lemma 7.4.7. For all $K \geq 1$, one has

$$\sup_{\theta \in \Theta_{K+1}} \mathbb{P}_n \ell_\theta - \sup_{\theta \in \Theta_K} \mathbb{P}_n \ell_\theta \underset{n \to \infty}{\longrightarrow} H(P^\star | \Pi_K) - H(P^\star | \Pi_{K+1}) > 0$$

in virtue of **C2**, because $P^\star \notin \Pi_K$. Therefore, **P0** yields that

$$\sup_{\theta \in \Theta_{K+1}} \ell_n(\theta) - \sup_{\theta \in \Theta_K} \ell_n(\theta) - \mathrm{pen}(n, K+1) + \mathrm{pen}(n, K) \underset{n \to \infty}{\longrightarrow} \infty,$$

hence $\widehat{K}_n^{\mathrm{L}} > K$ for n large enough. The latter being true for any $K \geq 1$, the proof is complete. □

The same kind of proof holds for the proposition below.

Proposition 7.4.8 (almost never underestimation). *Under Assumptions* **Com**, **P0**, **C3–C4** *and for* $P^\star \in \Pi_{K^\star} \setminus \Pi_{K^\star - 1}$*:*

- $\widehat{K}_n^{\mathrm{L}}$ *almost surely does not underestimate* $K^\star$ *whenever Assumption* **C2** *holds true;*
- $\widehat{K}_n^{\mathrm{G}}$ *almost surely does not underestimate* $K^\star$*.*

Remark 7.4.9. The assumptions of the proposition are fulfilled in the **MGM**, **RKO**, **ME**, **AC** examples for $\widehat{K}_n^{\mathrm{L}}$ and for the latter and also the **VR** example for $\widehat{K}_n^{\mathrm{G}}$, see Section 7.7.

Proof. • We aim at proving that $\mathrm{P}(\widehat{K}_n^{\mathrm{L}} < K^\star \text{ io}) = 0$ (io stands for infinitely often). Since

$$\left[\widehat{K}_n^{\mathrm{L}} < K^\star \text{ io}\right] \subset \bigcup_{K=1}^{K^\star - 1} \left[\widehat{K}_n^{\mathrm{L}} = K \text{ io}\right],$$

it suffices to prove that $\mathrm{P}(\widehat{K}_n^{\mathrm{L}} = K \text{ io}) = 0$ for any $1 \leq K < K^\star$. So, let us work on the event Ω_1 of probability 1 of Lemma 7.4.7 and denote $\delta = H(P^\star | \Pi_{K+1}) - H(P^\star | \Pi_K) < 0$ in virtue of Assumption **C2**:

$$\begin{aligned}
\left[\widehat{K}_n^{\mathrm{L}} = K \text{ io}\right] \cap \Omega_1 &\subset \left[\sup_{\theta \in \Theta_K} \mathbb{P}_n \ell_\theta - \sup_{\theta \in \Theta_{K+1}} \mathbb{P}_n \ell_\theta \geq n^{-1}\{\mathrm{pen}(n, K) - \mathrm{pen}(n, K+1)\} \text{ io}\right] \cap \Omega_1 \\
&\subset \left[\sup_{\theta \in \Theta_K} \mathbb{P}_n \ell_\theta - \sup_{\theta \in \Theta_{K+1}} \mathbb{P}_n \ell_\theta \geq -\delta/2 \text{ io}\right] \cap \Omega_1 \\
&\subset \left[\liminf_{n \to \infty} \left\{\sup_{\theta \in \Theta_K} \mathbb{P}_n \ell_\theta - \sup_{\theta \in \Theta_{K+1}} \mathbb{P}_n \ell_\theta\right\} \geq -\delta/2\right] \cap \Omega_1,
\end{aligned}$$

where the first inclusion derives from the very definition of $\widehat{K}_n^{\mathrm{L}}$, the second one stems form the condition **P0** on the penalty function and is satisfied for n large enough and the last inclusion is immediate.

Now, Lemma 7.4.7 yields that the last event has zero probability (because it implies $\delta = 0$), which concludes the proof of the first point of the proposition.

• The proof of the second point goes along the same lines: replace $(K+1)$ by $K^\star$, observe that $\delta = H(P^\star | \Pi_{K^\star}) - H(P^\star | \Pi_K) = -H(P^\star | \Pi_K) < 0$ and conclude as above. □

A la Huber trick

What we call perhaps familiarly the *à la Huber trick** is in fact a very general and powerful principle we may find a lot of examples of in the literature. The few examples we shall refer to are more or less closely related with our purpose (*i.e.* likelihood and model selection).

The idea is roughly that rescaling often enhances"performances" when empirical process methods and techniques are involved. Such a device *e.g.* allowed Pollard to propose some simple proofs of uniform central limit theorems, accordingly to the original idea of Huber as presented in its 1967 paper. One can find more recent examples in (Barron et al. 1999; Massart 2000) in a framework of model selection. Or in (Gassiat 2002), where the author derives the consistency of the maximum penalized marginal likelihood estimator of the order of a mixture with Markov regime (*i.e.* for instance, as the **MD** example, except that $\{X_i\}$ is a finite state space Markov chain and the Y_i's are conditionally independent).

The following proposition certainly casts some more light on the à la Huber trick, preparing its contribution to our purpose.

Proposition 7.4.10 (illustrating the à la Huber trick). *Set $K_2 > K_1 \geq K^\star$. Then, both inequalities below hold true, the second one providing an example of à la Huber trick result:*

$$\sup_{\theta\in\Theta_{K_2}} (\mathbb{P}_n - P^\star)(\ell_\theta - \ell^\star) \geq \sup_{\theta\in\Theta_{K_2}} \mathbb{P}_n\ell_\theta - \sup_{\theta\in\Theta_{K_1}} \mathbb{P}_n\ell_\theta \quad \text{and} \tag{7.4}$$

$$\sup_{\theta\in\Theta_{K_2}} \left((\mathbb{P}_n - P^\star)\frac{\ell_\theta - \ell^\star}{H(\theta)^{1/2}}\right)^2 \geq \sup_{\theta\in\Theta_{K_2}} \mathbb{P}_n\ell_\theta - \sup_{\theta\in\Theta_{K_1}} \mathbb{P}_n\ell_\theta. \tag{7.5}$$

Proof. The first inequality above is readily proved, since one has

$$\sup_{\theta\in\Theta_{K_2}} \mathbb{P}_n\ell_\theta - \sup_{\theta\in\Theta_{K_1}} \mathbb{P}_n\ell_\theta \leq \sup_{\theta\in\Theta_{K_2}} \mathbb{P}_n(\ell_\theta - \ell^\star) =$$
$$\sup_{\theta\in\Theta_{K_2}} \left\{(\mathbb{P}_n - P^\star)(\ell_\theta - \ell^\star) + P^\star(\ell_\theta - \ell^\star)\right\} \leq \sup_{\theta\in\Theta_{K_2}} (\mathbb{P}_n - P^\star)(\ell_\theta - \ell^\star).$$

The second one is more (à la Huber) tricky. Define for all $\theta \in \Theta_{K_2}$ such that $H(\theta) > 0$ (*i.e.* $P^\star \neq P_\theta$) the scaled likelihood ratio

$$g_\theta = \frac{\ell_\theta - \ell^\star}{H(\theta)^{1/2}}$$

and $g_\theta = 0$ otherwise. Now, for any $\theta \in \Theta_{K_2}$, $H(\theta)$ nonnegative yields

$$\mathbb{P}_n(\ell_\theta - \ell^\star) + H(\theta) = (\mathbb{P}_n - P^\star)(\ell_\theta - \ell^\star) \leq H(\theta)^{1/2} \sup_{\theta\in\Theta_{K_2}} (\mathbb{P}_n - P^\star)g_\theta. \tag{7.6}$$

Let us set some $\theta_0 \in \Theta_{K_2}$ such that both $\sup_{\theta\in\Theta_{K_2}} \mathbb{P}_n(\ell_\theta - \ell^\star) \leq \mathbb{P}_n(\ell_{\theta_0} - \ell^\star) + \varepsilon$ and $\mathbb{P}_n(\ell_{\theta_0} - \ell^\star) \geq 0$. Then, Inequality (7.6) gives for θ_0

$$\sup_{\theta\in\Theta_{K_2}} \mathbb{P}_n(\ell_\theta - \ell^\star) \leq H(\theta_0)^{1/2} \sup_{\theta\in\Theta_{K_2}} (\mathbb{P}_n - P^\star)g_\theta + \varepsilon. \tag{7.7}$$

*Our attention was kindly drawn on the à la Huber trick by Pascal Massart, who refers to it in (Barron et al. 1999) (see Proposition 7 therein and the attached remark).

Furthermore, $\mathbb{P}_n(\ell_{\theta_0} - \ell^\star) \geq 0$ combined with Inequality (7.6) implies in turn

$$H(\theta_0) \leq H(\theta_0)^{1/2} \sup_{\theta \in \Theta_{K_2}} (\mathbb{P}_n - P^\star) g_\theta, \tag{7.8}$$

hence

$$\sup_{\theta \in \Theta_{K_2}} \mathbb{P}_n \ell_\theta - \sup_{\theta \in \Theta_{K_1}} \mathbb{P}_n \ell_\theta \leq \sup_{\theta \in \Theta_{K_2}} \mathbb{P}_n(\ell_\theta - \ell^\star) \leq \left(\sup_{\theta \in \Theta_{K_2}} (\mathbb{P}_n - P^\star) g_\theta \right)^2 + \varepsilon$$

which completes the proof since $\varepsilon > 0$ is arbitrary. □

The two estimators almost surely do not overestimate

We shall hereafter conclude the proofs of Theorems 7.4.3 and 7.4.5 partly thanks to Proposition 7.4.10 and to the following bounded Law of the iterated logarithm (bounded LIL).

Theorem 7.4.11 (bounded LIL, Dudley and Philipp). *Let $\zeta_1, \ldots, \zeta_n$ be a n-tuple of independent identically distributed* random elements *with values in a seminormed vector space $(E, \|\cdot\|)$. Then, if*

(i) *there exists $C_1 > 0$ such that, for n large enough,* $\mathrm{P}(n^{-1/2} \| \sum_{i=1}^n \zeta_i \|^* > C_1) < 10^{-3}$ *and*

(ii) $\mathrm{E}(\varphi(\|\zeta_1\|))^* < \infty$,

there exists $C_2 > 0$ that depends only on C_1 such that, almost surely,

$$\limsup_{n \to \infty} \frac{\| \sum_{i=1}^n \zeta_i \|^*}{(n \log \log n)^{1/2}} \leq C_2.$$

Remark 7.4.12 (on Theorem 7.4.11**).** A subtlety deserve to be commented. The ζ_i's are not random variables but *random elements*: the terminology underlines that they are possibly non-measurable. Here, ξ^* denotes the smallest *random variable* that bounds above the random element ξ, hence both the probability $\mathrm{P}(\xi^* > a)$ and the expectation $\mathrm{E}\,\xi^*$ are well defined.

Let us introduce the class of function $\mathcal{G}'$

$$\mathcal{G}' = \left\{ \frac{\ell_\theta - \ell^\star}{H(\theta)^{1/2}} + H(\theta)^{1/2} : \theta \in \Theta_{K^*+1},\ H(\theta) > 0 \right\}$$

and apply the theorem above to the random elements

$$\zeta_i = \Big(g(Y_i) \Big)_{g \in \mathcal{G}'}$$

that take their values in $\ell^\infty(\mathcal{G}')$, vector space we equip with the uniform norm $\|\cdot\|_{\mathcal{G}'}$. In this setting,

$$\left\| \sum_{i=1}^n \zeta_i \right\|^* = n \sup_{g \in \mathcal{G}} |(\mathbb{P}_n - P^\star)\, g|$$

(in virtue of the assumption we made at the very beginning of this chapter concerning the measurability of such variables). Now, the condition (i) of the theorem is satisfied *e.g.* when the random variable above is tight. That is the reason why Assumptions **C5a** and **C5b** require that the class $\mathcal{G}$ is $P^\star$-Donsker: this is a stronger condition, but it is also quite a natural one. We could of course replace the $P^\star$-Donsker condition by a tightness condition and the result would still hold true.

We can finally proceed and prove the next two propositions:

Proposition 7.4.13 (almost never overestimation). *Under Assumptions* **P0** *and* **P1a** *on the one hand, Assumption* **C3** *on the other, for* $P^\star \in \Pi_{K^\star} \setminus \Pi_{K^\star-1}$*:*
- $\widehat{K}_n^{\mathrm{L}}$ *almost surely does not overestimate* $K^\star$ *whenever Assumption* **C5a** *holds true for* $K = K^\star + 1$*;*
- $\widehat{K}_n^{\mathrm{G}}$ *almost surely does not overestimate* $K^\star$ *whenever Assumption* **C5a** *holds true for* $K = K_{\max}$*, where* $K_{\max}$ *is a prior bound on* $K^\star$*.*

Proposition 7.4.14 (refined almost never overestimation).
Under Assumptions **P0** *and* **P1b** *on the one hand, Assumption* **C3** *on the other, for* $P^\star \in \Pi_{K^\star} \setminus \Pi_{K^\star-1}$*:*
- $\widehat{K}_n^{\mathrm{L}}$ *almost surely does not overestimate* $K^\star$ *whenever Assumption* **C5b** *holds true for* $K = K^\star + 1$*;*
- $\widehat{K}_n^{\mathrm{G}}$ *almost surely does not overestimate* $K^\star$ *whenever Assumption* **C5b** *holds true for* $K = K_{\max}$*, where* $K_{\max}$ *is a prior bound on* $K^\star$*.*

Remark 7.4.15. The assumptions of Theorem 7.4.13 hold true for the **MGM** and **VR** examples and the assumptions of Theorem 7.4.14 are satisfied in the **MGM** example, see Section 7.7.

The two proofs are very similar, thus we shall only present the first one, including a sole remark when the second proof would slightly differ. The scheme is simple: *primo*, apply Proposition 7.4.10 in order to translate the overestimation problem into a bounded LIL problem; *secundo*, apply the bounded LIL of Theorem 7.4.11 above; *tertio*, conclude.

Proof. (of Proposition 7.4.13*)* • We aim at proving that $\mathrm{P}(\widehat{K}_n^{\mathrm{L}} > K^\star \text{ io}) = 0$ (io stands for infinitely often). Since the following inclusions hold true (for the second inclusion, apply the first inequality of Proposition 7.4.10 – the second one when proving Proposition 7.4.14)

$$\begin{aligned}
\left[\widehat{K}_n^{\mathrm{L}} > K^\star \text{ io}\right] &\subset \left[\sup_{\theta\in\Theta_{K^\star}} \ell_n(\theta) - \mathrm{pen}(n,K^\star) \leq \sup_{\theta\in\Theta_{K^\star+1}} \ell_n(\theta) - \mathrm{pen}(n,K^\star+1) \text{ io}\right] \\
&= \left[\sup_{\theta\in\Theta_{K^\star+1}} \mathbb{P}_n\ell_\theta - \sup_{\theta\in\Theta_{K^\star}} \mathbb{P}_n\ell_\theta \geq n^{-1}\{\mathrm{pen}(n,K^\star+1) - \mathrm{pen}(n,K^\star)\} \text{ io}\right] \\
&\subset \left[\sup_{\theta\in\Theta_{K^\star+1}} (\mathbb{P}_n - P^\star)(\ell_\theta - \ell^\star) \geq n^{-1}\{\mathrm{pen}(n,K^\star+1) - \mathrm{pen}(n,K^\star)\} \text{ io}\right] \\
&\subset \left[\sup_{\theta\in\Theta_{K^\star+1}} |(\mathbb{P}_n - P^\star)(\ell_\theta - \ell^\star)| \geq n^{-1}\{\mathrm{pen}(n,K^\star+1) - \mathrm{pen}(n,K^\star)\} \text{ io}\right],
\end{aligned}$$

hence

$$\left[\widehat{K}_n^{\mathrm{L}} > K^\star \text{ io}\right] \subset \left[\left\{\frac{(n\log\log n)^{1/2}}{\mathrm{pen}(n,K^\star)}\right\} \cdot \left\{\frac{n\sup_{\theta\in\Theta_{K^\star+1}} |(\mathbb{P}_n - P^\star)(\ell_\theta - \ell^\star)|}{(n\log\log n)^{1/2}}\right\} \geq \frac{\mathrm{pen}(n,K^\star+1)}{\mathrm{pen}(n,K^\star)} - 1 \text{ io}\right].$$

Now, the latter event has zero probability for n large enough. Indeed, the right hand term of the inequality has a positive lim inf thanks to the first part of Assumption **P1a** while its second part ensures that the first expression between braces above tends to zero. Consequently, the whole left side of the inequality almost surely tends to zero, since Theorem 7.4.11 implies that the second expression between braces is bounded almost surely. This concludes the proof of this case.

• Let us consider now $\widehat{K}_n^{\mathrm{G}}$. The existence of the prior bound $K_{\max}$ allows to write

$$\left[\widehat{K}_n^{\mathrm{G}} > K^\star \text{ io}\right] = \bigcup_{K=K^\star+1}^{K_{\max}} \left[\widehat{K}_n^{\mathrm{G}} = K \text{ io}\right], \text{ with}$$

$$\left[\widehat{K}_n^{\mathrm{G}} = K \text{ io}\right] \subset \left[\sup_{\theta\in\Theta_{K^\star}} \ell_n(\theta) - \mathrm{pen}(n,K^\star) \leq \sup_{\theta\in\Theta_K} \ell_n(\theta) - \mathrm{pen}(n,K) \text{ io}\right].$$

Thus, it suffices to prove that the above events have zero probability for all $K^\star+1 \leq K \leq K_{\max}$. Minor changes in the previous study of $\widehat{K}_n^{\mathrm{L}}$ allow to conclude. □

Conclusion

The combination of Propositions 7.4.8, 7.4.13 and 7.4.14 yields the two Theorems 7.4.3 and 7.4.5.

7.5. Underestimation

This section addresses the study of the rate of underestimation of our estimators $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$. An application of the discussed Stein's lemma provides an universal lower bound. Then, Sanov's theorem in its refined form presented in Interlude 6 allows to get some upper bounds.

7.5.1. Stein's lemma for a lower bound on the rate

In this subsection, we shall derive in Proposition 7.5.1 an universal lower rate of underestimation for any estimator that *does not almost surely overestimate the order* (see Assumption **S1** below). The proof relies on a very general lemma known as *Stein's lemma*, whose result has been summarized as follows by Bahadur et al. (1980):

Given two probability measures on $(\Omega, \mathcal{A})$ and a sequence $\{A_n\}$ of measurable sets,

$$\liminf_{n\to\infty} Q(A_n) > 0 \quad \Longrightarrow \quad \liminf_{n\to\infty} n^{-1}\log P(A_n) \geq -H(Q\,|P). \tag{7.9}$$

The latter lemma will be commented in Remark 7.5.2 below.

Let $\widetilde{K}_n$ denote any estimator of the order and Assumption **S1** (**S** stands for Stein) be

S1 For all $K \geq 1$ and $P_\theta \in \Pi_K$,

$$\limsup_{n\to\infty} \mathrm{P}_{P_\theta}(\widetilde{K}_n > K) < 1.$$

Now,

Proposition 7.5.1 (yet another version of Stein's lemma, underestimation case). *Let $\widetilde{K}_n$ be any estimator of the true order $K^\star$ of $P^\star$. Suppose that Assumption* **S1** *holds true and that, for all $\theta \in \Theta_{K^\star-1}$, $\ell_\theta, \ell_{\theta^\star} \in L^1(P_\theta)$. Then*

$$\liminf_{n\to\infty} n^{-1} \log \mathrm{P}_{P^\star}(\widetilde{K}_n < K^\star) \geq - \inf_{K<K^\star} H(\Pi_K \,|P^\star).$$

Proof. Set $K < K^\star$, $\theta \in \Theta_K$ and $\varepsilon > 0$. Then

$$\begin{aligned}\mathrm{P}(\widetilde{K}_n < K^\star) = \mathrm{P}_{P^\star}(\widetilde{K}_n < K^\star) \geq \mathrm{P}_{P^\star}(\widetilde{K}_n \leq K) = \mathrm{E}_{P^\star}\mathbb{1}\{\widetilde{K}_n \leq K\} \\ \geq \mathrm{E}_{P_\theta}\mathbb{1}\{\widetilde{K}_n \leq K\} e^{\ell_n(\theta^\star)-\ell_n(\theta)} \geq e^{-n\left(H(P_\theta\,|P^\star)+\varepsilon\right)} \mathrm{E}_{P_\theta}\mathbb{1}\{\widetilde{K}_n \leq K, F_n\}\end{aligned}$$

by change of probability and for $F_n = \left\{|n^{-1}\big(\ell_n(\theta^\star) - \ell_n(\theta) + H(P_\theta\,|P^\star)\big)| < \varepsilon\right\}$. Since $\ell_\theta, \ell_{\theta^\star}$ are P_θ-integrable, the Law of large numbers yields that $P_\theta(F_n^c)$ tends to zero as $n \uparrow \infty$. Besides the previous lower bounding implies

$$\mathrm{P}(\widetilde{K}_n < K^\star) \geq e^{-n\left(H(P_\theta\,|P^\star)+\varepsilon\right)} \big(1 - P_\theta(\widetilde{K}_n > K) - P_\theta(F_n^c)\big)$$

where the second factor in the right hand term is bounded away from 0 as $n \uparrow \infty$ thanks to Assumption **S1** holds. We conclude by taking $n^{-1}\log$ and then $\liminf_n$ since K, θ and ε are arbitrary. □

The proof of Proposition 7.5.1 encloses the demonstration of Stein's lemma in its version by Bahadur et al.. Let us identify the probabilities P, Q and the events A_n according to (7.9) in the proof above: $P = \mathrm{P}_{P^\star}$, $Q = \mathrm{P}_{P_\theta}$ and $A_n = [\widetilde{K}_n \leq K] \cap F_n$.

Remark 7.5.2. Stein's lemma is quite famous in Statistics. We shall *mimic* it once again in this chapter, see Section 7.5.1. One can indeed refer to Stein's lemma as a versatile, powerful yet simple result due to Stein, though he never published it. Chernoff first mentions it in a framework of hypothesis testing. The corresponding version of the lemma is stated and proved *e.g.* in (Bahadur 1967; Bahadur 1971). It is intimately related to the notions of Bahadur efficiency and Hodges-Lehmann optimality, see *e.g.* (van der Vaart 1998; Kourouklis 1991; Kallenberg and Kourouklis 1992). There, Stein's lemma provides a benchmark (*via* a minoration) and efficiency or optimality roughly consists, for a test, of achieving the lower bound.

Bahadur, Zabell, and Gupta probably present the most concise – and therefore practical – version of Stein's lemma in (Bahadur et al. 1980): this is the preliminary implication (7.9) exhibited above. It is the core of Stein's lemma original proof and, in most cases, the key of its various versions, as in this work. It is also the case in another example which illustrates both the versatility and the power of the discussed lemma: Gassiat and Boucheron use Stein's lemma in order to characterize the best underestimation exponent when estimating the order of a Hidden Markov Model.

7.5.2. Upper bounds on the rate

We assume that the reader is familiar with the notations and results of the Interlude 6. Indeed, we shall derive hereafter an exponentially decreasing rate of underestimation from the Sanov's Theorem 6.2.4 of Interlude 6.

It is shown than $\widehat{K}_n^{\mathrm{G}}$ always achieves faster exponential rates than $\widehat{K}_n^{\mathrm{L}}$ and that the latter is true in spite of weaker assumptions for $\widehat{K}_n^{\mathrm{G}}$.

Nonetheless, both results hold true under mild assumptions, at least with regard to the typical conditions met in the corresponding literature and to the encompassed examples (see Remark 7.5.4 below).

The estimator underestimates with an exponentially decreasing rate

Let us state the three new assumptions required in the forthcoming theorem (**U** stands for Underestimation).

U1 The following three properties are satisfied:

(i) For any $P_\theta, P_t \in \Pi_{K^\star-1}$, $P_\theta << P_t$ and $H(P_\theta \,|P^\star) < \infty$.

(ii) $\{p_\theta \, \ell_\theta : \theta \in \Theta_{K^\star-1}\} \cup \{p_{\theta^\star} \, \ell_{\theta^\star}\} \subset L^1(\mu)$.

(iii) The functions ℓ_θ $(\theta \in \Theta_{K^\star})$ admit some exponential moments, or equivalently $\{\, \ell_\theta : \theta \in \Theta_{K^\star}\} \subset \mathcal{L}_\tau$.

U2 There exist $l, u \in \mathbb{R}^{\mathcal{Y}}$ such that $(u-l) \in \mathcal{L}_\tau$,

$$l \leq \ell_\theta \leq u \quad (\text{any } \theta \in \Theta_{K^\star})$$

and for all $\varepsilon > 0$ and $Q \in \mathcal{Q}$, there exists a compact set C of $\mathcal{Y}$ satisfying $Q\,(u-l)\mathbb{1}\{C^c\} < \varepsilon$.

U3 For all $K \leq K^\star$ and for any compact set C of $\mathcal{Y}$, the set $\{\, \ell_\theta \mathbb{1}\{C\} : \theta \in \Theta_K\}$ is precompact in the class $\mathrm{C}^0(C, \|\cdot\|_\infty)$ of all the continuous functions on C equipped with the uniform norm.

Let us introduce finally the sets $\Lambda_{\alpha,K}$ and $\Gamma_{\alpha,K}$ (Λ stands for Local and Γ for Global):

$$\Lambda_{\alpha,K} = \{Q \in \mathcal{Q} : \sup_{\theta\in\Theta_K} Q\,\ell_\theta - \sup_{\theta\in\Theta_{K+1}} Q\,\ell_\theta \geq -\alpha\} \quad (\text{any } \alpha \geq 0), \tag{7.10}$$

$$\Gamma_{\alpha,K} = \{Q \in \mathcal{Q} : \sup_{\theta\in\Theta_K} Q\,\ell_\theta - \sup_{\theta\in\Theta_{K^\star}} Q\,\ell_\theta \geq -\alpha\} \quad (\text{any } \alpha \geq 0). \tag{7.11}$$

Observe that $\Gamma_{\alpha,K} \subset \Lambda_{\alpha,K}$ $(K < K^\star)$, hence $-I(\Gamma_{\alpha,K}\,|P^\star) \leq -I(\Lambda_{\alpha,K}\,|P^\star)$ (the definition of $I(\cdot\,|P^\star)$ is given in Chapter 6, Section 6.2.2).

Theorem 7.5.3 (exponential rate of underestimation).
Whenever Assumptions **Com**, **P0** *and* **U1–U3** *hold true,*

- *if Assumption* **C2** *is also satisfied, then*

$$\limsup_{n\to\infty} n^{-1} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} < K^\star) \leq -\inf_{K<K^\star} I(\Lambda_{0,K}\,|P^\star) < 0; \tag{7.12}$$

- *without any further assumption,*

$$\limsup_{n\to\infty} n^{-1} \log \mathrm{P}(\widehat{K}_n^{\mathrm{G}} < K^\star) \leq -\inf_{K<K^\star} I(\Gamma_{0,K}\,|P^\star) < 0. \tag{7.13}$$

Remark 7.5.4.

- The assumptions of the theorem above hold true in the **MGM** and **VR** examples for $\widehat{K}_n^{\mathrm{G}}$ and only in the **MGM** example for $\widehat{K}_n^{\mathrm{L}}$, see Section 7.7.
- The main contribution of this theorem to the already existing collection of results that cope with rate of underestimation of an order relies in its mild assumptions. This is due to the nature of our approach: the analysis of underestimation events is based on functional arguments, *i.e.* are described as events concerning the empirical measure.

In a framework of finite mixtures, for a penalty function of the form $\mathrm{pen}(n,K) = v_n\, D(K)$ and for a number n of observations large enough, Dacunha-Castelle and Gassiat (1997) bound above the probability of misestimating the order (*i.e.* both underestimating and overestimating it) by an expression exponential in $n^{-1}v_n^2$ (this is actually the rate we get when overestimating in Theorem 7.6.3). However, this result crucially depends on the ability of estimating all the moments of the parameters wrt the mixing distribution thanks to the observations (see the paper for more comments).

Guyon and Yao (1999) give a description of the events on which underestimation (and also overestimation, see Remark 7.6.4) occurs. The authors get nonasymptotic upper bounds for the probability of underestimation. From that point of view, their result is stronger than ours. Nevertheless, the proof relies on exponential inequalities for a sample statistic $T_n \in \mathbb{R}^p$. Indeed, the main assumption of the paper is that the penalized empirical criterion $\sup_\theta U_n(\theta) + n^{-1}v_n\, K$ (the supremum is for θ ranging over Θ_K) whose minimization in $K \leq K_{\mathsf{max}}$ defines the estimator of the order is factorized by T_n, *i.e.* equals

$$\sup\left\{U(\theta, T_n) : \theta \in \Theta_K\right\} + n^{-1}v_n\, K,$$

where U is a deterministic function. The other assumptions of identifiability and smoothness are more standard. From this point of view, our result applies in frameworks where their does not (think *e.g.* of the **MD** example). For other references in the framework of (Guyon and Yao 1999), we refer the reader to the bibliography of the paper itself.

Haughton and Keribin (2001) also consider the rate of convergence of the probability of underestimating an order (and overestimating it), under more restrictive assumptions than ours. They particularly require twice continuously differentiability of the log–likelihood and nonsingularity of the Fisher information matrix, as well as an extra condition on the uniqueness of the projection (wrt the Kullback-Leibler divergence) of the distribution of the observations on the submodels (*i.e.* with our notations, of $P^\star$ on Π_K for $K < K^\star$). The required conditions limit the field of application of their result. The **MD** and **AC** examples provide two important cases where our result applies whereas their does not.

Finally, Gassiat and Boucheron (2001) have obtained upper bounds for the underestimation probability of the order of a Hidden Markov Model. These upper bounds coincide with the optimal lower bound yielded by Stein's lemma.

How close the exponential rate of underestimation can be to the best one ?

Once defined the following subsets of $\Lambda_{\alpha,K}$ and $\Gamma_{\alpha,K}$,

$$\Lambda_{0,K}^a = \Lambda_{0,K} \cap M_1(\mathcal{Y}), \tag{7.14}$$

$$\Gamma_{0,K}^a = \Gamma_{0,K} \cap M_1(\mathcal{Y}), \tag{7.15}$$

the following lemma allows comparisons:

Lemma 7.5.5. *Suppose that Assumption* **U1** (i), (ii) *holds true. Then for all* $K < K^\star$,
• $\Pi_K \subset \Lambda^a_{0,K} \subset \Lambda_{0,K}$ *and forwardly,*

$$-\inf_{K<K^\star} H(\Pi_K \,|P^\star) \leq -\inf_{K<K^\star} H(\Lambda^a_{0,K}\,|P^\star) \leq -\inf_{K<K^\star} I(\Lambda_{0,K}\,|P^\star);$$

• $\Pi_K \subset \Gamma^a_{0,K} \subset \Gamma_{0,K}$ *and forwardly,*

$$-\inf_{K<K^\star} H(\Pi_K \,|P^\star) \leq -\inf_{K<K^\star} H(\Gamma^a_{0,K}\,|P^\star) \leq -\inf_{K<K^\star} I(\Gamma_{0,K}\,|P^\star).$$

Finally, we strengthen Theorem 7.5.3 thanks to the additional assumptions hereafter. Let us emphasize that the structure of the spaces Θ_K is here (not so) implicitly reinforced: now, the distance d on Θ_K derives from a norm $\|\cdot\|$ on the vector space Θ_K, so that the notion of differentiability wrt to $\theta \in \Theta_K$ is available. Particularly, the **AC** example is therefore excluded since the distance d that equips Θ_K in that case involves the distance d_μ on Caccioppoli partitions and that the latter does not derive from a norm.

U4 Let $l, u \in \mathbb{R}^{\mathcal{Y}}$ as defined above in Assumption **U2**. There exists a compact set C_{lu} of $\mathcal{Y}$ such that $\mathbb{1}\{C^c_{lu}\}(u-l) \in \mathcal{M}_\tau$.

U5 For any $K \leq K^\star$ and $y \in \mathcal{Y}$, the functions $\theta \mapsto \ell_\theta(y)$ are differentiable on the interior of Θ_K, with derivative $\dot{\ell}_\theta(y)$. Moreover, the coordinates of $\dot{\ell}_\theta$ are elements of $\mathcal{M}_\tau$ and there exists $F \in \mathcal{L}_\tau$ such that

$$|\ell_{\theta+h} - \ell_\theta - \dot{\ell}_\theta{}^T h| \leq F \cdot o(h).$$

It suffices that h in the latter inequality has the particular form $h = \|h\|\, e_k$ where e_k is the k-th canonical base vector of Θ_K.

We can state the last theorem that copes with the underestimation rate. We exhibit again some exponential rates in n whose constants

$$-\inf_{K<K^\star} H(\Lambda^a_{0,K}\,|P^\star) \quad \text{and} \quad -\inf_{K<K^\star} H(\Gamma^a_{0,K}\,|P^\star)$$

are closer to the best possible than the previous constant obtained in Theorem 7.5.3, see (7.12), (7.13) and Lemma 7.5.5.

Theorem 7.5.6 (refined exponential rate of underestimation).
Suppose that Assumptions **P0**, **Com** *and* **U1–U5** *hold true. Then,*
• *if in addition Assumption* **C2** *is fulfilled,*

$$\limsup_{n\to\infty} n^{-1} \log \mathrm{P}(\widehat{K}^{\mathrm{L}}_n < K^\star) \leq -\inf_{K<K^\star} H(\Lambda^a_{0,K}\,|P^\star) < 0;$$

• *without any further assumption,*

$$\limsup_{n\to\infty} n^{-1} \log \mathrm{P}(\widehat{K}^{\mathrm{G}}_n < K^\star) \leq -\inf_{K<K^\star} H(\Gamma^a_{0,K}\,|P^\star) < 0.$$

Remark 7.5.7. The assumptions of the theorem above hold true in the **MGM** and **VR** examples for $\widehat{K}^{\mathrm{G}}_n$ and only in the **MGM** example for $\widehat{K}^{\mathrm{L}}_n$, see Section 7.7.

Remark 7.5.8 (finer yet refinement ?). One would have naturally preferred to get a more precise result involving the left hand quantity of the inequalities stated in Lemma 7.5.5, *i.e.* the best possible constant with a view to Proposition 7.5.1. Actually, we did not manage to reach it in great generality, though we proved that the best rate is achieved in exponential models by $\widehat{K}_n^{\mathrm{G}}$:

Proposition 7.5.9 (achievement of the optimal rate in exponential models). *Under the assumptions of Theorem* 7.5.6 *and* for exponential models, *the best underestimation rate with a view to Proposition* 7.5.1 *is achieved by* $\widehat{K}_n^{\mathrm{G}}$.

Remark 7.5.10. Proposition 7.5.9 particularly applies to the **VR** example, see Section 7.7. This is a new result.

In this case, the proof involves *H-projections* as defined and studied by Csiszár (1975): the H-projection $\overline{Q}$ of the probability measure Q on a convex set of probability measures $\mathcal{C}$ is defined (under conditions) by

$$\overline{Q} = \arg\min \Big\{ H(P|Q) : P \in \mathcal{C} \Big\}.$$

It enjoys some remarkable properties, including a useful Pythagorean equality. Nonetheless, the proof also involves much less tractable projections with reversed order of P and Q in the definition above. We can conclude in exponential models in the spirit of (Čencov 1982), where the author derives another Pythagorean equality for reversed projections.

On the contrary, Gassiat and Boucheron (2001) showed that their order estimator of a Hidden Markov Model did achieve the best possible rate without restriction.

7.5.3. Proofs and more comments, to be continued

This section is devoted to the presentation of the proofs we first omitted for sake of clarity. It is divided into four parts: the first is concerned with Lemma 7.5.5 which allows comparison of the rates of convergence we obtained with the best possible rate exhibited in Proposition 7.5.1. The second one encloses some useful properties of the sets $\Lambda_{\alpha,K}$ and $\Gamma_{\alpha,K}$, among which the fact that those sets are closed. The third one consists of comments on Assumptions **U2–U3**. The fourth one presents the proof of Theorem 7.5.3.

Proof of Lemma 7.5.5

According to what we announced earlier, the mutual core of the proofs of Theorems 7.5.3 and 7.5.6 relies on an application of the extended Sanov's theorem stated in Interlude 6, and therefore involves the space $\mathcal{Q}$ of all the linear forms on $\mathcal{L}_\tau$ which are nonnegative and integrate to one.

A simple application of Proposition 6.2.3 of Interlude 6 casts some light on the relationship between $\Pi_{K^\star-1}$ and $\mathcal{Q}$ on the one hand, and $H(\cdot|P^\star)$ and $I(\cdot|P^\star)$ restricted to $\Pi_{K^\star-1}$ on the other: as stated below, elements of $\Pi_{K^\star-1}$ are continuous elements of $\mathcal{Q}$ and the two functions coincide on $\Pi_{K^\star-1}$.

Lemma 7.5.11. *Under Assumption* **U1** (i), *any* $P_\theta \in \Pi_{K^\star-1}$, *satisfies* $\dfrac{\mathrm{d}P_\theta}{\mathrm{d}P^\star} \in M_{\tau^\star}$. *Consequently,* $\Pi_{K^\star-1} \subset L'_\tau \cap \mathcal{Q}$ *and the two functions* $H(\cdot|P^\star)$, $I(\cdot|P^\star)$ *coincides on* $\Pi_{K^\star-1}$.

The lemma above is an element of the postponed proof of Lemma 7.5.5.

Proof. (of Lemma 7.5.5) Set $K < K^\star$. It suffices to prove that $\Pi_K \subset \Lambda_{0,K}$ and $\Pi_K \subset \Gamma_{0,K}$. Now, Lemma 7.5.11 ensures that $\Pi_K \subset \mathcal{Q}$. Let us choose $P_\theta \in \Pi_K$ and prove that it belongs to $\Lambda^a_{0,K}$ and $\Gamma^a_{0,K}$.

Let T denote Θ_K, or Θ_{K+1}, or $\Theta_{K^\star}$ and take $t \in T$. Remember that $\ell_\theta \in L^1(P_\theta)$ thanks to Assumption **U1** (ii).

Now, observe that we can decompose the Kullback-Leibler divergence $H(P_\theta \,|P_t)$ as follows: $P_\theta << P_t$ (**U1** (i)) is equivalent to $P_\theta(p_t = 0) = 0$, so $H(P_\theta \,|P_t) = P_\theta \log(p_\theta/p_t) = P_\theta(\ell_\theta - \ell_t)$, hence finally $H(P_\theta \,|P_t) = P_\theta\, \ell_\theta - P_\theta \ell_t$ because $P_\theta\, \ell_\theta < \infty$. Particularly, $H(P_\theta \,|P_t) = \infty$ yields $P_\theta \ell_t = -\infty$.

Thus

$$\sup_{t\in T} P_\theta \ell_t = \sup\{P_\theta \ell_t : t \in T,\ H(P_\theta \,|P_t) < \infty\} = -\inf_{t\in T} H(P_\theta \,|P_t) + P_\theta\, \ell_\theta = P_\theta\, \ell_\theta$$

which allows to conclude that $P_\theta \in \Lambda_{0,K}$ and $P_\theta \in \Gamma_{0,K}$. □

Describing the sets $\Lambda_{\alpha,K}$ and $\Gamma_{\alpha,K}$

We point out hereafter some useful properties of the sets $\Lambda_{\alpha,K}$ and $\Gamma_{\alpha,K}$ defined by (7.10) and (7.11).

Lemma 7.5.12. *On the one hand, $\Lambda_{\alpha,K} \subset \Lambda_{\alpha',K}$ and $\Gamma_{\alpha,K} \subset \Gamma_{\alpha',K}$ if $\alpha \le \alpha'$. On the other hand, the sets $\Lambda_{\alpha,K}$ and $\Gamma_{\alpha,K}$ are $\mathcal{S}$-measurable as soon as Assumption* **U1** (iii) *holds.*

Proof. The inclusion property is obvious. The measurability derives from the fact that $Q \mapsto Q\, \ell_\theta$ is $\mathcal{S}$-measurable since $\ell_\theta \in \mathcal{L}_\tau$. □

Lemma 7.5.13. *Whenever Assumptions* **U1** (iii), **U2** *and* **U3** *hold true, the sets $\Lambda_{\alpha,K}$ and $\Gamma_{\alpha,K}$ are closed wrt the topology $\mathcal{T}$ for every $\alpha > 0$, $K < K^\star$.*

Remark 7.5.14. The scheme of proof of Theorem 7.5.3 involves the upper bound of the extended Sanov's theorem applied to $\Lambda_{\alpha,K}$ and $\Gamma_{\alpha,K}$, hence the necessary control of their closures $\mathrm{cl}(\Lambda_{\alpha,K})$ and $\mathrm{cl}(\Gamma_{\alpha,K})$. But the closure wrt the $\sigma(\mathcal{L}^*_\tau, \mathcal{L}_\tau)$-topology is quite difficult to handle, since this topology is not metrizable. Particularly, a set S is closed not only if all converging S-valued sequences converge in S (on the contrary, the latter is true when replacing sequences by nets). We finally propose a simple proof of the closeness of $\Lambda_{\alpha,K}$ and $\Gamma_{\alpha,K}$ that relies on simple, general assumptions, see **U2–U3** and the next subsection.

Proof. Set $\alpha > 0$, $K < K^\star$ and $\varepsilon > 0$. We shall actually prove that $\Lambda^c_{\alpha,K}$ is an open set. The same proof applies to $\Gamma_{\alpha,K}$, up to minor changes in indices.

Choose $Q_0 \in \Lambda^c_{\alpha,K}$ and $\alpha' > \alpha$ such that $\sup_{\theta\in\Theta_K} Q_0\, \ell_\theta - \sup_{\theta\in\Theta_{K+1}} Q_0\, \ell_\theta < -\alpha'$: can we construct an open neighbourhood of Q_0 in $\Lambda^c_{\alpha,K}$?

First, observe that for any subset A of $\mathcal{Y}$, T of Θ_{K+1} and $Q \in \mathcal{L}_\tau$, $Q \ge 0$,

$$\sup_{\theta\in T} Q\, \ell_\theta \mathbb{1}\{A\} \le \sup_{\theta\in T} Q\, \ell_\theta - Q\, l \mathbb{1}\{A^c\},$$
$$\sup_{\theta\in T} Q\, \ell_\theta \le \sup_{\theta\in T} Q\, \ell_\theta \mathbb{1}\{A\} + Q\, u \mathbb{1}\{A^c\}$$

and forwardly

$$\sup_{\theta\in\Theta_K} Q\,\ell_\theta \mathbb{1}\{A\} - \sup_{\theta\in\Theta_{K+1}} Q\,\ell_\theta \mathbb{1}\{A\} \le \sup_{\theta\in\Theta_K} Q\,\ell_\theta - \sup_{\theta\in\Theta_{K+1}} Q\,\ell_\theta + Q\,(u-l)\mathbb{1}\{A^c\}, \quad (7.16)$$

$$\sup_{\theta\in\Theta_K} Q\,\ell_\theta - \sup_{\theta\in\Theta_{K+1}} Q\,\ell_\theta \le \sup_{\theta\in\Theta_K} Q\,\ell_\theta \mathbb{1}\{A^c\} - \sup_{\theta\in\Theta_{K+1}} Q\,\ell_\theta \mathbb{1}\{A^c\} + Q\,(u-l)\mathbb{1}\{A^c\}. \quad (7.17)$$

Consequently, if C is a compact set of $\mathcal{Y}$ such that $Q_0\,(u-l)\mathbb{1}\{C^c\} < \varepsilon$ in virtue of Assumption **U2**, Inequality (7.16) with $Q = Q_0$ and $A = C$ gives

$$\sup_{\theta\in\Theta_K} Q_0\,\ell_\theta \mathbb{1}\{C\} - \sup_{\theta\in\Theta_{K+1}} Q_0\,\ell_\theta \mathbb{1}\{C\} \le \alpha' + \varepsilon. \quad (7.18)$$

Secondly, since $\{\ell_\theta \mathbb{1}\{C\} : \theta \in \Theta_K\}$ ($\{\ell_\theta \mathbb{1}\{C\} : \theta \in \Theta_{K+1}\}$, respectively) is precompact in $\mathrm{C}^0(C, \|\cdot\|_\infty)$, the Borel-Lebesgue property ensures that there exist $\theta_1, \ldots, \theta_r \in \Theta_K$ ($\theta_{r+1}, \ldots, \theta_{r+s} \in \Theta_{K+1}$, respectively) such that the union over $1 \le i \le r$ ($r+1 \le i \le r+s$, respectively) of the balls $B_{\|\cdot\|_\infty}(\ell_{\theta_i}\mathbb{1}\{C\}, \varepsilon)$ covers it. Observe now that $f\mathbb{1}\{A\} \in \mathcal{L}_\tau$ as soon as $f \in \mathcal{L}_\tau$ and introduce the open neighbourhood V of Q_0 defined by

$$V = \Big\{Q \in \mathcal{Q} : \big|Q\,(u-l)\mathbb{1}\{C^c\} - Q_0\,(u-l)\mathbb{1}\{C^c\}\big| < \varepsilon\Big\} \cap \bigcap_{i=1}^{r+s} \Big\{Q \in \mathcal{Q} : \big|Q\,\ell_{\theta_i}\mathbb{1}\{C\} - Q_0\,\ell_{\theta_i}\mathbb{1}\{C\}\big| < \varepsilon\Big\}.$$

Then for all $Q \in V$, since $Q1 = 1$ and $Q \ge 0$, one can verify easily that

$$\sup_{\theta\in\Theta_K} Q\,\ell_\theta \mathbb{1}\{C\} \le \sup_{\theta\in\Theta_K} Q_0\,\ell_\theta \mathbb{1}\{C\} + 2\varepsilon,$$

$$\sup_{\theta\in\Theta_{K+1}} Q_0\,\ell_\theta \mathbb{1}\{C\} \le \sup_{\theta\in\Theta_{K+1}} Q\,\ell_\theta \mathbb{1}\{C\} + 3\varepsilon.$$

Combining those two previous inequalities with Inequalities (7.17,7.18) implies that $V \subset \Lambda^c_{\alpha,K}$ since $\varepsilon > 0$ is arbitrarily small. □

Lemma 7.5.15. *Assume Assumptions* **Com** *and* **U1** (ii) *hold true and set* $K < K^\star$.
- *If Assumption* **C2** *is fulfilled, then* $P^\star \notin \Lambda_{0,K}$.
- *Without any additional assumption,* $P^\star \notin \Gamma_{0,K}$.

The latter lemma will allow to conclude the proof of Theorem 7.5.3 by ensuring that the obtained upper bound is negative.

Proof. • Indeed, if we had $P^\star \in \Lambda_{0,K}$, we would also have

$$0 \le \sup_{\theta\in\Theta_K} P^\star\,\ell_\theta - \sup_{\theta\in\Theta_{K+1}} P^\star\,\ell_\theta = \sup_{\theta\in\Theta_K} P^\star\,\ell_\theta - P^\star\,\ell_{\theta^\star} + P^\star\,\ell_{\theta^\star} - \sup_{\theta\in\Theta_{K+1}} P^\star\,\ell_\theta = -H(P^\star|\Pi_K) + H(P^\star|\Pi_{K+1}) \le 0,$$

and this contradicts Assumption **C2.**

• In the same spirit, if we had $P^\star \in \Gamma_{0,K}$, we would also have

$$0 \le \sup_{\theta\in\Theta_K} P^\star\,\ell_\theta - \sup_{\theta\in\Theta_{K^\star}} P^\star\,\ell_\theta \le \sup_{\theta\in\Theta_K} P^\star\,\ell_\theta - P^\star\,\ell_{\theta^\star} = -H(P^\star|\Pi_K).$$

Now, $H(\cdot\,|\,\cdot)$ is lower semicontinuous over $M_1(\mathcal{Y}) \times M_1(\mathcal{Y})$ equipped with the product of the weak topologies, so $H(P^\star\,|\,\cdot)$ achieves its infimum on compact sets and particularly on Π_K. Since then $P^\star \notin \Pi_K$, the right hand term above is negative. We have a contradiction. □

Comments on Assumptions U2–U4

If Assumption **U2** holds true, then $(u-l)$ is an envelope function for the set $\{\ell_\theta - \ell_t : \theta, t \in \Theta_{K^\star}\}$ that admits some exponential moments wrt $P^\star$. The additional condition on $(u-l)$, *i.e.* the existence for any $\varepsilon > 0$ and $Q \in \mathcal{L}_\tau^*$ of a compact subset C of $\mathcal{Y}$ such that $Q\,(u-l)\mathbb{1}\{C^c\} < \varepsilon$ is perhaps more restrictive than it seems at first look. A typical situation when it is satisfied is enclosed in the following lemma:

Lemma 7.5.16. *Let $\psi \in \mathbb{R}_+^{*\,\mathbb{R}}$ be an increasing positive function such that the ratio function $x \mapsto \psi(x)/x$ tends to infinity as $|x|$ also does. Whenever $\psi(u-l) \in \mathcal{L}_\tau$ and*

$$\Big\{y \in \mathcal{Y} : (u-l)(y) \le M\Big\} \text{ are compact sets (any } M > 0), \tag{7.19}$$

for all $\varepsilon > 0$ and $Q \in \mathcal{L}_\tau^*$, there exists a compact subset C of $\mathcal{Y}$ such that $Q\,(u-l)\mathbb{1}\{C^c\} < \varepsilon$.

In case $\mathcal{Y} = \mathbb{R}^q$ and $(u-l)$ is continuous, the condition stated in (7.19) is fulfilled. Furthermore, $(u-l)$ admits then all exponential moments, i.e. *$(u-l) \in M_\tau$, and therefore, Assumption* **U4** *holds true.*

Proof. • Set $Q \in \mathcal{L}_\tau^*$. If $(u-l)$ is bounded, then $(u-l) \in \mathcal{M}_\tau$ hence $Q^s(u-l) = 0$ and Lemma 6.1.5 of Interlude 6 ensures that $Q\,(u-l)\mathbb{1}\{u-l > M\}$ tends to zero as M tends to infinity, which is the expected result.

Suppose now that $(u-l)$ is not bounded. We have

$$M^{-1}\psi(M) \cdot Q\,(u-l)\mathbb{1}\{u-l > M\} \le Q\,\psi(u-l)\mathbb{1}\{u-l > M\} \le Q\,\psi(u-l),$$

and this again, yields the expected result.

• Now, assume that $\psi(u-l) \in \mathcal{L}_\tau$ with $\mathcal{Y} = \mathbb{R}^q$ and $(u-l)$ continuous. Then the envelope function $(u-l)$ admits any exponential moments, *i.e.* $(u-l) \in \mathcal{M}_\tau$. Indeed, if $(u-l)$ is bounded, then it belongs to $\mathcal{M}_\tau$. Thus, suppose that $(u-l)$ is unbounded, let $a > 0$ be such that $P^\star \exp\{a\psi(u-l)\} < \infty$ and set any $b > 0$. We can write

$$\begin{aligned} P^\star \exp\{b(u-l)\} = \\ P^\star \exp\{b(u-l)\}\mathbb{1}\{\psi(u-l) > b(u-l)/a\} + P^\star \exp\{b(u-l)\}\mathbb{1}\{\psi(u-l) \le b(u-l)/a\} \\ \le P^\star \exp\{a\psi(u-l)\} + P^\star \exp\{b(u-l)\}\mathbb{1}\{\psi(u-l)/(u-l) \le b/a\}, \end{aligned}$$

and the right hand term above is finite, *primo* by choice of a, *secundo* because the set in the indicator function is bounded. □

Finally, concerning Assumption **U3**, its verification typically relies on the application of Ascoli's theorem.

The two above points will be illustrated thanks to the examples in Section 7.7.

Proof of Theorem 7.5.3

• Naturally, $\mathrm{P}(\widehat{K}_n^{\mathrm{L}} < K^\star) = \sum_{K<K^\star} \mathrm{P}(\widehat{K}_n^{\mathrm{L}} = K)$ and Lemma 1.2.15 of (Dembo and Zeitouni 1998) ensures that

$$\limsup_{n\to\infty} n^{-1} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} < K^\star) = \sup_{K<K^\star} \limsup_{n\to\infty} n^{-1} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} = K).$$

Thus, it suffices to choose $K < K^\star$ and to prove that $\limsup_n n^{-1} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} = K)$ is bounded above by $-I(\Lambda_{0,K} \,|P^\star)$.

Now, for any $\alpha > 0$, Assumption **P** on the penalty function yields

$$\limsup_{n\to\infty} n^{-1} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} = K) \leq$$

$$\limsup_{n\to\infty} n^{-1} \log \mathrm{P}^\star\Big(\sup_{\theta\in\Theta_K} \ell_n(\theta) - \mathrm{pen}(n,K) \geq \sup_{\theta\in\Theta_{K+1}} \ell_n(\theta) - \mathrm{pen}(n,K^\star)\Big)$$

$$\leq \limsup_{n\to\infty} n^{-1} \log \mathrm{P}(\mathbb{P}_n \in \Lambda_{\alpha,K}).$$

Furthermore, Lemma 7.5.12 ensures that $\{\Lambda_{\alpha,K}\}$ nonincreases as $\alpha \downarrow 0$, so that $\{I(\Lambda_{\alpha,K} \,|P^\star)\}$ nondecreases as $\alpha \downarrow 0$, and it is bounded above by $I(\Lambda^a_{0,K} \,|P^\star) = H(\Lambda^a_{0,K} \,|P^\star)$ thanks to Proposition 6.2.3 of Interlude 6. Observe that Lemmas 7.5.5 and 7.5.11 finally ensure that the former upper bound is bounded itself by $H(\Pi_{K^\star-1} \,|P^\star)$, which is finite. Therefore, $\{I(\Lambda_{\alpha,K} \,|P^\star)\}$ converges as $\alpha \downarrow 0$ to some limit denoted L, $L \leq I(\Lambda_{0,K} \,|P^\star)$.

Let $\{\alpha_p\}$ be a decreasing sequence of positive numbers such that the sequence $\{I(\Lambda_{\alpha_p,K} \,|P^\star)\}$ increases to L. Since $I(\cdot \,|P^\star)$ is a good rate function and $\Lambda_{\alpha_p,K}$ is closed (see Lemma 7.5.13 and recall that a good lower semicontinuous function achieves its infimum on closed sets), there exists $Q_p \in \Lambda_{\alpha_p,K}$ such that $I(Q_p \,|P^\star) = I(\Lambda_{\alpha_p,K} \,|P^\star)$.

Let $\mathcal{K}$ be the compact subset of $\mathcal{Q}$ consisting of the Q's with $I(Q \,|P^\star) \leq H(\Pi_{K^\star-1} \,|P^\star)$. For any $q \geq 1$, the set $\mathrm{cl}(\{Q_p : p \geq q\})$ is a closed subset of $\Lambda_{\alpha_q,K} \cap \mathcal{K}$ which is compact as a closed subset of a compact set: thus $\{\mathrm{cl}(\{Q_p : p \geq q\})\}_q$ is a nonincreasing sequence of non void compact sets. The Borel Lebesgue property therefore ensures that their intersection is non void either: let $\overline{Q}$ be in the intersection.

Since $\overline{Q} \in \Lambda_{\alpha_p,K}$ for any $p \geq 1$, $\overline{Q} \in \Lambda_{0,K}$. Moreover, $I(Q_p \,|P^\star) \leq L$ (any $p \geq 1$) yields that $\{Q_p : p \geq q\}$ is included in the closed set of the Q's of $\mathcal{Q}$ with $I(Q \,|P^\star) \leq L$, and so does its closure. Forwardly, $\overline{Q}$ also satisfies $I(\overline{Q} \,|P^\star) \leq L$. Then $L \leq I(\Lambda_{0,K} \,|P^\star) \leq I(\overline{Q} \,|P^\star)$ concludes the main part of this proof.

It still remains to show that $I(\overline{Q} \,|P^\star) > 0$. Suppose on the contrary that equality holds true. Then

$$\sup\{\overline{Q}^s f : f \in \mathcal{L}_\tau, P^\star \exp(f) < \infty\} = 0. \tag{7.20}$$

Choose $f \in \mathcal{L}_\tau$: there exists $a > 0$ such that $P^\star \exp(a|f|) < \infty$. Then $\overline{Q}^s a|f| \geq 0$ (recall that $\overline{Q}^s \geq 0$ as observed in Remark 6.2.2 of Interlude 6) and equality does hold thanks to (7.20). Consequently, $\overline{Q}^s|f| = \overline{Q}^s f \vee 0 - \overline{Q}^s f \wedge 0 = 0$ and $\overline{Q}^s \geq 0$ implies that $\overline{Q}^s f \vee 0 = \overline{Q}^s f \wedge 0 = 0$, hence $\overline{Q}^s f = 0$ as their sum.

Thus, $\overline{Q} = \overline{Q}^a$. Following Remark 6.2.2 of Interlude 6, we can think of $\overline{Q}$ as a probability measure satisfying $H(\overline{Q} \,|P^\star) = I(\overline{Q} \,|P^\star) = 0$, *i.e.* $\overline{Q} = P^\star$. This contradicts Lemma 7.5.15 and concludes the proof for the estimator $\widehat{K}_n^{\mathrm{L}}$.

- The proof for the estimator $\widehat{K}_n^{\mathrm{G}}$ is almost identical, so we omit it.

7.5.4. Proofs, continued: Theorem 7.5.6

We present hereafter the proof of Theorem 7.5.6. Actually, the result of the latter stems from a more precise identification of the rate obtained in the first theorem, thanks to some more assumptions.

We start with a lemma:

Lemma 7.5.17. *If $Q \in \mathcal{Q} \cap L'_\tau$, then the function $\theta \mapsto Q\,\ell_\theta$ mapping $\Theta_{K^\star}$ to $\mathbb{R}$ is continuous over $\Theta_{K^\star}$ as soon as Assumptions* **U1** *part* (iii) *and* **U2–U4** *are satisfied.*

Proof. The function is well defined thanks to Assumption **U1** part (iii) and the inclusions (6.2). Since $Q \in L'_\tau$, it suffices to prove that $\|\ell_{\theta_q} - \ell_{\theta_0}\|_\tau \to 0$ when $\theta_q \to \theta_0$ $(\theta_q, \theta_0 \in \Theta_{K^\star})$.

In virtue of the definition of the norm $\|\cdot\|_\tau$, we have

$$\begin{aligned}\|\ell_{\theta_q} - \ell_{\theta_0}\|_\tau &= \inf\left\{a > 0 : P^\star\tau\left(a^{-1}\log\frac{p_{\theta_q}}{p_{\theta_0}}\right) \le 1\right\} \\ &\le \inf\left\{a > 0 : P^\star \exp\left(a^{-1}\left|\log\frac{p_{\theta_q}}{p_{\theta_0}}\right|\right) \le 2\right\}.\end{aligned}$$

Set $a > 0$, $\varepsilon > 0$ such that $e^{\varepsilon/a} \le 3/2$ and a compact set C of $\mathcal{Y}$ satisfying both $C_{lu} \subset C$ and $P^\star(C^c) < 1/2$ (where C_{lu} is defined in Assumption **U4**). Now,

$$\begin{aligned}P^\star \exp\left(a^{-1}\left|\log\frac{p_{\theta_q}}{p_{\theta_0}}\right|\right)\mathbb{1}\{C^c\} &= P^\star\left(\left(\frac{p_{\theta_q}}{p_{\theta_0}}\right)^{a^{-1}}\mathbb{1}\{p_{\theta_q} \ge p_{\theta_0}\} + \left(\frac{p_{\theta_0}}{p_{\theta_q}}\right)^{a^{-1}}\mathbb{1}\{p_{\theta_q} < p_{\theta_0}\}\right)\mathbb{1}\{C^c\} \\ &\le P^\star\left(\left(\frac{p_{\theta_q}}{p_{\theta_0}}\right)^{a^{-1}} + \left(\frac{p_{\theta_0}}{p_{\theta_q}}\right)^{a^{-1}}\right)\mathbb{1}\{C^c\}.\end{aligned}$$

Besides, Assumption **U3** ensures that the parameterization $\theta \mapsto p_\theta$ is pointwise continuous and therefore, $p_{\theta q} \to p_{\theta_0}$ pointwise as $q \uparrow \infty$ and the integrand of the right hand term of the previous inequality converges pointwise to $2\,\mathbb{1}\{C^c\}$. Furthermore, Assumption **U2** yields that p_θ/p_t is bounded above by $\exp(u-l)$ for any $\theta, t \in \Theta_K$, hence

$$\left(\frac{p_\theta}{p_t}\right)^{a^{-1}}\mathbb{1}\{C^c\} \le \exp\left\{a^{-1}(u-l)\right\}\mathbb{1}\{C^c\} \le \exp\left\{a^{-1}(u-l)\,\mathbb{1}\{C^c\}\right\},$$

and the right hand term in the former inequality is an element of $L^1(P^\star)$ thanks to Assumption **U4**. Thus, we conclude by dominated convergence that $P^\star \exp\left(a^{-1}\left|\log\frac{p_{\theta q}}{p_{\theta_0}}\right|\right)\mathbb{1}\{C^c\}$ is bounded above by $1/2$ for q large enough.

The pointwise continuous parameterization property together with Assumption **U3** ensure that the supremum of $|\ell_{\theta_q}(y) - \ell_{\theta_0}(y)|$ over $y \in C$ is bounded above by ε for q large enough. Consequently,

$$\begin{aligned}P^\star \exp\left(a^{-1}\left|\log\frac{p_{\theta_q}}{p_{\theta_0}}\right|\right)\mathbb{1}\{C\} &= P^\star\left(\left(\frac{p_{\theta_q}}{p_{\theta_0}}\right)^{a^{-1}}\mathbb{1}\{p_{\theta_q} \ge p_{\theta_0}\} + \left(\frac{p_{\theta_0}}{p_{\theta_q}}\right)^{a^{-1}}\mathbb{1}\{p_{\theta_q} < p_{\theta_0}\}\right)\mathbb{1}\{C\} \\ &\le e^{\varepsilon/a} \le 3/2\end{aligned}$$

and finally $P^\star \exp\left(a^{-1}\left|\log\frac{p_{\theta q}}{p_{\theta_0}}\right|\right) \le 2$, for q large enough. This concludes the proof, since $a > 0$ is arbitrarily small. □

Lemma 7.5.18. *Suppose that Assumptions* **U1** (iii) *and* **U5** *hold true. Set $Q \in \mathcal{Q}$ and $K \le K^\star$. Then the function $\theta \mapsto Q\,\ell_\theta$ mapping Θ_K to $\mathbb{R}$ is differentiable on the interior of Θ_K, with derivative $\theta \mapsto Q\,\dot{\ell}_\theta$.*

Proof. The function is well defined thanks to Assumption **U1** (iii). Moreover, since $Q \in \mathcal{Q}$ satisfies $Q \geq 0$, Assumption **U5** immediately yields

$$\left| Q\,\ell_{\theta+h} - Q\,\ell_\theta - Q\,\dot{\ell}_\theta^{\,T} h \right| \leq (Q\,F)\,o(h).$$

□

We can now finish the proof of Theorem 7.5.6.

Proof. (of Theorem 7.5.6*)* Let us start where we stopped in the proof of Theorem 7.5.3. Since $\overline{Q} \in L'_\tau$, so does $\overline{Q}^s$. Apply Lemma 7.5.18: $\theta \mapsto \overline{Q}^s\,\ell_\theta$ is differentiable on the interiors of Θ_K and Θ_{K+1}, with derivative $\overline{Q}^s\,\dot{\ell}_\theta$, which is zero since $\dot{\ell}_\theta \in \mathcal{M}_\tau$ in virtue of Assumption **U5**. So, $\overline{Q}^s\,\ell_\theta$ is constant over Θ_K and Θ_{K+1} with possibly distinct values on each of them. Now, Lemma 7.5.17 yields that $\overline{Q}^s\,\ell_\theta = \overline{Q}^s\,\ell_t$ for any $\theta \in \Theta_K$, $t \in \Theta_{K+1}$.

Finally, $\overline{Q}^a \in \Lambda^a_{0,K}$, Proposition 6.2.3 and the obvious inequality $I(Q\,|P^\star) \geq I(Q^a\,|P^\star)$ (any $Q \in \mathcal{Q} \cap L'_\tau$) give

$$H(\Lambda^a_{0,K}\,|P^\star) \leq H(\overline{Q}^a\,|P^\star) = I(\overline{Q}^a\,|P^\star) \leq I(\overline{Q}\,|P^\star) = I(\Lambda_{0,K}\,|P^\star) \leq H(\Lambda^a_{0,K}\,|P^\star)$$

and this completes the proof for the estimator $\widehat{K}^{\mathrm{L}}_n$. Once again, the proof when studying $\widehat{K}^{\mathrm{G}}_n$ is very similar, so we again omit it. □

7.5.5. Proofs, end: Proposition 7.5.9

Set $K < K^\star$. We can suppose that we have constructed a probability measure $\overline{Q}^a \in M_1(\mathcal{Y}) \cap L'_\tau$ such that

$$H(\overline{Q}^a\,|P^\star) = H(\Gamma^a_{0,K}\,|P^\star) < \infty$$

and we wonder whether $\overline{Q}^a \in \Pi_K$, which is a sufficient condition for achievement by $\widehat{K}^{\mathrm{G}}_n$ of the best rate with a view to Proposition 7.5.1, since then

$$H(\overline{Q}^a\,|P^\star) = H(\Pi_K\,|P^\star)$$

and K is arbitrarily lower than $K^\star$.

The main idea of the proof is quite simple: we try to take advantage of the nice Pythagorean characterization of the H-projection on convex sets. By H-projection of P on a convex set $\mathcal{C}$, we mean a probability $\overline{Q}$ such that

$$H(\overline{Q}\,|P) = \inf_{Q \in \mathcal{C}} H(Q\,|P).$$

Indeed, the following result holds true (it is a simple corollary of Theorem 2.2 in Csiszár 1975):

Proposition 7.5.19 (Csiszár). *Let $\mathcal{C}$ be a subset of $M_1(\mathcal{Y})$. A probability distribution $\overline{Q} \in M_1(\mathcal{Y})$ is the H-projection of $P \in M_1(\mathcal{Y})$ if and only if, for all $Q \in \mathcal{C}$,*

$$H(Q\,|P) \geq H(Q\,|\overline{Q}) + H(\overline{Q}\,|P).$$

We also try to use the equivalent characterization that can be derived in exponential models when the projection is meant with reversed order of P and Q in the definition above, *i.e.*

$$H(Q\,|\overline{P}) = \inf_{P\in\mathcal{C}} H(Q\,|P).$$

Such a projection $\overline{P}$ will be called in the sequel "reversed order projection" of Q on $\mathcal{C}$.

We first state and prove some useful lemmas.

Lemma 7.5.20. *$\overline{Q}^a$ is dominated by μ, with a density wrt μ denoted $\bar{q}$ such that $\overline{Q}^a \log \bar{q} < \infty$.*

Proof. On the one hand,

$$\overline{Q}^a \log(\bar{q}/p^\star) = H(\overline{Q}^a\,|P^\star) < \infty$$

and particularly, $\overline{Q}^a << P^\star << \mu$. On the other hand,

$$\overline{Q}^a \log p^\star \leq \sup_{\theta\in\Theta_{K^\star}} \overline{Q}^a\,\ell_\theta < \infty.$$

Indeed, the function $\theta \mapsto \overline{Q}^a\,\ell_\theta$ is continuous from $\Theta_{K^\star}$ to $\mathbb{R}$ (a consequence of Lemma 7.5.17) and $\Theta_{K^\star}$ is compact.

We conclude thanks to the two inequalities above. □

Lemma 7.5.21. *There exists a probability $\overline{P} \in \Pi_K$ with density denoted $\bar{p}$ such that $H(\overline{Q}^a\,|\overline{P}) = H(\overline{Q}^a\,|\Pi_K) = H(\overline{Q}^a\,|\Pi_{K^\star}) < \infty$.*

Proof. Since $H(\overline{Q}^a\,|P^\star)$ is finite, so is $H(\overline{Q}^a\,|\Pi_{K^\star})$. Furthermore, $\overline{Q}^a \in \Gamma_{0,K}$ and $\overline{Q}^a \log \bar{q} < \infty$, hence

$$-\sup_{\theta\in\Theta_K} \overline{Q}^a\,\ell_\theta + \overline{Q}^a \log \bar{q} = -\sup_{\theta\in\Theta_{K^\star}} \overline{Q}^a\,\ell_\theta + \overline{Q}^a \log \bar{q} = H(\overline{Q}^a\,|\Pi_{K^\star})$$

and forwardly the left hand expression equals $H(\overline{Q}^a\,|\Pi_K)$ and is finite, too.

The existence of $\overline{P}$ is a consequence of the compactness of Π_K and of the lower semicontinuity of $H(\overline{Q}^a\,|\cdot)$. □

Let us introduce now the sets

$$\begin{aligned}
\mathcal{C}_1 &= \Big\{Q \in M_1(\mathcal{Y}) : Q << \mu,\ Q = q\,\mu \text{ and } Q \log q < \infty\Big\},\\
\mathcal{C}_2 &= \Big\{Q \in M_1(\mathcal{Y}) : Q \log \bar{p} = \sup_{\theta\in\Theta_{K^\star}} Q\,\ell_\theta\Big\},\\
\mathcal{C}_3 &= \Big\{Q \in M_1(\mathcal{Y}) : H(Q\,|\overline{P}) < \infty\Big\} \text{ and}\\
\mathcal{C} &= \mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_3 \cap L'_\tau.
\end{aligned}$$

Lemma 7.5.22. *The following four properties hold true:*

(i) *$\mathcal{C}$ is a convex set;*

(ii) *$\overline{Q}^a \in \mathcal{C}$;*

(iii) $\mathcal{C} \subset \Gamma_{0,K}$;

(iv) *for all* $Q \in \mathcal{C}$, $H(Q\,|\overline{P}) = H(Q\,|\Pi_{K^\star}) < \infty$.

Proof. • The convexity of $\mathcal{C}_1$ stems from the convexity of $x \mapsto x \log x$ ($x \in \mathbb{R}_+$) and from elementary properties of domination wrt μ. The convexity of $\mathcal{C}_2$ is readily proved by linearity of $Q \mapsto Q\,\ell_\theta$. The convexity of $\mathcal{C}_3$ derives from the convexity of $Q \mapsto H(Q\,|\overline{P})$. Finally, $\mathcal{C}$ is a convex set as an intersection of three convex sets and a vector space.
• $\overline{Q}^a \in \mathcal{C}_1$ in virtue of Lemma 7.5.20. $\overline{Q}^a \in \mathcal{C}_3$ thanks to Lemma 7.5.21. Now, Lemmas 7.5.20 and 7.5.21 yield that $\overline{Q}^a \log \bar{q} = \sup_{\theta\in\Theta_K} \overline{Q}^a\,\ell_\theta$. Besides, $\overline{Q}^a \in \Gamma_{0,K}$ allows to conclude that $\overline{Q}^a$ also belongs to $\mathcal{C}_2$. Finally, $\overline{Q}^a \in L'_\tau$ in virtue of its previous construction, hence (ii).
• Choose $Q \in \mathcal{C}$. One has

$$\sup_{\theta\in\Theta_{K^\star}} Q\,\ell_\theta = Q \log \bar{p} \le \sup_{\theta\in\Theta_K} Q\,\ell_\theta \le \sup_{\theta\in\Theta_{K^\star}} Q\,\ell_\theta,$$

hence $Q \in \Gamma_{0,K}$.
• Choose $Q \in \mathcal{C}$: since $H(Q\,|\overline{P})$ and $Q \log q$ are finite, we can decompose the Kullback-Leibler divergence and apply the condition which appears in the definition of $\mathcal{C}_2$,

$$H(Q\,|\overline{P}) = Q \log(q/\bar{p}) = Q \log q - Q \log \bar{p} = Q \log q - \sup_{\theta\in\Theta_{K^\star}} Q\,\ell_\theta = H(Q\,|\Pi_{K^\star}),$$

hence the conclusion. □

The next lemma summarizes the first part of the proof of Proposition 7.5.9 (observe that we do not have used yet the exponential model assumption):

Lemma 7.5.23. *$\overline{Q}^a$ is the H-projection of $P^\star$ on the convex set $\mathcal{C}$. Particularly, the Pythagorean characterization of Proposition 7.5.19 holds true for $\overline{Q}^a$.*

Proof. It is readily proved:

$$H(\mathcal{C}\,|P^\star) \le H(\overline{Q}^a\,|P^\star) = H(\Gamma_{0,K}\,|P^\star) \le H(\mathcal{C}\,|P^\star),$$

where the first inequality stems from the fact that $\overline{Q}^a \in \mathcal{C}$, the next equality is by definition of $\overline{Q}^a$ and the last inequality derives from the fact that $\mathcal{C} \subset \Gamma_{0,K}$, in virtue of Lemma 7.5.22. Hence all the inequalities above are equalities, which completes the proof. □

Let us now describe the *exponential models* Π_K. Let $t = (t_1, \ldots, t_{K^\star})$ be a known function on $\mathcal{Y} \subset \mathbb{R}^q$ equipped with the Lebesgue measure μ on Borel sets. Let $h \in \mathbb{R}_+^{\mathcal{Y}}$ and Θ be the (convex and supposed open) natural set of parameters, *i.e.*

$$\Theta = \left\{\theta \in \mathbb{R}^{K^\star} : \mu(he^{\theta^T t}) < \infty\right\}.$$

We denote $\phi(\theta) = \log \mu(he^{\theta^T t})$, which defines a convex function on Θ.

We assume that Θ_K is a convex set strictly included in $\mathrm{int}(\Theta)$ and we define Π_K as the class of the distributions $P_\theta = p_\theta\,\mu$ for

$$p_\theta(y) = h(y)e^{\theta^T t(y) - \phi(\theta)} \text{ (any } y \in \mathcal{Y}).$$

We finally emphasize that ϕ is differentiable on Θ with gradient $\dot{\phi}(\theta) = P_\theta\, t$.

Let us choose $Q \in \mathcal{C}$. Lemma 7.5.22 (iv) ensures that $\overline{P}$ is its reversed order projection on $\Pi_{K^\star}$. The point is to prove that an analogue to the Pythagorean characterization for the H-projections is also satisfied when considering reversed order projections. Indeed

Lemma 7.5.24. *For all $Q \in \mathcal{C}$, for all $P_\theta \in \Pi_{K^\star}$, one has*

$$H(Q\,|P_\theta) \geq H(Q\,|\overline{P}) + H(\overline{P}\,|P_\theta). \tag{7.21}$$

The proof of the latter heavily relies on the exponential model assumption.

Proof. • We first emphasize that $\bar{p}$ satisfies an interesting minimization property. Indeed, since $Q \log q < \infty$ and $H(Q\,|\overline{P}) < \infty$, we can decompose the Kullback-Leibler divergence and get $H(Q\,|\overline{P}) = Q\,\log q - Q\,\log \bar{p}$. Furthermore, since $Q \in L'_\tau$ and all ℓ_θ belong to $\mathcal{L}_\tau$, we have

$$H(Q\,|P_\theta) = Q\,\log(q/p_\theta) = Q\,\log q - Q\,\ell_\theta \quad (\text{any } \theta \in \Theta_{K^\star}). \tag{7.22}$$

The equality $H(Q\,|\overline{P}) = H(Q\,|\Pi_{K^\star}) < \infty$ finally yields

$$0 \leq Q\,\log\frac{\bar{p}}{p_\theta} < \infty \ (\text{any } \theta \in \Theta_{K^\star}). \tag{7.23}$$

• Moreover, Inequality (7.21) is obviously satisfied when the left hand term is infinite. So, let us define the set Θ_c of all the parameters θ such that $H(Q\,|P_\theta) < \infty$: Θ_c is an open convex set and it suffices to verify whether Inequality (7.21) is fulfilled on Θ_c.

Indeed, Θ_c is open thanks to (7.22) and Lemma 7.5.17. For the convexity of Θ_c, set $\theta_1, \theta_2 \in \Theta_c$, $\lambda \in [0,1]$ and $\theta = \lambda\theta_1 + (1-\lambda)\theta_2$. One has simply

$$\begin{aligned} H(Q\,|P_\theta) &= Q\,(\log q - \log p_\theta) = Q\Big(\log q - \log h - \theta^T t + \phi(\theta)\Big) \\ &\leq Q\Big(\lambda \log(q/p_{\theta_1}) + (1-\lambda)Q(\log(q/p_{\theta_2}))\Big) = \lambda H(Q\,|P_{\theta_1}) + (1-\lambda)H(Q\,|P_{\theta_2}) < \infty \end{aligned}$$

in virtue of the linearity of the scalar product and of the convexity of ϕ, hence the result.

• Now, observe that $Q \log q < \infty$ and $H(Q\,|\overline{P}) < \infty$ (because $Q \in \mathcal{C}$) ensures that Inequality (7.21) is equivalent to

$$(Q - \overline{P})\log\frac{\bar{p}}{p_\theta} \geq 0 \quad (\text{any } \theta \in \Theta_c),$$

which is in turn equivalent to the following, thanks to the exponential model assumption (with notation $\overline{P} = P_{\bar{\theta}}$):

$$(\bar{\theta} - \theta)^T (Q - \overline{P})\, t \geq 0 \quad (\text{any } \theta \in \Theta_c). \tag{7.24}$$

• Finally, Inequality (7.24) is a straightforward consequence of Inequality (7.23) and of Lemma 7.5.25. Let f be the function on Θ_c defined for any $\theta \in \Theta_c$ by

$$f(\theta) = Q\,\log\frac{\bar{p}}{p_\theta} = (\bar{\theta} - \theta)^T\, Q\, t + \phi(\theta) - \phi(\bar{\theta}).$$

Then the convexity of ϕ and the property stated in (7.23) yield that f is a proper convex function on the convex set Θ_c. Besides, f is differentiable at $\bar{\theta}$, with gradient $\dot{f}(\bar{\theta}) = (\overline{P} - Q)\, t$. Finally, f attains a minimum in $\bar{\theta}$ in virtue of (7.23), hence

$$(\theta - \bar{\theta})^T \left(-\dot{f}(\bar{\theta})\right) = (\theta - \bar{\theta})^T (Q - \overline{P})\, t \leq 0$$

for any $\theta \in \Theta_c$, which is exactly the condition (7.24). This concludes the proof. □

We used in the proof the following simple corollary of Theorem 27.4 in (Rockafellar 1970):

Lemma 7.5.25 (Rockafellar). *Let f be a proper convex function,* i.e. *a convex function whose domain* $\mathrm{dom}(f) = \{x \in \mathbb{R}^q : f(x) < \infty\}$ *is nonempty, and such that the restriction of f to* $\mathrm{dom}(f)$ *is finite. Assume that* $\mathrm{dom}(f)$ *is an open set and let C be a nonempty, open convex set that intersetcs* $\mathrm{dom}(f)$. *In order that x be a point where the infimum of f relative to C is attained, with the additional assumption that f is differentiable at x with gradient $\dot{f}(x)$, it is necessary and sufficient that $-\dot{f}(x)$ is normal to C at x,* i.e. *that for any $y \in C$,*

$$(y - x)^T \left(-\dot{f}(x)\right) \leq 0.$$

The proof of Proposition 7.5.9 is now at hand.

Proof. (*of Proposition* 7.5.9) Let Q be an element of $\mathcal{C}$. Then Inequality (7.21) of Lemma 7.5.24 applied to $P^\star \in \Pi_{K^\star}$ yields

$$H(Q\,|P^\star) \geq H(Q\,|\overline{P}) + H(\overline{P}\,|P^\star).$$

Since the latter is true for all $Q \in \mathcal{C}$, $\overline{P}$ is the H-projection of $P^\star$ on the convex set $\mathcal{C}$ (thanks to Proposition 7.5.19), *i.e.* $\overline{Q}^a = \overline{P}$ and particularly, $\overline{Q}^a \in \Pi_K$, which is the final conclusion. □

7.6. Overestimation

This section copes with the problem of overestimation for our estimators $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$. Another application of Stein's lemma (see Section 7.5.1) yields that the rates of convergence of the overestimation probabilities are slower than exponential in n. It is then shown that both overestimation probabilities of $\widehat{K}_n^{\mathrm{L}}$ and $\widehat{K}_n^{\mathrm{G}}$ converge to zero *e.g.* faster than exponential in $n^{1-\delta}$ for any $0 < \delta < 1$. The latter result nevertheless requires an upper bound on $K^\star$ when considering $\widehat{K}_n^{\mathrm{G}}$. An à la Huber trick once again enhances the previous results (in terms of extension of the range of allowed penalty functions) but at the price of expensive (too expensive ?) additional assumptions.

7.6.1. Stein's lemma for a lower bound on the rate

In that subsection, we shall derive in Proposition 7.6.1 an universal lower bound for the rate of overestimation of any estimator that *does not almost surely underestimate the order* (see Assumption **S2** below). Once again, Stein's lemma and change of probability argument yield the result. We omit the proof of Proposition 7.6.1, for it is alike the proof of Proposition 7.5.1.

So, let $\widetilde{K}_n$ denote any estimator of the order and Assumption **S2** be

S2 For all $K \geq 2$ and $P_\theta \in \Pi_K \setminus \Pi_{K-1}$,

$$\limsup_{n\to\infty} \mathrm{P}_{P_\theta}(\widetilde{K}_n < K) < 1.$$

Now,

Proposition 7.6.1 (yet another version of Stein's lemma, overestimation case). *Let $\widetilde{K}_n$ be any estimator of the true order $K^\star$ of $P^\star$. Suppose that Assumption* **S2** *holds true and that, for all $\theta \in \Theta_{K^\star+1}$, $\ell_\theta, \ell_{\theta^\star} \in L^1(P_\theta)$. Then*

$$\liminf_{n\to\infty} n^{-1} \log \mathrm{P}_{P^\star}(\widetilde{K}_n < K^\star) = \limsup_{n\to\infty} n^{-1} \log \mathrm{P}_{P^\star}(\widetilde{K}_n < K^\star) = 0.$$

This result says that the rate of convergence is slower than exponential in n.

7.6.2. Upper bounds on the rate

Exponential rate of overestimation

Let us set the assumptions we shall need in the theorem (**O** stands for Overestimation).

Oa Let $K > K^\star$ be an integer to be chosen later. The following class of functions is $P^\star$-Donsker:

$$\mathcal{G} = \Big\{ g_\theta = (\ell_\theta - \ell_{\theta^\star}) : \theta \in \Theta_K \Big\}.$$

Besides, there exist $l, u \in \mathbb{R}^{\mathcal{Y}}$ such that $(u - l)$ admits some exponential moments, *i.e.* $(u - l) \in \mathcal{L}_\tau$, with

$$l \leq \ell_\theta \leq u \quad (\text{any } \theta \in \Theta_{K^\star+1}),$$

hence particularly, the family of functions $g_\theta \in \mathcal{G}$ admits an envelope function $(u-l) \in \mathcal{L}_\tau$.

P3a The penalty function is of the form $\mathrm{pen}(n, K) = v_n D(K)$ for $D \in \mathbb{R}^{\mathbb{N}}$ increasing and $\{v_n\}$ also increasing such that $n^{1/2} v_n^{-1} = o(1)$, $v_n = o(n)$ and finally,

$$v_{nk} \leq A k^{1-\delta} v_n$$

for some $A \geq 1$, $0 < \delta < 1$, any $n, k \geq 1$.

Remark 7.6.2 (on the penalty function). Assumption **P3a** restricts for sake of clarity the class of penalty functions by imposing the form $v_n\, D(K)$ though the scheme of proof of the theorem still applies with the more general penalty functions, up to slight changes. Typical examples of suitable sequences $\{v_n\}$ include the family of

$$v_n = n^{1-\delta} (\log n)^\beta$$

for any $0 < \delta < 1/2$ and $\beta \geq 0$.

Theorem 7.6.3 (exponential rate of overestimation). *Under Assumptions* **P3a***:*
• *if Assumption* **Oa** *holds true for* $K = K^\star + 1$*, then*

$$\limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} > K^\star) < 0;$$

• *if we know a prior bound* $K_{\max}$ *on* $K^\star$ *and Assumption* **Oa** *holds true for* $K = K_{\max}$*, then*

$$\limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}(\widehat{K}_n^{\mathrm{G}} > K^\star) < 0.$$

Remark 7.6.4.
• The theorem above applies to both the **MGM** and **VR** examples, see Section 7.7.
• For comparison with (Dacunha-Castelle and Gassiat 1997), see Remark 7.5.4. Indeed, the authors treat the under- and overestimation cases together, hence the exponential rate in $n^{-1}v_n^2$. We emphasize that this rate is precisely the lowest of the underestimation and overestimation rates exhibited by us in Theorems 7.5.6 and 7.6.3.

The latter paper does not require any prior bound on $K^\star$, contrarily to (Guyon and Yao 1999; Haughton and Keribin 2001). Guyon and Yao again prove a nonasymptotic result thanks to an exponential inequality, but under the factorization assumption presented in Remark 7.5.4.

The proof of Haughton and Keribin relies on a Taylor expansion and then on moderate and large deviations principles (respectively for the leading and the remainder terms), hence some strong assumptions of three times continuously differentiability of the log-likelihood, existence of some exponential moments for the gradient and for the supremum of the Hessian. Those conditions again limit the field of application of their result.

Refined exponential rate of overestimation

Now, the à la Huber trick again enhances the above result when strengthening Assumptions **Oa**, but relaxing Assumption **P3a**:

Ob Let $K > K^\star$ be an integer to be chosen later. The following class of functions is $P^\star$-Donsker:

$$\mathcal{G} = \left\{ g_\theta = \frac{\ell_\theta - \ell_{\theta^\star}}{H(\theta)^{1/2}} : \theta \in \Theta_K,\ H(\theta) > 0 \right\}.$$

Besides, the family of functions $g_\theta \in \mathcal{G}$ admits an envelope function G with some exponential moments, *i.e.* $G \in \mathcal{L}_\tau$.

P3b The penalty function is of the form $\mathrm{pen}(n, K) = v_n D(K)$ for $D \in \mathbb{R}^{\mathbb{N}}$ increasing and $\{v_n\}$ increasing to infinity such that both $(\log n)\, {v_n}^{-1} = o(1)$ and $v_n = o(n)$ with moreover

$$v_{nk} \leq A k^{1-\delta} v_n$$

for some $A \geq 1$, $0 < \delta < 1$, any $n, k \geq 1$.

Theorem 7.6.5 (refined exponential rate of overestimation). *Under Assumptions* **P3b***:*
• *if Assumption* **Ob** *holds true for* $K = K^\star + 1$*, then*

$$\limsup_{n\to\infty} v_n^{-1} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} > K^\star) < 0;$$

• *if we know a prior bound* $K_{\max}$ *on* $K^\star$ *and Assumption* **Ob** *holds true for* $K = K_{\max}$*, then*

$$\limsup_{n\to\infty} v_n^{-1} \log \mathrm{P}(\widehat{K}_n^{\mathrm{G}} > K^\star) < 0.$$

Remark 7.6.6. It occurs this time that the reinforcement of Theorem 7.6.3 that yields Theorem 7.6.5 is *expensive* whereas the improvement of the consistency result of Theorem 7.4.3 stated in Theorem 7.4.5 is not.

Indeed, for the latter, the original condition **C5a** yielded **C5b** that our benchmark example **MG** both satisfies (see Section 7.7). Thus, this transformation may be qualified of *constraining* with regard to the second transformation above of a **Oa** into **Ob**, since particularly, the **MG** example unfortunately satisfies the sole assumption **Oa**. Actually, the more typical example where both the original and the transformed assumptions are satisfied seems to be the case of bounded ℓ_θ.

7.6.3. Proof

The proof parallels that of consistency (see Theorems 7.4.3 and 7.4.5). The main difference is the application of a MDP theorem rather than a bounded LIL theorem in order to get rates of convergence. More precisely, the above mentioned MDP is namely Theorem 6.3.2, due to Wu – see the discussion in Interlude 6, Section 6.3.

The two proofs of the theorems above are very similar, thus we shall only present the first one, including remarks when the second proof would slightly differ. The scheme is simple: *primo*, apply Proposition 7.4.10 in order to translate the overestimation problem into a deviations problem; *secundo*, apply the MDP of Theorem 6.3.2 of Interlude 6; *tertio*, conclude.

Proof. (of Theorem 7.6.3*)* • Let us denote $C^\star = D(K^\star + 1) - D(K^\star) > 0$ and apply the first point of Proposition 7.4.10 (the second point when proving the refined theorem):

$$\begin{aligned}
\mathrm{P}(\widehat{K}_n^{\mathrm{L}} > K^\star) &\le \mathrm{P}\left(\sup_{\theta\in\Theta_{K^\star+1}} \mathbb{P}_n \ell_\theta - \sup_{\theta\in\Theta_{K^\star}} \mathbb{P}_n \ell_\theta \ge n^{-1} v_n C^\star \right) \\
&\le \mathrm{P}\left(\sup_{\theta\in\Theta_{K^\star+1}} (\mathbb{P}_n - P^\star)(\ell_\theta - \ell_{\theta^\star}) \ge n^{-1} v_n C^\star \right) \\
&\le \mathrm{P}\Big((n\, v_n^{-1})\ (\mathbb{P}_n - P^\star)^\infty \in \Lambda^\infty \Big)
\end{aligned}$$

where $\Lambda^\infty = \Big\{ B \in \ell^\infty(\mathcal{G}) : \|B\|_{\mathcal{G}} \ge C^\star \Big\}$.

Then, Theorem 6.3.2 (whose assumptions are satisfied here of course, see the sufficient conditions (a,b,c) in Section 6.3.3) ensures that

$$\limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} > K^\star) \le -\inf\{ J^\infty(B\,|P^\star) : B \in \Lambda^\infty \},$$

so we have to prove that the right hand term is negative.

We first emphasize that Λ^∞ is closed for the uniform topology on $\ell^\infty(\mathcal{G})$ equipped with $\|\cdot\|_{\mathcal{G}}$ Let us now suppose on the contrary that the infimum is zero. Then, there exists a sequence $\{B_p\}$ of elements of $\ell^\infty(\mathcal{G})$ such that $B_p \in \Lambda^\infty$ and $J^\infty(B_p\,|P^\star) \le p^{-1}$. Consequently, there exists a sequence $\{Q_p\}$ of elements of $M(\mathcal{Y})$ such that both

$$J(Q_p\,|P^\star) \le J^\infty(B_p\,|P^\star) + p^{-1} \le 2\,p^{-1} \quad \text{and} \quad Q_p^\infty = B_p$$

for any $p \geq 1$. In particular, $Q_p << P^\star$ thus, for any $g \in \mathcal{G}$, one has

$$(B_p^\infty\, g)^2 = (Q_p\, g)^2 = \left(P^\star \frac{\mathrm{d}Q_p}{\mathrm{d}P^\star}\, g\right)^2 \leq (P^\star\, g^2)\, P^\star \left(\frac{\mathrm{d}Q_p}{\mathrm{d}P^\star}\right)^2 \leq 2 \left(\sup_{g\in\mathcal{G}} P^\star\, g^2\right)\, J(Q_p\,|P^\star)$$

thanks to Cauchy-Schwarz's inequality. Now, Assumption **Oa** ensures that $\mathcal{G}$ is (totally) bounded in $L^2(P^\star)$, hence

$$\|B_p^\infty\|_{\mathcal{G}} = o(1).$$

We have proved that, if $\inf\{J^\infty(B\,|P^\star) : B \in \Lambda^\infty\} = 0$, then $0 \in \Lambda^\infty$ as a limit of a sequence of elements of the closed set Λ^∞. Since this is false, the infimum must be positive and the proof for $\widehat{K}_n^{\mathrm{L}}$ is complete.

- As for $\widehat{K}_n^{\mathrm{G}}$, we can write

$$\mathrm{P}(\widehat{K}_n^{\mathrm{G}} > K^\star) = \sum_{K=K^\star+1}^{K_{\max}} \mathrm{P}(\widehat{K}_n^{\mathrm{G}} = K),$$

so that Lemma 1.2.15 of (Dembo and Zeitouni 1998) yields the equality below (we denote $C_K^\star = D(K) - D(K^\star)$)

$$\limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}(\widehat{K}_n^{\mathrm{G}} > K^\star) = \sup_{K^\star<K\leq K_{\max}} \limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}(\widehat{K}_n^{\mathrm{G}} = K) \leq$$

$$\sup_{K^\star<K\leq K_{\max}} \limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}\left(\sup_{\theta\in\Theta_K} \mathbb{P}_n \ell_\theta - \sup_{\theta\in\Theta_{K^\star}} \mathbb{P}_n \ell_\theta \geq n^{-1} v_n\, C_K^\star\right),$$

while the inequality stems from the application of the first point of Proposition 7.4.10 (we would apply the second point when proving the refined theorem). The remainder of the proof goes along the same lines than above, so we omit it. □

Alternative uncompleted result

It is possible to apply Theorem 6.3.1 instead of Theorem 6.3.2 in order to use weaker assumptions. In the following proposition, cl denotes the closure wrt the topology induced on $M(\mathcal{Y})$ by the $\sigma(\mathcal{L}_\tau^*, \mathcal{L}_\tau)$-topology (*i.e.* the topology $\mathcal{T}'$ with the notations of Section 6.3 of Interlude 6).

Proposition 7.6.7 (an uncompleted result on overestimation).
Let us suppose that the penalty function is of the form $\mathrm{pen}(n,K) = v_n D(K)$ *for* $D \in \mathbb{R}^{\mathbb{N}}$ *increasing and* $\{v_n\}$ *also increasing such that* $n^{1/2} v_n^{-1} = o(1)$ *and* $v_n = o(n)$. *For any* $K > K^\star$, *let us denote* $C_K^\star = D(K) - D(K^\star)$ *and introduce*

$$\mathcal{C}_K = \mathrm{cl}\left(\left\{Q \in M(\mathcal{Y}) : \sup_{\theta\in\Theta_K} Q(\ell_\theta - \ell_{\theta^\star}) \geq C_K^\star\right\}\right).$$

- *Whenever the functions* ℓ_θ $(\theta \in \Theta_{K^\star+1})$ *admit some exponential moments, or equivalently* $\{\ell_\theta : \theta \in \Theta_{K^\star+1}\} \subset \mathcal{L}_\tau$, *one has*

$$\limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}(\widehat{K}_n^{\mathrm{L}} > K^\star) \leq -J(\mathcal{C}_{K^\star+1}\,|P^\star) \leq 0.$$

• *If we know a prior bound* $K_{\max}$ *on* $K^\star$ *and if the functions* ℓ_θ $(\theta \in \Theta_{K_{\max}})$ *admit some exponential moments, or equivalently* $\{\ell_\theta : \theta \in \Theta_{K_{\max}}\} \subset \mathcal{L}_\tau$, *then one has*

$$\limsup_{n\to\infty} n v_n^{-2} \log \mathrm{P}(\widehat{K}_n^{\mathrm{G}} > K^\star) \leq - \min_{K^\star < K \leq K_{\max}} J(\mathcal{C}_K \,|P^\star) \leq 0.$$

The difficulty is to evaluate the upper bounds and in particular, to prove they are positive. Indeed, the sets

$$\left\{ Q \in M(\mathcal{Y}) : \sup_{\theta \in \Theta_K} Q\,(\ell_\theta - \ell_{\theta^\star}) \geq C_K^\star \right\}$$

are no longer closed for the topology $\mathcal{T}'$. We did not manage to.

7.7. Back to the three examples

This section is devoted to a careful study of the three examples we have already introduced in Section 7.2.2. They were chosen for their own interest as models (the three of them), for the historical attention they were addressed in the field of order estimation (especially **MD** and **VR**) but also for their intrinsic difficulties (especially **MD** and **AC**, with its very general underlying class of partitions).

7.7.1. The mixture of distributions

Short introduction

This example is the more classical of the three we shall consider. It has attracted a lot of attention, see among others (Henna 1985; Leroux 1992; Dacunha-Castelle and Gassiat 1997; Dacunha-Castelle and Gassiat 1999; Keribin 2000; James et al. 2001; Gassiat 2002). For it does not belong to the exponential model family, nor to regular families; for its lack of identifiability when overestimating (and the subsequent degeneration of the Fisher information matrix that avoids classical method based on a Taylor expansion – nonetheless, the already mentioned locally conic parameterization of Dacunha-Castelle and Gassiat allows to cope with Taylor expansions in such situations, method employed by Keribin); for the unboundedness of the support of its distributions; even more, for the fact that the envelope function of the log-densities of the mixture distributions does not admit any exponential moment, this example is both technically very challenging and a benchmark in the order estimation field of research.

The remainder of this section is divided into three subsections, dedicated to the verification of the assumptions related to the results of consistency, underestimation and overestimation, respectively. We shall prove that for the **MGM** example, all results hold true but the refined overestimation Theorem 7.6.5 (for a comment of this deficiency, refer to Remark 7.6.6).

Consistency

Proposition 7.7.1. *Assumptions* **C1–C2** *hold true for the* **MGM**, **RKO** *and* **ME** *examples when* $P^\star$ *is a possibly infinite mixture.*

Proof. First observe that in both cases, $\ell^\star \in L^1(P^\star)$ and **C1** holds true. Concerning **C2**, suppose on the contrary that

$$H(P^\star|\Pi_K) \leq H(P^\star|\Pi_{K+1}), \tag{7.25}$$

i.e. that we have an equality. Lower semicontinuity of $H(P^\star|\cdot)$ and compactness of Π_K ensure the existence of $P_0 \in \Pi_K$ such that $H(P^\star|\Pi_K) = H(P^\star|P_0)$. Then, (7.25) yields that for any $Q \in \Pi_{K+1}$,

$$H(P^\star|Q) \geq H(P^\star|P_0). \tag{7.26}$$

Now let γ denote any element of $\mathcal{D}$ and let us define $Q_h = (1-h)P_0 + h\,\gamma\mu \in \Pi_{K+1}$ (any $h \in [0,1]$). Then $H(P^\star|Q_h) \leq (1-h)H(P^\star|P_0) + h\,H(P^\star|\gamma\mu)$ by a convexity argument, hence $H(P^\star|Q_h)$ is finite. Furthermore, $\ell^\star \in L^1(P^\star)$, so we can decompose the Kullback-Leibler divergences in (7.26) in order to get, for any h,

$$h^{-1}P^\star \log \frac{(1-h)p_0 + h\,\gamma}{p_0} \leq 0,$$

where p_0 is the density of P_0 wrt μ. Applying Fatou's lemma, we derive from the previous inequality that $P^\star(\gamma - p_0)/p_0 \leq 0$. Write F the distribution of the mixture (the support of F may be infinite), so that

$$p^\star(y) = \int_{\mathcal{U}} \gamma_u(y)\,\mathrm{d}F(u).$$

The Fubini theorem ensures that

$$\int_{\mathcal{U}} P^\star(\gamma_u - p_0)/p_0\,\mathrm{d}F(u) = P^\star(p^\star - p_0)/p_0 \leq 0.$$

Finally, $\log x \leq x - 1$ (any $x > 0$) yields

$$0 \leq H(P^\star|P_0) = P^\star \log(p^\star/p_0) \leq P^\star(p^\star/p_0) - 1 \leq 0,$$

and forwardly $P^\star = P_0 \in \Pi_K$. This completes the proof. □

We focus now on Assumptions **C3–C4**:

Proposition 7.7.2. *Assumptions* **C3–C4** *are satisfied for the* **MGM**, **RKO** *and* **ME** *examples.*

Proof. The proof is elementary. We omit the proofs for the **MGM** and **ME** example, which are very similar. Assumption **C4** is immediate. For Assumption **C3**, denote $\underline{m} = \inf \mathcal{M}$, $\overline{m} = \sup \mathcal{M}$, $\underline{s} = \inf \mathcal{S}$, $\overline{s} = \sup \mathcal{S}$ and observe that, for any $v = (m, \sigma^2) \in \mathcal{M} \times \mathcal{S}$, one has for a compact set $\mathcal{C}$ (which depends on $\underline{m}, \overline{m}, \underline{s}, \overline{s}$)

$$\gamma_{(\overline{m},\underline{s})}\,\mathbb{1}\{\mathcal{C}^c\} \leq \gamma_v\,\mathbb{1}\{\mathcal{C}^c\} \leq \gamma_{(\underline{m},\overline{s})}\,\mathbb{1}\{\mathcal{C}^c\},$$

while there exist $a, b > 0$ such that $a\,\mathbb{1}\{\mathcal{C}\} \leq \gamma_v\,\mathbb{1}\{\mathcal{C}\} \leq b\,\mathbb{1}\{\mathcal{C}\}$. Consequently, for any distribution F on $\mathcal{U}$ (possibly nonfinite), we do have for any $y \in \mathbb{R}$

$$l(y) \leq \log\left(\int_{\mathcal{U}} \gamma_v(y)\mathrm{d}F(v)\right) \leq u(y),$$

where $l(y) = \log[a\,\mathbb{1}\{y \in \mathcal{C}\} + \gamma_{(\overline{m},\underline{s})}(y)\,\mathbb{1}\{y \in \mathcal{C}^c\}]$ and $u(y) = [\gamma_{(\underline{m},\overline{s})}(y)\,\mathbb{1}\{y \in \mathcal{C}^c\} + b\,\mathbb{1}\{y \in \mathcal{C}\}]$.

We conclude because $(u - l)$ belongs to $L^1(P^\star)$ as a difference of two elements of $\mathcal{L}_\tau$. □

Remark 7.7.3 (on the envelope function $(u-l)$ in the MGM example).
We actually prove in the **MGM** example that $(u-l) \in \mathcal{M}_\tau \subset \mathcal{L}_\tau$. Indeed, one then has for any $y \in \mathcal{C}^c$

$$2\,\sigma^2\,(u-l)(y) = (\overline{m} - \underline{m})(2y - \overline{m} - \underline{m}). \tag{7.27}$$

and $y \mapsto y$, $\mathbb{1}$ belong to $\mathcal{M}_\tau$. So, one has particularly also verified that the first half of Assumption **U2** and Assumption **U4** hold true in that case. Moreover, $(u-l)^{1+\delta} \in \mathcal{M}_\tau$ for any $\delta \in [0,1[$. This is the key of the verification of the second half of Assumption **U2**, thanks to Lemma 7.5.16.

Besides, Assumptions **C5a** and **C5b** hold true at least for the **MGM** example. It is for **C5a** an almost straightforward application of the parametric class example of (van der Vaart 1998); the result of a quite natural though painful study for **C5b**. In both cases, we can indeed prove that the class $\mathcal{G}$ has finite bracketing integral. We shall explain below why verifying whether Assumption **C5a** (and *a fortiori* Assumption **C5b**) holds true is not as straightforward for the **RKO** and **ME** examples.

Let us remind briefly for sake of completeness what it means (for a luminous and comprehensive introduction, see as usual van der Vaart 1998): given two functions l and u, the *bracket* $[l,u]$ is the set of all functions g with $l \leq g \leq u$; an *ε-bracket* in $L^2(P^\star)$ is a bracket $[l,u]$ such that $P^\star\,(u-l)^2 < \varepsilon^2$. The *bracketing number* $N_{[]}(\varepsilon, \mathcal{G}, L^2(P^\star))$ is the minimum number of ε-brackets needed to cover $\mathcal{G}$. The *entropy with bracketing* is the logarithm of the bracketing number and the *bracketing integral* $J_{[]}(\delta, \mathcal{G}, L^2(P^\star))$ is given by

$$J_{[]}(\delta, \mathcal{G}, L^2(P^\star)) = \int_0^\delta \Big(\log N_{[]}(\varepsilon, \mathcal{G}, L^2(P^\star))\Big)^{1/2} \mathrm{d}\varepsilon.$$

Those tools are intimately related to the Donsker property thanks to the following classical result and its corollary dealing with the so-called Lipschitz parametric class example:

Theorem 7.7.4. *Any class $\mathcal{G}$ with $J_{[]}(\delta, \mathcal{G}, L^2(P^\star)) < \infty$ is $P^\star$-Donsker. Moreover, in this particular setting, the class admits an envelope function that belongs to $L^2(P^\star)$.*

Corollary 7.7.5. *Any class $\mathcal{G} = \{g_\theta : \theta \in \Theta\}$ indexed by a bounded subset Θ of $\mathbb{R}^r$ is $P^\star$-Donsker with a $L^2(P^\star)$-integrable envelope function whenever there exists a function $G \in L^2(P^\star)$ such that*

$$|g_\theta(y) - g_t(y)| \leq G(y)\,\|\theta - t\| \quad (\text{any } \theta, t \in \Theta \text{ and } y \in \mathcal{Y}).$$

The latter allows to prove the next proposition:

Proposition 7.7.6. *Assumption* **C5a** *is satisfied for any $K > K^\star$ in the* **MGM** *example.*

Proof. Let us consider the class $\mathcal{G} = \{g_\theta = (\ell_\theta - \ell^\star) : \theta \in \Theta_K\}$ for some $K > K^\star$. Careful calculus yields that, for any $y \in \mathcal{Y}$, $\theta \in \mathrm{int}(\Theta_K)$ and $1 \leq k \leq K-1$

$$\begin{aligned} \left|\frac{\partial g_\theta}{\partial \pi_k}(y)\right| &= \left|\frac{\gamma_{m_k}(y) - \gamma_{m_K}(y)}{p_\theta(y)}\right|, \\ \left|\frac{\partial g_\theta}{\partial m_k}(y)\right| &= \left|\sigma^{-2}\,(y - m_k)\frac{\pi_k \gamma_{m_k}(y)}{p_\theta(y)}\right| \end{aligned} \tag{7.28}$$

and the right hand expression above is dominated by $\sigma^{-2}\Big(|y| + |\overline{m}| + |\underline{m}|\Big)$, which belongs to $\mathcal{M}_\tau$, hence to $L^2(P^\star)$.

The control of the first derivative requires more care. First, one always has an upper bound $(\pi_k^{-1} + \pi_K^{-1})$, with $\pi_K = \sum_{h<K} \pi_h$. Thus, the derivative particularly belongs to $\mathcal{M}_\tau$ because it is a bounded function of $y \in \mathcal{Y}$.

Now, choose $0 < \varepsilon < 1/2$. If π_k is bounded below by ε, then the left hand expression of (7.28) is bounded by $\varepsilon^{-1}\, \gamma_{m_k}^{-1}(y) |\gamma_{m_k}(y) - \gamma_{m_K}(y)|$.

Otherwise, one has either $\pi_K \leq \varepsilon$, or $\varepsilon \leq \pi_K$. In the first situation, there necessarily exists $h < K$ such that π_h is bounded below by $(1-\varepsilon)/(K-1)$, while in the second one, there exists $1 \leq h \leq K$ such that π_h is bounded below by $(1-\varepsilon)/K$. Thus, we get

$$\begin{aligned} \left|\frac{\partial g_\theta}{\partial \pi_k}(y)\right| &\leq \varepsilon^{-1}\, K \left|\gamma_{m_k}(y) - \gamma_{m_K}(y)\right| \sup_{1\leq h\leq K} \gamma_{m_h}^{-1}(y), \\ &\leq C\Big(\gamma_{\underline{m}}(y) + \gamma_{\overline{m}}(y)\Big)\Big(\gamma_{\underline{m}}^{-1}(y) + \gamma_{\overline{m}}^{-1}(y)\Big) \end{aligned} \tag{7.29}$$

for some positive constant C which depends on ε, K, $\underline{m}$ and $\overline{m}$. The upper bound defines a function that belongs to $L^2(P^\star)$. Thus, Taylor–Lagrange Inequality yields that Corollary 7.7.5 applies. □

Remark 7.7.7.

• We have proved that the derivatives at an inner point of Θ_K belong to $\mathcal{M}_\tau$. This point will be used later when proving that the **MGM** example also satisfies Assumption **U5**.

• The scheme of proof we have used for Proposition 7.7.6 does not work for the **RKO** and **ME** examples. Indeed, we would get final upper bounds in the spirit of (7.29), *i.e.* of the form

$$\Big(\sup_{u\in\mathcal{U}} \gamma_u\Big)\Big(\inf_{u\in\mathcal{U}} \gamma_u\Big)^{-1}$$

and such functions are not integrable wrt $P^\star$ for those examples.

Besides, as we announced earlier:

Proposition 7.7.8. *Assumption* **C5b** *holds true for the* **MGM** *example.*

The proof of the Proposition 7.7.8 is technical and fastidious. It is nevertheless worth to present it, even in a simpler yet relevant model, particularly because the scheme of proof gives an insight into the involved mechanisms that allow to forge an intuition. The mentioned proof is postponed in Appendix 7.8.

Therefore, we have shown that

- Theorems 7.4.1, 7.4.3 and 7.4.5 about consistency apply to the **MGM** example;
- Theorem 7.4.1 and the result of almost never underestimation of Proposition 7.4.8 apply to the **RKO** and **ME** examples.

Underestimation in the MGM example

As a straightforward verification yields,

Proposition 7.7.9. *Assumption* **U1** *holds true for the* **MGM** *example.*

Now, when investigating whether Assumptions **U2** and **U4** also do, Remark 7.7.3 ensures that both the first half of **U2** and **U4** are satisfied. Furthermore, Lemma 7.5.16 applied with the function $\psi(x) = x^{1+\delta}$ ($\delta \in [0, 1[$) yields the last half of Assumption **U2**, since we emphasized in Remark 7.7.3 that $\psi(u - l) \in \mathcal{M}_\tau$, hence:

Proposition 7.7.10. *Both Assumptions* **U2** *and* **U4** *are satisfied for the* **MGM** *example.*

We deal with Assumption **U3** thanks to Ascoli's theorem.

Proposition 7.7.11. *Assumption* **U3** *holds true for the* **MGM** *example.*

Proof. Choose indeed $K \leq K^\star$ and a compact set C of $\mathcal{Y}$. Let us denote $\mathcal{H}$ the class of functions

$$\mathcal{H} = \{ \ell_\theta \mathbb{1}\{C\} : \theta \in \Theta_K \}.$$

Primo, $\mathcal{H}$ is a bounded subset of $\mathrm{C}^0(C, \|\cdot\|_\infty)$ because the functions ℓ_θ are continuous. Actually, these functions $y \mapsto \ell_\theta(y)$ are continuously differentiable wrt y, with

$$\frac{\partial \ell_\theta}{\partial y}(y) = \sigma^{-2} \left(-y + \sum_{i=1}^{K} \frac{\pi_i m_i \gamma_{m_i}(y)}{p_\theta(y)} \right).$$

Now, the first term of the expression between parentheses is bounded since y belongs to the compact set C while the second term is bounded by $|\overline{m}| \vee |\underline{m}|$. Consequently, the Taylor–Lagrange Inequality ensures that $\mathcal{H}$ is equicontinuous, hence the final result. □

We finally cope with Assumption **U5** thanks to a crude yet careful application of Taylor's integral remainder theorem.

Proposition 7.7.12. *Assumption* **U5** *is also satisfied for the* **MGM** *example.*

Proof. Consider e_k the k-th canonical base vector of Θ_K corresponding to the weight π_k ($k < K$). Denoting $\pi_0 = (\pi_k^{-1} + \pi_K^{-1})$ and $h = \|h\| e_k$, one has whenever $\|h\| < {\pi_0}^{-1}$

$$|\ell_{\theta+h} - \ell_\theta - \dot{\ell_\theta}^T h| \leq {\pi_0}^2 \left(1 + (1 - \pi_0 \|h\|)^{-1} + (1 + \pi_0 \|h\|)^{-1} \right) \|h\|^2.$$

Moreover, if e_{K-1+k} is the $(K-1+k)$-th canonical base vector of Θ_K that corresponds to the mean m_k ($k < K$), one has for $\|h\|$ small enough

$$|\ell_{\theta+h}(y) - \ell_\theta(y) - \dot{\ell_\theta}^T(y) h| \leq \frac{\pi_k \gamma_{m_k}(y)}{p_\theta(y)} \left(1 + \sigma^{-2} + \sigma^{-4} (\|h\| + |y - m_k|)^2 \right) \|h\|^2.$$

Finally, the first factor in the right hand expression is bounded by 1 and the expression between parentheses defines a function $f_{m,\|h\|}(y)$ of $m \in \mathcal{M}$, $\|h\| \in \mathbb{R}_+$ and $y \in \mathcal{Y}$ such that the supremum function $\sup_{m \in \mathcal{M}, \|h\| \leq 1} f_{m,\|h\|}$ belongs to $\mathcal{L}_\tau$. This concludes the proof, since we have already noticed that the derivative functions $\dot{\ell_\theta}$ belong to $\mathcal{M}_\tau$ when showing Proposition 7.7.6. □

Therefore, Theorem 7.5.3 and 7.5.6 about underestimation apply to the **MGM** example, but not Proposition 7.5.9.

Overestimation in the MGM example

First, as a consequence of Propositions 7.7.6, 7.7.8 and 7.7.10, one can state

Proposition 7.7.13. *Assumption* **Oa** *holds true for the* **MGM** *example.*

On the contrary, the condition on an envelope function in Assumption **Ob** is certainly not satisfied, though we do not have a precise argument to prove this deficiency.

Consequently, Theorem 7.6.3 about overestimation applies to the **MGM** example.

7.7.2. Abrupt changes and various regressions

Short introduction

The **AC** example is meant to illustrate the mildness of the assumptions of the results about consistency. However, we shall only prove that the assumptions of Proposition 7.4.8 on the almost never underestimation hold true. Actually, the Donsker property asserted in Assumption **C5a** is too demanding regarding the sophistication of the class $\mathcal{T}_{\leq K^\star+1}$, although it is compact for the distance d_P. In Chapter 5, we proved the weak consistency of an estimator of the order (that coincided with the cardinality of the true partition) under a strong assumption on the class of allowed partitions which, informally, have to be decomposable into a finite number of elementary sets from a basic class of sets.

The regression framework is classical. One should refer to (Haughton 1988; Haughton 1989; Guyon and Yao 1999; Haughton and Keribin 2001) among our usual references for earlier results and methods related to order estimation.

Consistency

We denote $\|\cdot\|_2$ the $L^2(P)$-norm.

Lemma 7.7.14. *In the* **AC** *or* **VR** *examples, one has for any* P_θ, $P_t \in \Pi_\infty$,

$$2\sigma^2 H(P_\theta \,|P_t) = \|f_\theta - f_t\|_2^2.$$

Proposition 7.7.15. *The* **VR** *example satisfies Assumption* **C1**. *In the* **AC** *example, both Assumptions* **C1**–**C2** *hold true.*

Proof. The verification of Assumption **C1** is immediate.

Let us prove that **C2** is satisfied for the **AC** example. Suppose on the contrary that

$$H(P^\star \,|\Pi_K) \leq H(P^\star \,|\Pi_{K+1}), \tag{7.30}$$

i.e. we have an equality. Lower semicontinuity of $H(P^\star\,|\cdot)$ and compactness of Π_K ensure the existence of $P \in \Pi_K$ such that $H(P^\star\,|P) = H(P^\star\,|\Pi_K)$. Denote f the corresponding mean function of the form

$$f(x) = \sum_{k=1}^{K} m_k \mathbb{1}\{x \in \tau_k\} \quad (\text{any } x \in \mathcal{X})$$

for some vector $\mathbf{m} \in \mathcal{M}^K$ and partition $\tau \in \mathcal{T}_{\leq K}$. Then one can identify the mean vector $\mathbf{m}$ in virtue of Lemma 7.7.14 and thanks to the following simple decomposition ($f^\star$ stands for the

extended mean vector of $P^\star$ – "extended" because it may be not piecewise constant, see Remark 7.4.2). Indeed,

$$2\sigma^2 H(P^\star|P) = \|f^\star - f\|^2 = \sum_{k=1}^{K} P(f^\star - m_k)^2 \mathbb{1}\{\tau_k\},$$

hence (remember that the definition of a partition requires that $P(\tau_k) > 0$ for any k)

$$m_k = \frac{Pf^\star \mathbb{1}\{\tau_k\}}{P(\tau_k)} \quad (\text{any } 1 \le k \le K).$$

We shall prove below that $f^\star$ is constant on every τ_{k_0} with value m_{k_0} by considering the piecewise constant functions on the subdivisions obtained from τ when dividing τ_{k_0} into two subsets S and $\tau_{k_0} \setminus S$.

The latter equalities and Inequality (7.30) yield that, for any $1 \le k_0 \le K$, for any subset S of τ_{k_0} with positive P-measure

$$Pf^{\star 2} - \sum_{k=1}^{K} \frac{(Pf^\star \mathbb{1}\{\tau_k\})^2}{P(\tau_k)} \le Pf^{\star 2} - \sum_{1 \le k \ne k_0 \le K} \frac{(Pf^\star \mathbb{1}\{\tau_k\})^2}{P(\tau_k)} - \left(\frac{(Pf^\star \mathbb{1}\{S\})^2}{P(S)} + \frac{(Pf^\star \mathbb{1}\{\tau_{k_0} \setminus S\})^2}{P(\tau_{k_0} \setminus S)} \right)$$

or equivalently,

$$\frac{(Pf^\star \mathbb{1}\{S\})^2}{P(S)} + \frac{(Pf^\star \mathbb{1}\{\tau_{k_0} \setminus S\})^2}{P(\tau_{k_0} \setminus S)} \le \frac{(Pf^\star \mathbb{1}\{\tau_{k_0}\})^2}{P(\tau_{k_0})}.$$

Thus, we get, expanding first the right hand term and then factorizing

$$\frac{P(\tau_{k_0} \setminus S)}{P(S)} \Big(Pf^\star \mathbb{1}\{S\}\Big)^2 + \frac{P(S)}{P(\tau_{k_0} \setminus S)} \Big(Pf^\star \mathbb{1}\{\tau_{k_0} \setminus S\}\Big)^2 \le 2 \Big(Pf^\star \mathbb{1}\{S\}\Big)\Big(Pf^\star \mathbb{1}\{\tau_{k_0} \setminus S\}\Big). \tag{7.31}$$

Now, the basic inequality $2ab \le (au)^2 + (bu^{-1})^2$ (any $a, b \in \mathbb{R}$ and positive u) together with inequality (7.31) ensure (take $u^2 = P(\tau_{k_0} \setminus S)/P(S)$) that the equality holds in (7.31). Consequently, for any subset S of τ_{k_0} with positive P-measure:

$$\frac{Pf^\star \mathbb{1}\{S\}}{P(S)} = \frac{Pf^\star \mathbb{1}\{\tau_{k_0} \setminus S\}}{P(\tau_{k_0} \setminus S)} = \frac{Pf^\star \mathbb{1}\{\tau_{k_0}\} - Pf^\star \mathbb{1}\{S\}}{P(\tau_{k_0 \setminus S})},$$

hence, for any subset S of τ_{k_0}:

$$Pf^\star \mathbb{1}\{S\} = \frac{P(S)}{P(\tau_{k_0})} Pf^\star \mathbb{1}\{\tau_{k_0}\}.$$

Let us choose $S = S_+ = \{x \in \tau_{k_0} : f^\star(x) > Pf^\star \mathbb{1}\{\tau_{k_0}\} P(\tau_{k_0})^{-1}\}$: we obtain $P(S_+) = 0$. Let us choose now $S = S_- = \{x \in \tau_{k_0} : f^\star(x) < Pf^\star \mathbb{1}\{\tau_{k_0}\} P(\tau_{k_0})^{-1}\}$: we obtain $P(S_-) = 0$, hence finally $P(S_0) = P(\tau_{k_0})$ where $S_0 = \{x \in \tau_{k_0} : f^\star(x) = Pf^\star \mathbb{1}\{\tau_{k_0}\} P(\tau_{k_0})^{-1}\}$ (*i.e.* $f^\star$ P-almost surely constant on τ_{k_0}). We conclude because k_0 is arbitrary. □

Let us denote $\underline{f}$ and $\overline{f}$ two functions on $\mathcal{X}$ that respectively bound below and above the functions f_θ ($\theta \in \Theta_\infty$). In the **AC** example, one can simply take $\underline{f} = \inf \mathcal{U}$ and $\overline{f} = \sup \mathcal{U}$. The latter will be useful when proving

Proposition 7.7.16. *Assumptions* **C3** *and* **C4** *are satisfied in the* **AC** *and* **VR** *examples.*

Proof. The proof for **C4** is obvious. Verification of **C3** is readily proved. Let us define

$$2\sigma^2 l(x,y) = \sigma^2 \log(2\pi\sigma^2) - y^2 - \left(\underline{f}(y)^2 + \overline{f}(y)^2\right) + 2y\left(\underline{f}(y)\mathbb{1}\{y \geq 0\} + \overline{f}(y)\mathbb{1}\{y < 0\}\right),$$
$$2\sigma^2 u(x,y) = \sigma^2 \log(2\pi\sigma^2) - y^2 + 2y\left(\overline{f}(y)\mathbb{1}\{y \geq 0\} + \underline{f}(y)\mathbb{1}\{y < 0\}\right),$$

so that, for any $\theta \in \Theta_\infty$ and $(x,y) \in \mathcal{X} \times \mathcal{Y}$

$$l \leq \ell^\star,\ \ell_\theta \leq u \quad \text{and}$$
$$2\sigma^2 (u-l)(x,y) = \left(\underline{f}^2 + \overline{f}^2\right) + 2|y|\left(\overline{f} - \underline{f}\right).$$

Furthermore, $(u-l)$ belongs to $\mathcal{M}_\tau$, hence to $L^1(P^\star)$, which concludes the proof. □

Remark 7.7.17 (on the envelope function $(u-l)$).
We actually prove in that $(u-l) \in \mathcal{M}_\tau \subset \mathcal{L}_\tau$. So, one has particularly also verified that the first half of Assumption **U2** and Assumption **U4** hold true in that case. Moreover, $(u-l)^{1+\delta} \in \mathcal{M}_\tau$ for any $\delta \in [0,1[$. This is the key of the verification of the second half of Assumption **U2**, thanks to Lemma 7.5.16.

Proposition 7.7.18. *Assumption* **C5a** *holds true for any $K > K^\star$ in the* **VR** *example.*

Proof. We mimic the proof of Proposition 7.7.6. Set $K > K^\star$. Here, one has simply for $\theta \in \mathrm{int}(\Theta_K)$

$$\sigma^2 \left|\frac{\partial \ell_\theta}{\partial \theta_k}(x,y)\right| = \left|\Big(y - f_\theta(x)\Big)t_k(x)\right| \leq C_1|y| + C_2$$

for some positive constants C_1, C_2 that only depends on the family $\{t_h\}$ and $\mathcal{U}$. This completes the proof, since the right hand expression above belongs to $\mathcal{M}_\tau$, hence to $L^2(P^\star)$ and that the Taylor-Lagrange Inequality allows to apply Corollary 7.7.5. □

Remark 7.7.19. We have proved that the derivatives at an inner point of $\Theta_{K^\star}$ belong to $\mathcal{M}_\tau$. This point will be used later when proving that the **MGM** example also satisfies Assumption **U5**.

Finally, we have shown that

- Theorem 7.4.3 about consistency applies to the **VR** example;
- Theorem 7.4.1 and the result of almost never underestimation of Proposition 7.4.8 apply to the **AC** example.

Underestimation in the VR example

As a straightforward verification yields,

Proposition 7.7.20. *Assumption* **U1** *is fulfilled in the* **VR** *example.*

Now, when investigating whether Assumptions **U2** and **U4** also do, Remark 7.7.17 ensures that both the first half of **U2** and **U4** are satisfied. Furthermore, Lemma 7.5.16 applied with the function $\psi(x) = x^{1+\delta}$ ($\delta \in [0, 1[$) yields the last half of Assumption **U2**, since we emphasized in Remark 7.7.17 that $\psi(u - l) \in \mathcal{M}_\tau$, hence:

Proposition 7.7.21. *Both Assumptions* **U2** *and* **U4** *hold true in the* **VR** *example.*

We deal with Assumption **U3** thanks to Ascoli's theorem and a straightforward application of Taylor-Lagrange inequality. The verification of Assumption **U5** is also easy. It relies again on Taylor-Lagrange Inequality and use the fact that the functions t_K are uniformly bounded. Thus, we omit the proof of the following result:

Proposition 7.7.22. *Assumptions* **U3** *and* **U5** *are satisfied in the* **VR** *example.*

Finally, since the **VR** example is an example of exponential model, we have verified that Theorems 7.5.3, 7.5.6 as well as Proposition 7.5.9 apply to the **VR** example. In particular, $\widehat{K}_n^{\mathrm{G}}$ attains the optimal rate of underestimation with a view to Proposition 7.5.1. This is a new result.

Overestimation in the VR example

Propositions 7.7.6, 7.7.8 and 7.7.10 yield

Proposition 7.7.23. *Assumption* **Oa** *holds true for the* **VR** *example.*

Consequently, Theorem 7.6.3 about overestimation applies to the **VR** example.

7.8. Appendix

This Appendix is devoted to the proof of Proposition 7.7.8. We shall actually present the proof of a simpler result. Its scheme is nonetheless very relevant – and its complication justifies the simplification of the framework: we shall indeed consider below the case of a mixture of $K^\star + 1 = 2$ Gaussian distributions of variance 1 when the true distribution is Gaussian mean 0, variance 1. A general proof is beyond the scope of this Appendix, but follows the same lines.

A simpler framework

Hereafter, we assume that $K^\star = 1$ and that $P^\star$ is the standard Gaussian distribution, *i.e.*

$$\begin{aligned}\sigma^2 &= 1,\\ \ell^\star &= \log \gamma_0,\\ \ell_\theta &= \log\Big((1-\pi)\gamma_{m_1} + \pi\gamma_{m_2}\Big).\end{aligned}$$

The parameter θ denotes the triplet (π, m_1, m_2) that ranges over $\Theta = [0,1] \times [\underline{m}, \overline{m}]^2$ (with $\underline{m} \leq 0 \leq \overline{m}$) which is a bounded subset of $\mathbb{R}^3$.

We are interested in the normalized likelihood ratios

$$g_\theta = \frac{\ell_\theta - \ell^\star}{H(\theta)^{1/2}}$$

for any θ such that $H(\theta) > 0$: we indeed want to prove that the class $\mathcal{G}$ of all such g_θ's is a $P^\star$-Donsker class with envelope function G satisfying $\varphi(G) \in L^1(P^\star)$. Now, since $\mathcal{G}$ is a subset of $L^2(P^\star)$, it is sufficient to show that the assumption of Theorem 7.7.4 holds true, hence for instance that there exists a positive integer ρ such that

$$N_{[]}(\varepsilon, \mathcal{G}, L^2(P^\star)) = O(\varepsilon^{-\rho}).$$

We shall construct our ε-brackets in two steps. The subclass of all functions g_θ with parameter θ ε-away from the boundary (*i.e.* $|m_1|, |m_2|, \pi, 1-\pi \geq \varepsilon$) is straightforwardly handled thanks to Lemma 7.8.1 and Corollary 7.7.5. In order to cover the functions g_θ for θ ε-close to the boundary (*i.e.* not ε-away from it), we construct three kinds of grids of meshwidth ε and, for each θ ranging over the grid, an ε-bracket. They complete the covering of the class $\mathcal{G}$.

Away from the boundary

Lemma 7.8.1. *$H(\theta)$ is bounded below by a positive polynomial expression in ε^{-1} when θ is bounded away from the boundary of Θ by ε,* e.g.

$$H(\theta) \geq O(\varepsilon^{-7}) > 0$$

whenever $|m_1|, |m_2| \geq \varepsilon$ and $\varepsilon \leq \pi \leq 1 - \varepsilon$.

Now, since the class of likelihood ratios $(\ell_\theta - \ell^\star)$ is a Lipschitz parameter class, *i.e.* satisfies the assumption of Corollary 7.7.5 (apply the Taylor–Lagrange Inequality), the lemma above ensures that the class of g_θ's for θ bounded away from the boundary is a $P^\star$-Donsker class with envelope in $L^2(P^\star)$.

Proof. (of Lemma 7.8.1*)* The key of the proof is the next inequality

$$\log(1+u) \leq u - \frac{u^2}{2} \mathbb{1}\{u \leq 0\} \quad (\text{any } u \in \mathbb{R}).$$

It yields straightforwardly that, for any $\theta \in \Theta$ with $\varepsilon < \pi < 1 - \varepsilon$,

$$\begin{aligned} 2\,H(\theta) &\geq P^\star \left[(1-\pi) \left(e^{m_1 x - {m_1}^2/2} - 1 \right) + \pi \left(e^{m_2 x - {m_2}^2/2} - 1 \right) \right]^2 \mathbb{1}\{x \in A\} \\ &\geq \varepsilon^2 \left\{ P^\star \left(e^{m_1 x - {m_1}^2/2} - 1 \right)^2 \mathbb{1}\{x \in A\} + P^\star \left(e^{m_2 x - {m_2}^2/2} - 1 \right)^2 \mathbb{1}\{x \in A\} \right. \\ &\qquad \left. + 2P^\star \left(e^{m_1 x - {m_1}^2/2} - 1 \right) \left(e^{m_2 x - {m_2}^2/2} - 1 \right) \mathbb{1}\{x \in A\} \right\}, \end{aligned}$$

where A is the set of the real numbers such that the product of the two terms in the expression between brackets above is nonnegative.

• Assume (without loss of generality) that m_1 and m_2 are both positive and bounded below by ε (the case of negative means goes along the same lines). Then $\mathbb{R}_- \subset A$ and for any $x \leq 0$,

$$1 - e^{m_1 x - {m_1}^2/2} \geq 1 - e^{\varepsilon x} \geq 0$$
$$1 - e^{m_2 x - {m_2}^2/2} \geq 1 - e^{\varepsilon x} \geq 0$$

hence

$$H(\theta) \geq 2\varepsilon^2 h(\varepsilon) \quad \text{for} \quad h(\varepsilon) = P^\star(1 - e^{\varepsilon x})^2 \mathbb{1}\{x \leq 0\}.$$

Thanks to the Taylor–Young theorem, one has $h(\varepsilon) = \varepsilon^2/2 + o(\varepsilon^2)$ and consequently, $H(\theta)$ is bounded below by a polynomial in ε^4 for ε small enough.

• Assume (without loss of generality) that $m_1 \geq \varepsilon$ and $m_2 \leq -\varepsilon$. First, notice that $1 - e^{-u} \geq u/2$ (any $0 \leq u \leq 1$). Now, the interval $[m_2/4, m_1/4]$ is included in A. Choose $x \in [m_2/4, m_1/4]$ and consider below, first the case of $m_1 \leq 2$, then $m_1 \geq 2$:

$$1 - e^{m_1 x - {m_1}^2/2} \geq 1 - e^{m_1^2/4} \geq \frac{{m_1}^2}{8} \geq \frac{\varepsilon^2}{8},$$
$$1 - e^{m_1 x - {m_1}^2/2} \geq 1 - e^{-1} > 0.$$

The same inequalities also hold true when replacing m_1 by m_2, hence

$$H(\theta) \geq 128^{-1}\varepsilon^2 \varepsilon^4 P^\star([m_2/4, m_1/4]) \geq C\varepsilon^7$$

for ε small enough, where C is a positive constant independent of ε. □

Application of the previous lemma and Corollary 7.7.5 yields that the set of functions g_θ for θ ε-away from the boundary is $P^\star$-Donsker. So, as a basic example of permanence of the Donsker property, it suffices to prove now that the remainder of the class is also $P^\star$-Donsker.

Grids near the boundary – scheme of proof

The table below describes 22 regions whose union covers the set of all $\theta \in \Theta$ which are ε-close to the boundary:

$\pi, \lvert m_1\rvert \leq \varepsilon, \lvert m_2\rvert \geq b$		$1-\pi, \lvert m_2\rvert \leq \varepsilon, \lvert m_1\rvert \geq b$	
$\lvert m_1\rvert/\pi \leq 1$	$\lvert m_1\rvert/\pi > 1$	$\lvert m_2\rvert/\pi \leq 1$	$\lvert m_2\rvert/\pi > 1$
R_1	R_2	R_3	R_4

$\lvert m_1\rvert, \lvert m_2\rvert \leq \varepsilon, b \leq \pi \leq 1-b$	
$\lvert m_1\rvert/\lvert m_2\rvert \leq 1$	$\lvert m_1\rvert/\lvert m_2\rvert > 1$
R_5	R_6

$\pi, \lvert m_1\rvert, \lvert m_2\rvert \leq \varepsilon$	$1-\pi, \lvert m_1\rvert, \lvert m_2\rvert \leq \varepsilon$
$R_7, \ldots, R_{14}$	$R_{15}, \ldots, R_{22}$

We do not give the definition of $R_7, \ldots, R_{22}$ for sake of legibility: on each of them, the three ratios $|m_1|/|m_2|$, $|m_1|/\pi$ and $|m_2|/\pi$ are bounded below or above by 1.

Symmetry arguments justify that it is not necessary to study the 22 regions one by one. Even then, it would be quite tedious to present an accurate study for each symmetric case. Thus, we only present hereafter two comprehensive studies and prepare another, but do not complete it. The other cases would go along the same lines.

For the regions $R = R_1, R_5$ (with an additional assumption on the signs of m_1 and m_2 for sake of clarity – other cases easily follow), we construct below a grid of meshwidth ε over R whose cardinality is bounded above by a $O(\varepsilon^{-2})$; for each θ in the grid, we construct an ε-bracket; the union of all those ε-brackets covers the set $\{g_\theta : \theta \in R\}$.

First region: $\pi, |m_1| \leq \varepsilon$, $|m_2| \geq b$ **and** $0 \leq |m_1|/\pi \leq 1$

Actually, we shall only consider the case of nonnegative m_1. The case of negative m_1 is similar.

Let us introduce some temporary notations for sake of legibility. For any real numbers x, u, v, mean m_2 such that $|m_2| \geq b$, integer k and $\varepsilon > 0$, define

$$\psi_{m_2}(x) = e^{m_2 x - {m_2}^2/2}, \quad 2\, q_{m_2}(u,v) = u^2 + 2 m_2\, uv + (e^{{m_2}^2} - 1)v^2.$$

Now, the Taylor–Young theorem and the Taylor–Lagrange Inequality yield the lemma below, whose proof is not reproduced here:

Lemma 7.8.2. *Suppose that $\pi \to 0$, $m_1 \to 0$ and $|m_2| \geq b$. The following Taylor–Young expansion holds true*

$$H(\theta) = q_{m_2}(m_1, \pi) + o(\pi^2 + {m_1}^2).$$

Moreover, there exist two positive constant B (depending on b) and ε_0 such that, for any $\varepsilon < \varepsilon_0$, if $0 \leq m_1/\pi \leq 1$, $m_1 \leq \varepsilon$ and $\pi \leq \varepsilon$,

$$\left| \frac{H(\theta)}{q_{m_2}(m_1, \pi)} - 1 \right| \leq B\,\varepsilon < 1.$$

Furthermore, there exists a positive constant C such that, under the same conditions on m_1, m_2, π and ε than above, for any $x \in \mathbb{R}$,

$$\left| (\ell_\theta - \ell^\star)(x) - \Big(m_1\, x + \pi\, \psi_{m_2}(x) \Big) \right| \leq ({m_1}^2 + \pi^2)\, e^{C|x|}.$$

Let $\varepsilon_0 > 0$ of Lemma 7.8.2 be small enough so that, for any $0 \leq u \leq \varepsilon_0$,

$$(1-u)^{-1/2} \leq 1 + u \quad \text{and} \quad (1+u)^{-1/2} \geq 1 - u. \tag{7.32}$$

Choose an $\varepsilon \leq \varepsilon_0$. We shall construct some ε-brackets for the class

$$\mathcal{G}_1 = \Big\{ g_\theta : 0 \leq m_1, \pi \leq \varepsilon,\ m_1/\pi \leq 1,\ |m_2| \geq b \Big\},$$

and evaluate the bracketing number. We shall finally conclude that the bracketing integral is finite.

Choose any m_1, π, m_2 under the condition of Lemma 7.8.2 and let k denote the sole integer such that $k\varepsilon \leq m_1/\pi < (k+1)\varepsilon$. Denote also

$$\overline{f}_{m_2}(x) = \overline{f}_{m_2,\varepsilon,k}(x) = \psi_{m_2}(x) + \varepsilon x \Big((k+1)\mathbb{1}\{x \geq 0\} + k\mathbb{1}\{x < 0\} \Big) + \varepsilon\, e^{C|x|},$$

$$\underline{f}_{m_2}(x) = \underline{f}_{m_2,\varepsilon,k}(x) = \psi_{m_2}(x) + \varepsilon x \Big(k\mathbb{1}\{x \geq 0\} + (k+1)\mathbb{1}\{x < 0\} \Big) - \varepsilon\, e^{C|x|}.$$

As a consequence of Lemma 7.8.2, one has for any $x \in \mathbb{R}$

$$\frac{(\ell_\theta - \ell^\star)(x)}{H(\theta)^{1/2}} \leq \overline{f}_{m_2,\varepsilon,k}(x) \Big\{ (1 - B\,\varepsilon)^{-1/2}\, q_{m_2}(k\varepsilon, 1)^{-1/2} \mathbb{1}\left\{\overline{f}_{m_2}(x) \geq 0\right\} + (1 + B\,\varepsilon)^{-1/2}\, q_{m_2}((k+1)\varepsilon, 1)^{-1/2} \mathbb{1}\left\{\overline{f}_{m_2}(x) < 0\right\} \Big\},$$

$$\frac{(\ell_\theta - \ell^\star)(x)}{H(\theta)^{1/2}} \geq \underline{f}_{m_2,\varepsilon,k}(x) \Big\{ (1 + B\,\varepsilon)^{-1/2}\, q_{m_2}((k+1)\varepsilon, 1)^{-1/2} \mathbb{1}\{\underline{f}_{m_2}(x) \geq 0\} + (1 - B\,\varepsilon)^{-1/2}\, q_{m_2}(k\varepsilon, 1)^{-1/2} \mathbb{1}\{\underline{f}_{m_2}(x) < 0\} \Big\}.$$

We then apply (7.32) and get

$$\frac{(\ell_\theta - \ell^\star)(x)}{H(\theta)^{1/2}} \leq \overline{f}_{m_2,\varepsilon,k}(x) \Big\{ (1 + B\,\varepsilon)\, q_{m_2}(k\varepsilon, 1)^{-1/2} \mathbb{1}\left\{\overline{f}_{m_2}(x) \geq 0\right\} + (1 - B\,\varepsilon)\, q_{m_2}((k+1)\varepsilon, 1)^{-1/2} \mathbb{1}\left\{\overline{f}_{m_2}(x) < 0\right\} \Big\} = f^+_{m_2,\varepsilon,k}(x) = f^+_{m_2}(x)$$

$$\frac{(\ell_\theta - \ell^\star)(x)}{H(\theta)^{1/2}} \geq \underline{f}_{m_2,\varepsilon,k}(x) \Big\{ (1 - B\,\varepsilon)\, q_{m_2}((k+1)\varepsilon, 1)^{-1/2} \mathbb{1}\{\underline{f}_{m_2}(x) \geq 0\} + (1 + B\,\varepsilon)\, q_{m_2}(k\varepsilon, 1)^{-1/2} \mathbb{1}\{\underline{f}_{m_2}(x) < 0\} \Big\} = f^-_{m_2,\varepsilon,k}(x) = f^-_{m_2}(x)$$

and this defines a bracket $[f^-_{m_2,\varepsilon,k}, f^+_{m_2,\varepsilon,k}]$: thus, we have to study the difference of the superior and the inferior functions in order to verify whether this defines an ε-bracket. We propose to decompose the difference as the following sum of two expressions:

$$\begin{aligned}
\left(f^+_{m_2,\varepsilon,k} - f^-_{m_2,\varepsilon,k}\right)(x) = \\
B\varepsilon \Big[& q_{m_2}(k\varepsilon, 1)^{-1/2} \left(|\overline{f}_{m_2}(x)| \mathbb{1}\{\overline{f}_{m_2}(x) \geq 0\} + |\underline{f}_{m_2}(x)| \mathbb{1}\{\underline{f}_{m_2}(x) < 0\} \right) + \\
& q_{m_2}((k+1)\varepsilon, 1)^{-1/2} \left(|\overline{f}_{m_2}(x)| \mathbb{1}\{\overline{f}_{m_2}(x) < 0\} + |\underline{f}_{m_2}(x)| \mathbb{1}\{\underline{f}_{m_2}(x) \geq 0\} \right) \Big] \\
+ \Big[& q_{m_2}(k\varepsilon, 1)^{-1/2} \left(|\overline{f}_{m_2}(x)| \mathbb{1}\{\overline{f}_{m_2}(x) \geq 0\} + |\underline{f}_{m_2}(x)| \mathbb{1}\{\underline{f}_{m_2}(x) < 0\} \right) - \\
& q_{m_2}((k+1)\varepsilon, 1)^{-1/2} \left(|\overline{f}_{m_2}(x)| \mathbb{1}\{\overline{f}_{m_2}(x) < 0\} + |\underline{f}_{m_2}(x)| \mathbb{1}\{\underline{f}_{m_2}(x) \geq 0\} \right) \Big].
\end{aligned}$$

Now, the first expression between brackets defines a function of m_2, ε, k and x which is bounded above uniformly (in $|m_2| \geq b$, $\varepsilon < \varepsilon_0$ and k such that $k\varepsilon \leq 1 + \varepsilon_0$ – we shall simply write *uniformly* in the sequel) by a function (of x) that belongs to $L^2(P^\star)$. Hence, the factor $B\varepsilon$ in front of the latter expression ensures the control of this first term.

For the second term, we apply again (7.32) twice, respectively for

$$q_{m_2}((k+1)\varepsilon, 1)^{-1/2} = q_{m_2}(k\varepsilon, 1)^{-1/2} \left(1 + \frac{\varepsilon\,[(k+1/2)\varepsilon + m_2]}{q_{m_2}(k\varepsilon, 1)} \right)^{-1/2},$$

$$q_{m_2}(k\varepsilon, 1)^{-1/2} = q_{m_2}((k+1)\varepsilon, 1)^{-1/2} \left(1 - \frac{\varepsilon\,[(k+1/2)\varepsilon + m_2]}{q_{m_2}((k+1)\varepsilon, 1)} \right)^{-1/2}$$

This yields the two following upper bounds

$$q_{m_2}(k\varepsilon,1)^{-1/2}\overline{f}_{m_2}(x) - q_{m_2}((k+1)\varepsilon,1)^{-1/2}\underline{f}_{m_2}(x) + \varepsilon\,[(k+1/2)\varepsilon + m_2]\left(q_{m_2}(k\varepsilon,1)^{-3/2}|\overline{f}_{m_2}| + q_{m_2}((k+1)\varepsilon,1)^{-3/2}\right),$$

$$q_{m_2}((k+1)\varepsilon,1)^{-1/2}\overline{f}_{m_2}(x) - q_{m_2}(k\varepsilon,1)^{-1/2}\underline{f}_{m_2}(x) + \varepsilon\,[(k+1/2)\varepsilon + m_2]\left(q_{m_2}((k+1)\varepsilon,1)^{-3/2}|\overline{f}_{m_2}| + q_{m_2}(k\varepsilon,1)^{-3/2}\right),$$

whose sum is still an upper bound, which enjoys some nice properties of symmetry:

$$\left[q_{m_2}(k\varepsilon,1)^{-1/2} + q_{m_2}((k+1)\varepsilon,1)^{-1/2}\right]\left(\overline{f}_{m_2}(x) - \underline{f}_{m_2}(x)\right) + \varepsilon\left[\left((k+1/2)\varepsilon + m_2\right)\left(q_{m_2}(k\varepsilon,1)^{-3/2} + q_{m_2}((k+1)\varepsilon,1)^{-3/2}\right)\right]\left(|\overline{f}_{m_2}(x)| + |\underline{f}_{m_2}(x)|\right).$$

The hard work is done. One can conclude at this stage that the integral wrt $P^\star$ of the latter upper bound is bounded above by ε, up to a constant. Indeed, on the one hand,

$$(\overline{f}_{m_2} - \underline{f}_{m_2})(x) = \varepsilon\left(2e^{C|x|} + |x|\right)$$

(which defines ε times an element of $L^2(P^\star)$) and the expression between brackets that precedes it is *uniformly* bounded by a constant; on the other hand, the second expression is also *uniformly* bounded by ε times a function (of x) that belongs to $L^2(P^\star)$.

Hence, we have proved that the bracket $[f^-_{m_2,\varepsilon,k}, f^+_{m_2,\varepsilon,k}]$ is an ε-bracket. The bracketing number that corresponds to those brackets is a $O(\varepsilon^{-1})$. This is our temporary conclusion.

It suffices now to notice (apply again Taylor–Lagrange Inequality) that both classes of functions

$$\left\{f^+_{m_2} : |m_2| \geq b\right\} \quad \text{and} \quad \left\{f^-_{m_2} : |m_2| \geq b\right\}$$

are Lipschitz parametric classes (see Corollary 7.7.5), *i.e.* that there exist two functions G^+, G^- in $L^2(P^\star)$ such that, for any m_2, m_2' and x,

$$|f^+_{m_2}(x) - f^+_{m_2'}(x)| \leq G^+(x)\,|m_2 - m_2'| \quad \text{and} \quad |f^-_{m_2}(x) - f^-_{m_2'}(x)| \leq G^-(x)\,|m_2 - m_2'|.$$

Let us consider the $O(\varepsilon^{-2})$ brackets

$$[f^-_{(b+h\,\varepsilon),\varepsilon,k} - \varepsilon\,G^-, f^+_{(b+h\,\varepsilon),\varepsilon,k} + \varepsilon\,G^+] \quad \text{and} \quad [f^-_{(-b-h'\,\varepsilon),\varepsilon,k} - \varepsilon\,G^-, f^+_{(-b-h'\,\varepsilon),\varepsilon,k} + \varepsilon\,G^+].$$

(where $k\varepsilon \leq 1+\varepsilon$, $b + h\,\varepsilon \leq \overline{m}$ and $\underline{m} \leq -b - h'\,\varepsilon$) Those brackets are ε-brackets in virtue of the temporary conclusion above and their union covers the whole class $\mathcal{G}_1$.

Consequently, the bracketing integral is finite: this completes the proof.

Second region: $|m_1|, |m_2| \leq \varepsilon$, $b \leq \pi \leq 1 - b$ and $0 \leq |m_1|/|m_2| \leq 1$

Actually, we shall only consider the case of nonnegative m_1 and m_2. The complementary cases are very similar.

Let us introduce again a temporary notation for sake of legibility. For any weight π such that $b \leq \pi \leq 1 - \pi$ and any real numbers u, v, define

$$2\, q_\pi(u,v) = \Big((1-\pi)\, m_1 + \pi\, m_2\Big)^2.$$

In the same spirit than Lemma 7.8.2, one can state (the proof is omitted)

Lemma 7.8.3. *Suppose that $m_1 \to 0$, $m_2 \to 0$ and $b \leq \pi \leq 1 - b$. The following Taylor–Young expansion holds true*

$$H(\theta) = q_\pi(m_1, m_2) + o({m_1}^2 + {m_2}^2).$$

Moreover, there exist two positive constant B and ε_0 such that, for any $\varepsilon < \varepsilon_0$, if $0 \leq m_1/m_2 \leq 1$, $0 \leq m_1, m_2 \leq \varepsilon$ and $b \leq \pi \leq 1 - b$,

$$\left| \frac{H(\theta)}{q_\pi(m_1, m_2)} - 1 \right| \leq B\, \varepsilon < 1.$$

Furthermore, there exists a positive constant C such that, under the same conditions on m_1, m_2, π and ε than above, for any $x \in \mathbb{R}$,

$$\Big| (\ell_\theta - \ell^\star)(x) - \Big((1-\pi)\, m_1 + \pi\, m_2\Big)\, x \Big| \leq ({m_1}^2 + {m_2}^2)\, e^{C|x|}.$$

Choose an $\varepsilon \leq \varepsilon_0$. We shall construct some ε-brackets for the class

$$\mathcal{G}_2 = \Big\{ g_\theta : 0 \leq m_1, m_2 \leq \varepsilon,\ m_1/m_2 \leq 1,\ b \leq \pi \leq 1 - b \Big\},$$

and evaluate the bracketing number. We shall finally conclude that the bracketing integral is finite.

Let us set some m_1, m_2, π under the condition of Lemma 7.8.3 and let k denote the sole integer such that $k\varepsilon \leq m_1/m_2 < (k+1)\varepsilon$. Denote also

$$\overline{f}_\pi(x) = \overline{f}_{\pi,\varepsilon,k}(x) = \pi\, x + \varepsilon\, (1-\pi)\, x \Big((k+1) \mathbb{1}\{x \geq 0\} + k \mathbb{1}\{x < 0\} \Big) + \varepsilon\, e^{C|x|},$$
$$\underline{f}_\pi(x) = \underline{f}_{\pi,\varepsilon,k}(x) = \pi\, x + \varepsilon\, (1-\pi)\, x \Big(k \mathbb{1}\{x \geq 0\} + (k+1) \mathbb{1}\{x < 0\} \Big) - \varepsilon\, e^{C|x|}.$$

Then, the proof would be written exactly as in the previous case, replacing everywhere $\overline{f}_{m_2}$ by $\overline{f}_\pi$, $\underline{f}_{m_2}$ by $\underline{f}_\pi$ and q_{m_2} by q_π. The "uniform" upper bounds would be uniform in $b \leq \pi \leq 1 - b$, $\varepsilon < \varepsilon_0$ and k such that $k\varepsilon \leq 1 + \varepsilon_0$, of course. The final argument would be of the same kind than the former one. We conclude again that $\mathcal{G}_2$ is $P^\star$-Donsker, since it has finite bracketing integral.

Third region: $0 \leq \pi, |m_1|, |m_2| \leq \varepsilon$

We only state here the basic lemma that allows a proof in the same spirit than the previous ones. Observe that this time, we have to expand further our expressions.

Let us define $q(u,v,w) = \frac{1}{2}u^2 + uvw$.

Lemma 7.8.4. *Suppose that $m_1, m_2, \pi \to 0$ simultaneously. The following Taylor–Young expansion holds true*

$$H(\theta) = q(m_1, m_2, \pi) + o(|m_1|^3 + |m_2|^3 + \pi^3).$$

Moreover, there exist two positive constant B and ε_0 such that, for any $\varepsilon < \varepsilon_0$, whenever $0 \leq m_1, m_2, \pi \leq \varepsilon$,

$$\left| \frac{H(\theta)}{q(m_1, m_2, \pi)} - 1 \right| \leq B\varepsilon < 1.$$

Furthermore, there exists a positive constant C such that, under the same conditions on m_1, m_2, π and ε than above, for any $x \in \mathbb{R}$,

$$\left| (\ell_\theta - \ell^\star)(x) - x(m_1 - m_1\pi + m_2\pi - {m_1}^2) - \frac{1}{2}{m_1}^2 \right| \leq ({m_1}^3 + {m_2}^3 + \pi^3)\, e^{C|x|}.$$

A

Glossaire

- **Affaiblissement** : perte en puissance d'un champ émis. Se calcule en dB ou en dBm. Voir champ, dB et dBm.
- **APE** : *Activité Principale Exercée* par une entreprise ou par un établissement. Ce code est attribué par l'INSEE à chaque entreprise et à chaque établissement selon son activité principale. Il est constitué d'après la NAF. Voir APET60 et APET700.
- **APET60** : *Activité Principale exercée par les ETablissements* de finesse 60, classification des APE suivant 60 divisions, moins fine que la classification APET700.
- **APET700** : *Activité Principale exercée par les ETablissements* de finesse 700, classification des APE suivant 696 postes. S'obtient par raffinement de chacune des 60 divisions du codage APET60.
- **BSC** : *Base Station Controller*, l'un des deux éléments constitutifs du BSS avec la BTS. Le BSC contrôle plusieurs BTS. Il a pour fonction principale de gérer la ressource radio. Le BSC est relié au NSS. Voir la Figure 2.
- **BSS** : *Base Station Sub-System*, sous-système radio qui constitue le premier des trois sous-ensembles (BSS/NSS/OSS) du réseau GSM. Voir la Figure 2.
- **BTS** : *Base Transceiver Station*, l'un des deux éléments constitutifs du BSS avec le BSC. La BTS, ou *station de base*, est un émetteur-récepteur. Elle est composée de plusieurs TRX. La BTS a la charge de la transmission radio. Elle est reliée à un BSC. Voir la Figure 2.
- **Cellule** : zone de desserte d'une BTS. Pour une exemple de cellule standard, voir la Figure 1.6.
- **Champ** : champ induit par un signal. Si le signal est émis à une puissance P, le champ au voisinage de l'émetteur vaut $10 \log_{10} P$. Un champ est par ailleurs soumis à un affaiblissement à mesure que l'on s'éloigne de l'émetteur. Se calcule en dB ou en dBm Voir affaiblissement, dB et dBm. Pour une illustration, voir la Figure 1.5.
- **Cigale** : acronyme pour *Contrôle de l'Interface Généralisée A partir des Lectures d'Enregistrements*. Jeu de données exhaustif sur le trafic mobile. Il est constitué des relevés d'échanges pour toutes les connexions sur une journée. Son extraction au niveau des MSC est automatisée au sein de l'OSS. Ce jeu de données se substitue au jeu CRA et complète le jeu HC2. Voir connexion, CRA, échange et HC2. Pour un échantillon, voir le Tableau 1.2.
- **Connexion** : une connexion est une série d'échanges (qui en constituent les parties élémentaires) entre un mobile, le BSS et le NSS. Ces échanges sont à l'origine des relevés Cigale. Voir BSS, Cigale, échange, mobile et NSS.
- **CONTOURSILOTS** : base de données de contours des îlots créés pour procéder au recensement.
- **CRA** : *Comptes-Rendus d'Appels*, jeu de donnée utilisé à l'origine à des fins de facturation des appels. Ce jeu de données est insatisfaisant. Il s'est vu substituer les jeux Cigale et HC2. Voir Cigale et HC2. Pour un échantillon, voir le Tableau 1.1.
- **dB** : unité de champ et d'affaiblissement. Le champ $10 \log_{10} P$ induit par une puissance P est exprimé en dB si la puissance est en Watt, en dBm si elle est en mWatt. Voir affaiblissement et champ.
- **dBm** : voir dB.
- **DMR** : *Direction des services Mobiles et Systèmes Radio*, l'une des neuf unités de recherche de FTR&D. Le laboratoire IIM en dépend.
- **Echange** : partie élémentaire d'une connexion. Chaque échange émet un rapport lorsqu'il s'achève. Ces rapports sont à l'origine des relevés Cigale. Voir Cigale et connexion.
- **Entreprise** : unité économique, juridiquement autonome, organisée pour pro-

duire des biens ou des services pour le marché. On distingue : (i) l'entreprise individuelle (personne physique) qui ne possède pas de personnalité juridique distincte de celle de la personne physique de son exploitant ; (ii) l'entreprise sociétaire, SA ou SARL.

- **Erlang** : unité de mesure de quantité de trafic téléphonique. Un trafic d'un Erlang correspond à un appel sur un canal radio pendant 3600 secondes. Voir la Section 1.2.2 du Chapitre 1.
- **Etablissement** : unité de production localisée géographiquement, individualisée mais dépendant juridiquement d'une entreprise. L'établissement constitue le niveau le mieux adapté à une approche géographique de l'économie. Il est relativement homogène et son activité principale apparaît proche du produit.
- **Exploitation principale** : première des deux exploitations statistiques auxquelles sont soumis les bulletins individuels et les feuilles de logement issus d'un recensement. Elle est qualifiée de légère au regard de l'exploitation complémentaire, qui necessite une lecture plus approfondie des bulletins et feuilles.
- **France Télécom R&D** : centre de recherche du groupe français de télécommunications France Télécom.
- **GPRS** : *General Packet Radio Service*, technologie de sophistication du GSM. Permet la transmission de données par paquets sur la voie radio ; ouvre ainsi la voie aux applications mobiles multimédia, *i.e.* à la téléphonie mobile de troisième génération. Voir UMTS.
- **GSM** : *Global System for Mobile communications*, norme de téléphonie de seconde génération, *i.e.* numérique et celllaire. Le réseau téléphonique mobile Orange en est un exemple d'application avec sophistication due à la technologie GPRS.
- **Handover** : transfert intercellulaire.
- **HC2** : acronyme pour 2ème Heure Chargée. Ce jeu de données de trafic est constitué, pour chaque cellule $\mathcal{C}$ de la quantité HC2 sur chaque semaine $\mathcal{S}$ de la période d'observation. La quantité HC2 de $\mathcal{C}$ pour $\mathcal{S}$ est déterminée comme suit : *primo*, calcul des sept plus importantes quantités quotidiennes de trafic écoulé à l'heure ; *secundo*, choix de la seconde plus grande de ces quantités. Ce jeu de données complète le jeu Cigale et se substitue au jeu CRA. Voir Cigale et CRA.
- **HLR** : *Home Location Register*, base de données sur les abonnements, offrant aussi aux MSC qui lui sont reliés des informations de plus récente localisation grossière des mobiles. L'un des trois éléments constitutifs du NSS avec le MSC et le VLR. Voir la Figure 2.
- **Identifiant** : attribué par l'INSEE à toute personne juridique, physique ou morale, introduite dans le répertoire SIRENE sur demande des organismes habilités. Voir SIREN et SIRET.
- **IIM** : *Interface radio et Ingiénierie pour réseaux Mobiles*, laboratoire de la direction DMR dans le cadre duquel cette thèse a été faite.
- **Ilot** : zone de surface réduite, aussi homogène que possible en termes socio-démographique (le zonage est fixé en partenariat avec les collectivités locales), peuplée de moins de 800 habitants. En milieu urbain (cas qui nous intéresse), il s'agit de la plus petite surface délimitée par des voies publiques et/ou privées : l'îlot correspond ainsi à ce que l'on entend ordinairement par "pâté de maisons".
- **Itinérance** : ou *roaming*, possibilité de téléphoner de n'importe où depuis son mobile. A ne pas confondre avec la mobilité.
- **ILOTS15** : base de données issues de l'exploitation principale du recensement 1999. Offre 24 indicateurs concernant la répartition de la population par tranches

d'âges, celle des logements et la population des résidences principales à l'échelle des îlots.

- **IMEI** : *International Mobile Equipment Identity*, numéro d'identification d'un téléphone mobile, attribué à la fabrication de façon unique et définitive.
- **INSEE** : *Institut National de la Statistique et des Etudes Economiques*, un des organismes de la statistique publique française.
- **Logement** : local séparé et indépendant utilisé pour l'habitation. On distingue les logements *principaux*, *secondaires*, *ocassionnels* et *vacants*.
- **Macrocellule** : cellule à grande superficie en comparaison de celle d'une microcellule. Correspond à des BTS placées au-dessus des toits en zone urbaine. Voir la Figure 3.
- **Maille** : zone élémentaire carrée sur lesquelles sont évalués les données d'affaiblissement de champ. Les cellules de desserte sont décrites en termes de mailles. Voir affaiblissement et cellule. Pour une illustration, voir les Figures 1.5 et 1.6.
- **Microcellule** : cellule à petite superficie en comparaison de celle d'une macrocellule. Correspond à des BTS placées audessous du niveau des toits en zone urbaine. Voir la Figure 3.
- **Mobile** : terminal portatif de télécommunication itinérante et mobile.
- **Mobilité** : possibilité de se déplacer au cours d'un appel depuis son mobile. A ne pas confondre avec l'itinérance.
- **MSC** : *Mobiles-services Switching Centers*, commutateur mobile en liaison avec le BSS et le RTC, élément constitutif du NSS. Voir la Figure 2.
- **NAF** : *Nomenclature d'Activités Française.*
- **NOTICES** : nom des fichiers d'extraction du répertoire SIRENE.
- **NSS** : *Network Sub-System*, sous-système d'acheminement (ou réseau fixe) qui constitue le second des trois sous-ensembles (BSS/NSS/OSS) du réseau GSM. Voir la Figure 2.
- **Occasionnel (Logement)** : logement (ou pièce indépendante) utilisé une partie de l'année pour des raisons professionnelles.
- **Orange** : Branche Mobile française du groupe France Télécom. Orange SA est la holding qui regroupe tous les opérateurs mobiles du groupe France Télécom.
- **OSS** : *Operation Sub-System*, sous-système d'exploitation et de maintenance qui constitue le dernier des trois sous-ensembles (BSS/NSS/OSS) du réseau GSM.
- **Principal (logement)** : logement occupé de façon permanente et à titre principal par un ménage. Voir Logement.
- **RTC** : *Réseau Téléphonique Commuté*, réseau de téléphonie fixe. Il est en relations avec le réseau GSM/GPRS *via* les MSC.
- **SA** : *Société Anonyme*, voir Entreprise.
- **SARL** : *Société à Responsabilité Limitée*, voir Entreprise.
- **Secondaire (logement)** : logement occupé de façon temporaire. Voir Logement.
- **SIREN** : numéro d'immatriculation d'une entreprise vue comme personne juridique. Il est unique et n'est attribué qu'une seule fois.
- **SIRENE** : répertoire *Système Informatique pour le Répertoire des ENtreprises et de leurs Etablissements* dont la gestion incombe à l'INSEE. Il enregistre l'état civil de toutes les entreprises et leurs établissements en France, quelle que soit leur forme juridique et quel que soit leur secteur d'activité.

  L'INSEE attribue à chaque entreprise un identifiant numérique SIREN et à chaque établissement un identifiant numérique SIRET. La base de données SIRENE reprend, pour les seuls entreprises

et établissements administrativement actifs, les informations contenues dans le répertoire SIRENE.

- **SIRET** : numéro d'immatriculation d'un établissement d'une entreprise en tant qu'unité géographiquement localisée. Il est obtenu par complétion du numéro SIREN correspondant.
- **TRX** : émetteur-récepteur qui est l'élément constitutif d'une station de base, ou BTS.
- **UMTS** : *Universal Mobile Telecommunications System*, technologie de troisième génération de téléphonie mobile caractérisée par l'offre et la diffusion de contenus multimédia sur les mobiles.
- **Vacant (logement)** : logements sans occupant, autres que les résidences secondaires, disponibles à la vente ou à la location. Voir Logement.
- **VLR** : *Visitor Location Register*, très semblable au HLR.

B

Caccioppoli partitions*

*I would like to thank Raphaël Cerf for the useful reference (Leonardi and Tamanini 2002).

This appendix is essentially based on Leonardi and Tamanini's paper (2002) on metric spaces of partitions. Let $(\mathcal{X}, \mathcal{B}, P)$ be a measure space, with $\mathcal{X} \subset \mathbb{R}^p$, $\mathcal{B}$ the trace of the Borel σ-field on $\mathcal{X}$ and $P << \mu$ (μ is the Lebesgue measure) a probability measure whose support is the whole $\mathcal{X}$ (*i.e.* $\{x \in \mathcal{X} : x \in \mathcal{O} \Rightarrow P(\mathcal{O}) > 0\} = \mathcal{X} - \mathcal{O}$ denotes an open set).

Perimeter and Caccioppoli sets

Let $\mathcal{O}$ be an open subset of $\mathcal{X}$ and $B \in \mathcal{B}$ any Borel set. The *perimeter* of B in $\mathcal{O}$ is defined by

$$\mathrm{Per}(B, \mathcal{O}) = \sup \left\{ \mu \left(\operatorname{div} \varphi \, \mathbb{1}\{B \cap \mathcal{O}\} \right) : \varphi \in \mathrm{C}_c^1(\mathcal{O}, \mathbb{R}^p), \ \|\varphi\|_\infty \leq 1 \right\}$$

where $\mathrm{C}_c^1(\mathcal{O}, \mathbb{R}^p)$ denotes the set of all the continuously differentiable functions defined on $\mathcal{O}$ with values in $\mathbb{R}^p$ that have a compact support.

The perimeter enjoys the basic properties we expect it does, namely (denoting $A \,\Delta\, B$ the symmetric difference of A and B, *i.e.* $A \,\Delta\, B = (A \setminus B) \cup (B \setminus A)$), for sake of curiosity

(i) $\mathrm{Per}(B, \mathcal{O}) = \mathrm{Per}(B^c, \mathcal{O})$,

(ii) $\mathrm{Per}(B, \mathcal{O}) = \mathrm{Per}(B', \mathcal{O})$ whenever $\mu((B \,\Delta\, B') \cap \mathcal{O}) = 0$,

(iii) $\mathrm{Per}(B \,\Delta\, B', \mathcal{O}) \leq \mathrm{Per}(B, \mathcal{O}) + \mathrm{Per}(B', \mathcal{O})$ and

(iv) $\mathrm{Per}(B \cup B', \mathcal{O}) + \mathrm{Per}(B \cap B', \mathcal{O}) \leq \mathrm{Per}(B, \mathcal{O}) + \mathrm{Per}(B', \mathcal{O})$,

for any $B, B' \in \mathcal{B}$ and $\mathcal{O}$ open set.

Moreover, one can extend the perimeter $\mathrm{Per}(B, \mathcal{O})$ as a function of the open set $\mathcal{O}$ to a measure on the Borel sets of $\mathcal{X}$ *via* the natural extension rule

$$\mathrm{Per}(B, F) = \inf \left\{ \mathrm{Per}(B, \mathcal{O}) : \mathcal{O} \supset F \right\}$$

for any $F \in \mathcal{B}$. The latter measure (denoted $\mathrm{Per}(B, \cdot)$) enjoys a remarkable property (see Lemma B.1 below) when B is a "smooth boundary" Caccioppoli set.

A set $B \in \mathcal{B}$ is a *Caccioppoli set* if, for any bounded open subset $\mathcal{O}$ of $\mathcal{X}$, one has

$$\mathrm{Per}(B, \mathcal{O}) < \infty.$$

The following lemma provides a simple class of Caccioppoli sets. We denote ∂B the topological boundary of B (*i.e.* $\partial B = \mathrm{cl}(B) \cap \mathrm{cl}(B^c)$) and H^q the q-dimensional Hausdorff dimension (see *e.g.* Ziemer).

Lemma B.1. *Let B be an open set with C^2 boundary. Then the extended Borel measure $\mathrm{Per}(B, \cdot)$ coincides with the restriction of the Hausdorff measure H^{p-1} to the boundary of B,* i.e.

$$\mathrm{Per}(B, F) = H^{p-1}(\partial B \cap F) \quad (\text{any } F \in \mathcal{B}).$$

B.2. Remark. Thus, let B be an open set with C^2 boundary: whenever B also satisfies

$$H^{p-1}(\partial B) < \infty,$$

B is a Caccioppoli set, since $\mathrm{Per}(B, \mathcal{O}) \leq H^{p-1}(\partial B) < \infty$ for any open set $\mathcal{O}$.

Partitions

A *subdivision* of $(\mathcal{X}, \mathcal{B}, P)$ is a sequence $\mathbb{T} = \{T_h\}_{h\geq 1}$ of elements of $\mathcal{B}$ (with possibly P-measure 0) such that, for any $h \neq k$,

$$P(T_k \cap T_h) = 0 \quad \text{and} \quad P\left(\mathcal{X} \setminus \bigcup_{h\geq 1} T_h\right) = 0.$$

We denote shortly (*i.e.* omitting $\mathcal{X}$, $\mathcal{B}$ and P) the collection Subd of all subdivisions of $(\mathcal{X}, \mathcal{B}, P)$. Moreover, we shall generally denote X_h the elements of the subdivision $\mathbb{X}$.

Let now $\mathrm{I}_{\mathbb{T}}$ be the *set of effective indices of* $\mathbb{T}$ (with implicitly $\mathbb{T} = \{T_h\}$):

$$\mathrm{I}_{\mathbb{T}} = \{h \geq 1 : P(T_h) > 0\},$$

each T_h for $h \in \mathrm{I}_{\mathbb{T}}$ being an *effective component* of $\mathbb{T}$. Two subdivisions $\mathbb{T}$ and $\mathbb{T}'$ are said to be *equivalent* if and only if there exists a bijection $\phi : \mathrm{I}_{\mathbb{T}} \to \mathrm{I}_{\mathbb{T}'}$ such that, for any $h \in \mathrm{I}_{\mathbb{T}}$,

$$T_h = T'_{\phi(h)}.$$

This defines an equivalence relation, and the set $\mathcal{P}$ of all the *partitions* of $\mathcal{X}$ is the quotient of Subd wrt this equivalence relation.

Thus, a partition $\tau \in \mathcal{P}$ is an equivalence class of subdivisions, *i.e.*, roughly speaking, that τ is a subdivision up to changes of P-measure 0 and rearrangements of its effective components. We shall later write sometimes $T_h \in \tau$ while we should have written $T_h \in \mathbb{T} \in \tau$ when considering effective components.

A metric on partitions

Let us introduce now

$$\delta_P(\mathbb{T}, \mathbb{T}') = \sum_{h\geq 1} P(T_h \,\Delta\, T'_h) \quad \text{and}$$
$$d_P(\tau, \tau') = \inf\left\{\delta_P(\mathbb{T}, \mathbb{T}') : \mathbb{T} \in \tau,\ \mathbb{T}' \in \tau'\right\}$$

for any $\tau, \tau' \in \mathcal{P}$ and $\mathbb{T} \in \tau$, $\mathbb{T}' \in \tau'$. The following result describes the properties of d_P:

Theorem B.3 (Leonardi, Tamanini). *Primo, d_P is a bounded distance on $\mathcal{P}$. Secundo, the space $(\mathcal{P}, d_P)$ of all partitions equipped with the topology induced by d_P is a complete metric space.*

Moreover, one can compare the metric convergence with the *columnwise convergence* for a suitable choice of the representatives, *i.e.*

Proposition B.4 (Leonardi, Tamanini). *Let $\{\tau^{(r)}\}$ be a sequence of elements of $\mathcal{P}$ and $\tau \in \mathcal{P}$. Then $d_P(\tau^{(r)}, \tau)$ goes to zero as r goes to infinity if and only if there exists some subdivisions $\mathbb{T} \in \tau$ and $\mathbb{T}^{(r)} \in \tau^{(r)}$ such that $P(T_h^{(r)} \,\Delta\, T_h)$ also does for all $h \geq 1$.*

Caccioppoli partitions

Finally, we define the set $\mathcal{C}$ of all the Caccioppoli partitions as the subset of $\mathcal{P}$ constituted of the partitions τ whose *perimeter* is locally finite, *i.e.* by definition such that

$$\mathrm{Per}(\tau, \mathcal{O}) = \sum_{h \geq 1} \mathrm{Per}(T_h, \mathcal{O}) < \infty,$$

for all open set $\mathcal{O}$ relatively compact in $\mathcal{X}$ and $\mathbb{T}$, an arbitrary representative of τ (of course, this definition is consistent). One can then prove that

Theorem B.5 (Leonardi, Tamanini). *Suppose that the sequence $\{\tau^{(r)}\}$ of elements of $\mathcal{C}$ satisfies*

$$\sup_{r \geq 1} \mathrm{Per}(\tau^{(r)}, \mathcal{O}) < \infty$$

for all open relatively compact set $\mathcal{O}$. Then $\{\tau^{(r)}\}$ converges along some subsequence to a partition $\tau \in \mathcal{C}$. Particularly, for any $\zeta > 0$, the subspace $\mathcal{T}$ of $\mathcal{C}$ constituted of all Caccioppoli partitions whose perimeters are uniformly bounded by ζ is a compact subspace of $\mathcal{P}$ when equipped with d_P.

Finally, Proposition B.4 allows to derive readily from the previous theorem that

Proposition B.6. *Let ζ be a positive number and $K \geq 1$ be an integer. The subspace $\mathcal{T}_{\leq K}$ constituted of all the Caccioppoli partitions τ whose perimeters are uniformly bounded by ζ and that satisfy the extra condition*

$$\mathrm{card}(\mathrm{I}_{\mathbb{T}}) \leq K \quad (\text{for some } \mathbb{T} \in \tau)$$

is a compact metric space when equipped with d_P.

Final comment

This particular class of partitions is one of the keys of the **AC** example of model where some results of ours do apply in spite of the great generality of the involved objects. Roughly speaking (with a view to the definition of a partition as set in Chapter 5), an element τ of $\mathcal{T}_{\leq K}$ is a collection $(\tau_k)_{1 \leq k \leq K}$ of at most K Caccioppoli sets of $\mathcal{X}$ with positive P-measure, satisfying for any $h \neq k$,

$$P(\tau_k \cap \tau_h) = 0 \quad \text{and} \quad P\left(\mathcal{X} \setminus \bigcup_{k=1}^{K} \tau_k\right) = 0.$$

Moreover, Proposition B.6 ensures that $\mathcal{T}_K$ is a compact metric space when equipped with d_P.

We did not know this elegant setting while we were working on Chapter 5. We think that the results should also hold true when replacing the "gap" between two partitions (see Definition 5.2.3 in Chapter 5 – this object is both a natural extension of its 1-dimensional ancestor and a technically useful tool that quantifies the likeness of two partitions, or two finite vectors of ordered times of abrupt changes in the 1-dimensional case) by their d_P-distance according to the setting above. Some future work ?

C

Divers tableaux

Les notations pour les indicateurs ont partiellement été introduites dans le Chapitre 4. Nous complétons ici la liste : `PM` et `PF` dénotent respectivement les populations masculine et féminine sans considération d'âge, avec les prolongements évidents `PM0, PM20, PM40, PM60, PM75` (idem pour la population féminine en substituant un `F` au `M`) ; `RP, RS, LV` et `LO` dénotent quant à eux respectivement les nombres de logements principaux, secondaires, vacants et occasionnels.

PT	PT0	PT20	PT40	PT60	PT75	PM	PM0	PM20	PM40	PM60
6	5	7	6	7	5	7	5	8	7	8
4	4	5	5	4	2	5	4	6	6	6
PM75	**PF**	**PF0**	**PF20**	**PF40**	**PF60**	**PF75**	**LOG**	**RP**	**RS**	**LV**
6	6	5	7	5	6	4	6	6	6	6
4	4	4	5	4	3	1	4	4	7	5
LO	**PLOG**	**158**	**#158**	**521**	**#521**	**522**	**#522**	**523**	**#523**	**524**
7	6	8	5	2	-4	4	0	11	3	1
4	4	11	8	5	-3	6	0	13	3	1
#524	**525**	**#525**	**527**	**#527**	**551**	**#551**	**552**	**#552**	**553**	**#553**
-2	10	9	12	3	12	5	3	1	9	12
-4	13	12	14	3	11	1	2	0	10	8
554	**#554**	**555**	**#555**	**633**	**#633**	**651**	**#651**	**714**	**#714**	**721**
6	7	19	9	8	3	2	3	6	6	7
11	9	16	9	11	7	2	5	6	9	8
#721	**722**	**#722**	**723**	**#723**	**724**	**#724**	**725**	**#725**	**726**	**#726**
16	8	5	9	9	0	1	2	10	?	?
10	8	7	10	8	6	1	5	10	?	?
731	**#731**	**732**	**#732**	**741**	**#741**	**742**	**#742**	**743**	**#743**	**744**
1	3	1	-2	3	2	7	15	4	-1	10
2	6	3	2	4	4	7	11	2	-2	13
#744	**745**	**#745**	**801**	**#801**	**802**	**#802**	**803**	**#803**	**804**	**#804**
6	19	20	7	4	6	0	-2	4	6	3
9	13	15	6	6	9	2	2	7	7	6
851	**#851**	**853**	**#853**	**921**	**#921**	**922**	**#922**	**923**	**#923**	**925**
2	0	4	4	2	1	2	1	10	5	8
2	-2	5	5	1	0	5	2	15	8	10

#925	926	#926	927	#927	930	#930
9	8	4	3	-3	10	6
7	6	2	8	-4	11	9

Tableau C.1 – Corrélations sur une échelle de -100 à 100 pour chacun des indicateurs INSEE vis à vis de la moyenne et de l'écart-type, calculés sur l'ensemble des cellules pour les données Cigale restreintes à la matinée. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à l'indicateur INSEE, à la moyenne et à l'écart-type.

PT	PT0	PT20	PT40	PT60	PT75	LOG	158	#158	521	#521
33	3	2	4	1	1	0	100	97	63	66
37	4	3	5	1	1	0	94	100	67	58
40	5	3	5	1	1	0	88	100	71	50
522	**#522**	**523**	**#523**	**524**	**#524**	**525**	**#525**	**527**	**#527**	**551**
9	65	43	53	66	46	15	17	2	1	5
7	75	44	51	71	33	11	16	1	0	4
5	85	45	47	72	21	7	13	1	0	4
#551	**552**	**#552**	**553**	**#553**	**554**	**#554**	**633**	**#633**	**651**	**#651**
27	0	0	26	16	4	9	31	1	12	4
26	0	0	27	20	3	9	27	1	8	3
25	0	0	25	23	2	10	23	1	5	2
714	**#714**	**721**	**#721**	**722**	**#722**	**723**	**#723**	**724**	**#724**	**725**
0	7	42	16	9	25	0	5	1	9	0
0	5	58	16	7	27	0	4	1	5	0
0	3	76	16	5	29	0	2	1	3	0
#725	**726**	**#726**	**731**	**#731**	**732**	**#732**	**741**	**#741**	**742**	**#742**
0	0	0	6	12	12	0	51	3	0	5
0	0	0	5	12	8	0	50	3	0	4
0	0	0	4	12	5	0	47	3	0	3
743	**#743**	**744**	**#744**	**745**	**#745**	**801**	**#801**	**802**	**#802**	**803**
0	0	1	3	0	18	2	0	0	0	2
0	0	1	2	0	18	2	0	0	0	2
0	0	0	1	0	18	2	0	0	0	1
#803	**804**	**#804**	**851**	**#851**	**853**	**#853**	**921**	**#921**	**922**	**#922**
0	0	4	4	8	11	16	0	0	0	0
0	0	4	3	7	10	13	0	0	0	0
0	0	3	2	6	8	10	0	0	0	0

923	#923	925	#925	926	#926	927	#927	930	#930
1	0	0	0	14	0	10	0	4	6
1	0	0	0	10	0	7	0	4	6
1	0	0	0	7	0	5	0	4	5

Tableau C.2 – Importances Δ_δ des variables INSEE échelonnées entre 0 et 100 et calculées lors de la construction d'un unique arbre de régression **CART** pour les données Cigale restreintes à la matinée. L'arbre en question est présenté dans la Figure 4.1. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à un coefficient δ de pénalisation de la profondeur égal à 1, 0.9 et 0.8.

PT	PT0	PT20	PT40	PT60	PT75	LOG	158	#158	521	#521
37	9	5	5	12	15	2	88	66	100	76
39	10	6	5	12	14	2	85	63	100	73
42	11	6	6	12	13	2	82	59	100	70
522	**#522**	**523**	**#523**	**524**	**#524**	**525**	**#525**	**527**	**#527**	**551**
23	44	54	28	39	27	17	10	8	3	17
22	47	54	26	41	27	15	9	8	2	17
21	52	54	25	42	26	14	8	7	2	18
#551	**552**	**#552**	**553**	**#553**	**554**	**#554**	**633**	**#633**	**651**	**#651**
11	0	0	30	13	6	10	7	3	6	4
10	0	0	30	13	6	10	7	3	5	4
10	0	0	29	13	5	10	7	2	4	3
714	**#714**	**721**	**#721**	**722**	**#722**	**723**	**#723**	**724**	**#724**	**725**
4	4	27	8	7	5	3	3	4	1	1
4	4	29	8	7	5	2	3	4	1	1
3	3	32	8	7	4	2	2	3	1	1
#725	**726**	**#726**	**731**	**#731**	**732**	**#732**	**741**	**#741**	**742**	**#742**
1	0	0	5	5	4	1	35	4	11	4
1	0	0	4	5	4	1	35	4	11	4
1	0	0	3	5	4	1	35	4	10	3
743	**#743**	**744**	**#744**	**745**	**#745**	**801**	**#801**	**802**	**#802**	**803**
1	1	5	5	6	6	2	0	1	0	2
1	1	5	5	7	7	2	0	1	0	2
1	1	4	5	8	8	2	0	1	0	2
#803	**804**	**#804**	**851**	**#851**	**853**	**#853**	**921**	**#921**	**922**	**#922**
1	6	2	13	4	2	2	1	0	1	1
1	6	2	12	4	2	2	0	0	0	1
1	6	2	11	3	2	2	0	0	0	1

923	#923	925	#925	926	#926	927	#927	930	#930
3	1	1	0	1	0	1	0	13	4
4	1	1	0	1	0	2	0	13	4
4	1	1	0	1	0	2	1	13	4

Tableau C.3 – Importances Δ_δ des variables INSEE échelonnées entre 0 et 100 et calculées lors de la procédure de **Bagging** appliquées aux données Cigale restreintes à la matinée. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à un coefficient δ de pénalisation de la profondeur égal à 1, 0.9 et 0.8.

PT	PT0	PT20	PT40	PT60	PT75	PM	PM0	PM20	PM40	PM60
4	3	6	4	4	2	5	2	6	5	5
7	5	8	7	8	7	8	5	9	8	10
PM75	**PF**	**PF0**	**PF20**	**PF40**	**PF60**	**PF75**	**LOG**	**RP**	**RS**	**LV**
3	3	3	5	3	3	1	5	5	10	9
8	7	5	8	5	7	7	8	8	8	12
LO	**PLOG**	**158**	**#158**	**521**	**#521**	**522**	**#522**	**523**	**#523**	**524**
8	4	17	11	7	-1	14	5	19	6	8
9	7	12	9	6	-2	9	3	9	3	8
#524	**525**	**#525**	**527**	**#527**	**551**	**#551**	**552**	**#552**	**553**	**#553**
3	19	19	18	4	19	2	4	3	18	17
1	6	6	13	5	8	2	1	1	12	7
554	**#554**	**555**	**#555**	**633**	**#633**	**651**	**#651**	**714**	**#714**	**721**
15	12	19	9	18	6	10	5	12	6	16
12	7	8	-1	11	3	2	-1	11	1	15
#721	**722**	**#722**	**723**	**#723**	**724**	**#724**	**725**	**#725**	**726**	**#726**
15	17	7	17	11	10	3	9	9	?	?
3	12	3	14	10	5	0	9	4	?	?
731	**#731**	**732**	**#732**	**741**	**#741**	**742**	**#742**	**743**	**#743**	**744**
1	2	1	-1	7	4	14	15	2	-1	18
2	1	1	-2	4	1	10	2	5	4	17
#744	**745**	**#745**	**801**	**#801**	**802**	**#802**	**803**	**#803**	**804**	**#804**
14	22	22	12	4	11	0	2	5	12	6
12	13	11	14	4	11	0	1	6	8	4
851	**#851**	**853**	**#853**	**921**	**#921**	**922**	**#922**	**923**	**#923**	**925**
5	-1	5	3	6	3	6	2	18	11	12
2	-1	8	7	2	0	5	0	11	3	2
#925	**926**	**#926**	**927**	**#927**	**930**	**#930**				
10	14	1	13	-2	19	14				
2	8	-2	11	-4	14	7				

Tableau C.4 – Corrélations sur une échelle de -100 à 100 pour chacun des indicateurs INSEE vis à vis de la moyenne et de l'écart-type, calculés sur l'ensemble des cellules pour les données Cigale restreintes à la mi-journée. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à l'indicateur INSEE, à la moyenne et à l'écart-type.

PT	PT0	PT20	PT40	PT60	PT75	LOG	158	#158	521	#521
7	2	0	0	1	0	0	53	76	100	82
8	3	0	0	1	0	0	45	62	100	71
9	3	0	0	2	0	0	38	48	100	60
522	**#522**	**523**	**#523**	**524**	**#524**	**525**	**#525**	**527**	**#527**	**551**
11	69	7	15	32	50	25	7	13	0	42
9	53	7	14	25	43	24	6	12	0	58
7	38	7	13	18	36	22	5	11	0	79
#551	**552**	**#552**	**553**	**#553**	**554**	**#554**	**633**	**#633**	**651**	**#651**
8	0	0	7	6	28	14	10	7	8	3
5	0	0	6	6	22	16	8	5	6	3
3	0	0	4	6	17	17	6	3	5	3
714	**#714**	**721**	**#721**	**722**	**#722**	**723**	**#723**	**724**	**#724**	**725**
4	10	7	12	1	10	13	12	3	10	0
3	9	7	10	1	9	9	15	3	6	0
2	8	6	8	1	7	5	18	2	3	0
#725	**726**	**#726**	**731**	**#731**	**732**	**#732**	**741**	**#741**	**742**	**#742**
0	0	0	12	2	3	0	30	8	3	4
0	0	0	9	1	3	0	28	5	2	3
0	0	0	7	1	2	0	24	2	1	2
743	**#743**	**744**	**#744**	**745**	**#745**	**801**	**#801**	**802**	**#802**	**803**
9	0	4	13	1	5	9	0	2	0	3
8	0	3	12	1	5	6	0	1	0	2
7	0	3	10	0	4	3	0	1	0	2
#803	**804**	**#804**	**851**	**#851**	**853**	**#853**	**921**	**#921**	**922**	**#922**
0	1	2	0	2	9	0	1	0	0	9
0	1	2	0	2	6	0	1	0	0	6
0	1	1	0	1	4	0	1	0	0	3

923	#923	925	#925	926	#926	927	#927	930	#930
2	0	0	0	0	0	3	0	17	0
2	0	0	0	0	0	2	0	14	0
1	0	0	0	0	0	2	0	11	0

Tableau C.5 – Importances Δ_δ des variables INSEE échelonnées entre 0 et 100 et calculées lors de la construction d'un unique arbre de régression **CART** pour les données Cigale restreintes à la mi-journée. L'arbre en question est présenté dans la Figure 4.5. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à un coefficient δ de pénalisation de la profondeur égal à 1, 0.9 et 0.8.

PT	PT0	PT20	PT40	PT60	PT75	LOG	158	#158	521	#521
23	9	5	3	6	9	1	100	65	60	72
26	10	5	3	6	9	1	100	69	57	72
29	11	6	4	6	9	1	99	72	52	70
522	**#522**	**523**	**#523**	**524**	**#524**	**525**	**#525**	**527**	**#527**	**551**
35	76	67	17	25	27	36	14	16	3	46
34	87	76	16	23	25	38	14	17	3	58
33	100	87	15	20	23	41	13	18	3	74
#551	**552**	**#552**	**553**	**#553**	**554**	**#554**	**633**	**#633**	**651**	**#651**
12	0	0	23	9	5	13	14	8	9	6
12	0	0	23	9	4	14	14	8	9	6
11	0	0	23	10	4	14	14	7	8	6
714	**#714**	**721**	**#721**	**722**	**#722**	**723**	**#723**	**724**	**#724**	**725**
4	7	9	7	12	7	6	7	3	3	2
4	7	8	7	15	8	6	7	3	3	1
3	7	8	7	18	8	5	7	2	3	1
#725	**726**	**#726**	**731**	**#731**	**732**	**#732**	**741**	**#741**	**742**	**#742**
1	0	0	5	4	6	2	19	5	5	5
0	0	0	4	4	6	2	19	4	5	5
0	0	0	4	4	5	2	18	4	4	4
743	**#743**	**744**	**#744**	**745**	**#745**	**801**	**#801**	**802**	**#802**	**803**
2	2	4	7	9	6	3	0	3	1	2
1	1	4	7	11	7	3	0	3	1	2
1	1	5	7	13	8	3	0	2	1	1
#803	**804**	**#804**	**851**	**#851**	**853**	**#853**	**921**	**#921**	**922**	**#922**
2	3	3	5	4	2	1	1	1	1	1
2	3	3	5	4	2	1	1	1	1	1
2	3	2	4	4	2	0	1	0	1	1
923	**#923**	**925**	**#925**	**926**	**#926**	**927**	**#927**	**930**	**#930**	
3	1	1	0	1	0	1	1	5	4	
3	1	1	0	1	0	1	1	5	4	
3	1	1	1	1	0	0	0	5	4	

Tableau C.6 – Importances Δ_δ des variables INSEE échelonnées entre 0 et 100 et calculées lors de la procédure de **Bagging** appliquées aux données Cigale restreintes à la mi-journée. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à un coefficient δ de pénalisation de la profondeur égal à 1, 0.9 et 0.8.

PT	PT0	PT20	PT40	PT60	PT75	PM	PM0	PM20	PM40	PM60
5	3	6	5	5	4	6	3	7	6	6
9	7	11	8	10	10	10	7	11	9	12
PM75	**PF**	**PF0**	**PF20**	**PF40**	**PF60**	**PF75**	**LOG**	**RP**	**RS**	**LV**
5	4	3	5	4	4	3	6	5	10	8
11	9	7	10	7	9	10	11	10	8	14
LO	**PLOG**	**158**	**#158**	**521**	**#521**	**522**	**#522**	**523**	**#523**	**524**
9	5	14	11	6	-1	8	2	17	4	4
13	9	15	8	6	-2	12	4	8	1	6
#524	**525**	**#525**	**527**	**#527**	**551**	**#551**	**552**	**#552**	**553**	**#553**
-3	18	18	17	2	17	6	7	4	20	16
-3	2	1	17	4	8	2	2	3	12	6
554	**#554**	**555**	**#555**	**633**	**#633**	**651**	**#651**	**714**	**#714**	**721**
14	13	17	9	17	7	6	5	7	4	14
8	7	11	-1	9	2	0	0	15	9	14
#721	**722**	**#722**	**723**	**#723**	**724**	**#724**	**725**	**#725**	**726**	**#726**
16	15	8	14	11	9	2	6	11	?	?
10	15	6	16	8	6	5	9	17	?	?
731	**#731**	**732**	**#732**	**741**	**#741**	**742**	**#742**	**743**	**#743**	**744**
3	4	2	-2	7	4	15	17	4	-2	17
5	3	2	-3	3	-1	14	5	4	-4	16
#744	**745**	**#745**	**801**	**#801**	**802**	**#802**	**803**	**#803**	**804**	**#804**
7	22	20	10	3	7	0	2	6	8	5
2	12	7	12	1	11	8	1	0	7	3
851	**#851**	**853**	**#853**	**921**	**#921**	**922**	**#922**	**923**	**#923**	**925**
2	2	4	3	7	1	6	-1	19	10	11
3	5	7	3	3	1	10	-2	6	1	6

#925	926	#926	927	#927	930	#930
10	9	2	7	-2	18	13
10	7	1	9	-4	16	7

Tableau C.7 – Corrélations sur une échelle de -100 à 100 pour chacun des indicateurs INSEE vis à vis de la moyenne et de l'écart-type, calculés sur l'ensemble des cellules pour les données Cigale restreintes à l'après-midi. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à l'indicateur INSEE, à la moyenne et à l'écart-type.

PT	PT0	PT20	PT40	PT60	PT75	LOG	158	#158	521	#521
33	3	2	4	1	1	0	100	97	63	66
37	4	3	5	1	1	0	94	100	67	58
40	5	3	5	1	1	0	88	100	71	50
522	**#522**	**523**	**#523**	**524**	**#524**	**525**	**#525**	**527**	**#527**	**551**
9	65	43	53	66	46	15	17	2	1	5
7	75	44	51	71	33	11	16	1	0	4
5	85	45	47	72	21	7	13	1	0	4
#551	**552**	**#552**	**553**	**#553**	**554**	**#554**	**633**	**#633**	**651**	**#651**
27	0	0	26	16	4	9	31	1	12	4
26	0	0	27	20	3	9	27	1	8	3
25	0	0	25	23	2	10	23	1	5	2
714	**#714**	**721**	**#721**	**722**	**#722**	**723**	**#723**	**724**	**#724**	**725**
0	7	42	16	9	25	0	5	1	9	0
0	5	58	16	7	27	0	4	1	5	0
0	3	76	16	5	29	0	2	1	3	0
#725	**726**	**#726**	**731**	**#731**	**732**	**#732**	**741**	**#741**	**742**	**#742**
0	0	0	6	12	12	0	51	3	0	5
0	0	0	5	12	8	0	50	3	0	4
0	0	0	4	12	5	0	47	3	0	3
743	**#743**	**744**	**#744**	**745**	**#745**	**801**	**#801**	**802**	**#802**	**803**
0	0	1	3	0	18	2	0	0	0	2
0	0	1	2	0	18	2	0	0	0	2
0	0	0	1	0	18	2	0	0	0	1
#803	**804**	**#804**	**851**	**#851**	**853**	**#853**	**921**	**#921**	**922**	**#922**
0	0	4	4	8	11	16	0	0	0	0
0	0	4	3	7	10	13	0	0	0	0
0	0	3	2	6	8	10	0	0	0	0
923	**#923**	**925**	**#925**	**926**	**#926**	**927**	**#927**	**930**	**#930**	
1	0	0	0	14	0	10	0	4	6	
1	0	0	0	10	0	7	0	4	6	
1	0	0	0	7	0	5	0	4	5	

Tableau C.8 – Importances Δ_δ des variables INSEE échelonnées entre 0 et 100 et calculées lors de la construction d'un unique arbre de régression **CART** pour les données Cigale restreintes à l'après-midi. L'arbre en question est présenté dans la Figure 4.9. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à un coefficient δ de pénalisation de la profondeur égal à 1, 0.9 et 0.8.

PT	PT0	PT20	PT40	PT60	PT75	LOG	158	#158	521	#521
25	8	6	4	9	11	1	100	83	60	87
24	8	6	4	8	10	1	95	79	51	80
21	7	6	4	7	9	1	80	67	38	63
522	**#522**	**523**	**#523**	**524**	**#524**	**525**	**#525**	**527**	**#527**	**551**
40	91	69	26	22	26	35	15	29	6	45
41	100	72	24	19	23	32	13	32	5	50
37	100	68	19	15	17	25	10	31	4	51
#551	**552**	**#552**	**553**	**#553**	**554**	**#554**	**633**	**#633**	**651**	**#651**
15	1	0	27	8	8	15	15	11	10	6
14	1	0	26	8	7	15	14	10	8	6
12	0	0	22	7	6	14	12	8	6	5
714	**#714**	**721**	**#721**	**722**	**#722**	**723**	**#723**	**724**	**#724**	**725**
6	10	9	8	22	9	5	10	5	3	1
5	8	8	7	27	9	5	9	4	3	1
4	6	6	6	29	7	4	8	3	2	0
#725	**726**	**#726**	**731**	**#731**	**732**	**#732**	**741**	**#741**	**742**	**#742**
1	0	0	7	7	8	3	18	5	5	6
1	0	0	6	6	6	2	17	4	5	5
0	0	0	4	5	5	2	14	3	4	4
743	**#743**	**744**	**#744**	**745**	**#745**	**801**	**#801**	**802**	**#802**	**803**
2	2	4	8	9	7	4	0	3	0	3
1	1	4	7	10	7	3	0	2	0	2
1	1	3	6	11	7	2	0	2	0	2
#803	**804**	**#804**	**851**	**#851**	**853**	**#853**	**921**	**#921**	**922**	**#922**
2	2	4	4	3	2	1	1	0	1	1
2	2	3	4	3	2	1	1	0	1	1
2	1	2	3	2	1	1	1	0	1	1
923	**#923**	**925**	**#925**	**926**	**#926**	**927**	**#927**	**930**	**#930**	
2	3	2	1	1	0	1	1	7	4	
2	3	2	2	1	0	1	1	7	3	
1	2	2	2	1	0	1	0	5	3	

Tableau C.9 – Importances Δ_δ des variables INSEE échelonnées entre 0 et 100 et calculées lors de la procédure de **Bagging** appliquées aux données Cigale restreintes à l'après-midi. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à un coefficient δ de pénalisation de la profondeur égal à 1, 0.9 et 0.8.

PT	PT0	PT20	PT40	PT60	PT75	PM	PM0	PM20	PM40	PM60
12	10	15	11	12	12	13	10	15	13	12
10	8	12	9	9	9	10	8	12	11	10
PM75	**PF**	**PF0**	**PF20**	**PF40**	**PF60**	**PF75**	**LOG**	**RP**	**RS**	**LV**
12	12	10	14	11	12	11	15	14	17	18
9	9	8	11	8	9	9	12	11	10	13
LO	**PLOG**	**158**	**#158**	**521**	**#521**	**522**	**#522**	**523**	**#523**	**524**
14	12	25	17	15	-3	19	0	20	4	2
11	10	18	12	14	-4	16	2	11	0	2
#524	**525**	**#525**	**527**	**#527**	**551**	**#551**	**552**	**#552**	**553**	**#553**
-2	9	9	24	10	16	5	5	4	24	13
-5	5	5	14	5	17	3	11	8	12	2
554	**#554**	**555**	**#555**	**633**	**#633**	**651**	**#651**	**714**	**#714**	**721**
19	12	11	3	19	6	8	2	15	4	24
11	6	5	-1	11	4	3	7	12	0	18
#721	**722**	**#722**	**723**	**#723**	**724**	**#724**	**725**	**#725**	**726**	**#726**
11	22	5	19	7	13	3	12	6	?	?
3	14	4	14	14	8	2	12	10	?	?
731	**#731**	**732**	**#732**	**741**	**#741**	**742**	**#742**	**743**	**#743**	**744**
-1	1	3	-1	9	3	23	11	9	2	23
0	1	0	0	5	3	18	6	5	2	17
#744	**745**	**#745**	**801**	**#801**	**802**	**#802**	**803**	**#803**	**804**	**#804**
7	19	1	22	1	9	-1	-1	3	9	1
9	12	2	15	0	5	-5	-2	-1	4	-2
851	**#851**	**853**	**#853**	**921**	**#921**	**922**	**#922**	**923**	**#923**	**925**
6	3	13	7	8	2	8	1	21	7	8
3	1	9	7	5	0	3	2	14	3	4

#925	926	#926	927	#927	930	#930
6	15	2	10	-3	28	14
5	11	3	8	-3	17	7

Tableau C.10 – Corrélations sur une échelle de -100 à 100 pour chacun des indicateurs INSEE vis à vis de la moyenne et de l'écart-type, calculés sur l'ensemble des cellules pour les données Cigale restreintes à la soirée. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à l'indicateur INSEE, à la moyenne et à l'écart-type.

PT	PT0	PT20	PT40	PT60	PT75	LOG	158	#158	521	#521
18	5	6	2	2	8	0	55	38	83	100
16	5	6	3	3	10	0	57	30	83	100
16	5	5	4	3	12	0	62	23	83	100
522	**#522**	**523**	**#523**	**524**	**#524**	**525**	**#525**	**527**	**#527**	**551**
30	21	16	35	42	38	25	18	11	0	9
24	18	13	38	42	33	25	15	13	0	7
19	14	11	41	42	28	24	12	17	0	5
#551	**552**	**#552**	**553**	**#553**	**554**	**#554**	**633**	**#633**	**651**	**#651**
8	0	0	11	4	0	4	6	7	3	1
7	0	0	13	4	0	3	6	4	2	1
6	0	0	15	3	0	2	6	2	2	1
714	**#714**	**721**	**#721**	**722**	**#722**	**723**	**#723**	**724**	**#724**	**725**
12	0	12	1	1	2	1	8	0	17	0
9	0	10	1	1	1	1	5	0	15	0
6	0	7	0	1	1	1	4	0	12	0
#725	**726**	**#726**	**731**	**#731**	**732**	**#732**	**741**	**#741**	**742**	**#742**
2	0	0	0	3	1	3	6	6	9	2
1	0	0	0	2	0	3	6	5	9	2
1	0	0	0	1	0	3	5	4	8	2
743	**#743**	**744**	**#744**	**745**	**#745**	**801**	**#801**	**802**	**#802**	**803**
1	0	22	2	1	0	17	0	2	0	0
0	0	26	1	0	0	12	0	1	0	0
0	0	31	1	0	0	8	0	1	0	0
#803	**804**	**#804**	**851**	**#851**	**853**	**#853**	**921**	**#921**	**922**	**#922**
0	2	12	2	9	0	0	0	0	0	1
0	3	14	2	6	0	0	0	0	0	1
0	3	17	1	4	0	0	0	0	0	1
923	**#923**	**925**	**#925**	**926**	**#926**	**927**	**#927**	**930**	**#930**	
2	5	0	0	0	0	5	4	24	4	
2	5	0	0	0	0	4	3	32	3	
2	5	0	0	0	0	3	2	43	2	

Tableau C.11 – Importances Δ_δ des variables INSEE échelonnées entre 0 et 100 et calculées lors de la construction d'un unique arbre de régression **CART** pour les données Cigale restreintes à la soirée. L'arbre en question est présenté dans la Figure 4.13. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à un coefficient δ de pénalisation de la profondeur égal à 1, 0.9 et 0.8.

PT	PT0	PT20	PT40	PT60	PT75	LOG	158	#158	521	#521
18	6	8	3	9	10	2	100	42	44	58
20	7	10	4	9	9	2	100	39	41	55
23	7	12	5	10	9	2	100	36	38	52
522	**#522**	**523**	**#523**	**524**	**#524**	**525**	**#525**	**527**	**#527**	**551**
28	27	38	46	22	30	31	11	13	3	10
27	26	39	47	21	27	32	11	13	3	10
26	26	40	49	19	25	32	10	12	3	9
#551	**552**	**#552**	**553**	**#553**	**554**	**#554**	**633**	**#633**	**651**	**#651**
17	0	0	19	8	12	5	12	7	13	3
17	0	0	20	7	12	5	13	7	13	3
16	0	0	20	6	12	5	13	7	12	2
714	**#714**	**721**	**#721**	**722**	**#722**	**723**	**#723**	**724**	**#724**	**725**
3	2	9	5	5	6	4	4	5	6	2
3	2	9	4	5	5	4	3	6	5	2
2	2	9	4	5	5	4	3	6	4	2
#725	**726**	**#726**	**731**	**#731**	**732**	**#732**	**741**	**#741**	**742**	**#742**
3	0	0	8	4	4	2	25	4	6	3
3	0	0	7	4	3	2	27	3	6	2
2	0	0	7	3	3	2	29	3	6	2
743	**#743**	**744**	**#744**	**745**	**#745**	**801**	**#801**	**802**	**#802**	**803**
1	1	6	3	3	2	3	0	2	0	2
1	1	6	3	3	2	3	0	2	0	2
1	0	7	2	3	2	3	0	2	0	2
#803	**804**	**#804**	**851**	**#851**	**853**	**#853**	**921**	**#921**	**922**	**#922**
2	3	4	6	2	2	1	1	1	0	1
1	3	4	6	2	1	1	1	1	0	1
1	2	4	6	1	1	0	1	1	0	1
923	**#923**	**925**	**#925**	**926**	**#926**	**927**	**#927**	**930**	**#930**	
3	2	1	0	1	0	1	5	26	2	
3	2	0	0	1	0	1	5	30	2	
3	1	0	0	0	0	1	6	35	2	

Tableau C.12 – Importances Δ_δ des variables INSEE échelonnées entre 0 et 100 et calculées lors de la procédure de **Bagging** appliquées aux données Cigale restreintes à la soirée. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à un coefficient δ de pénalisation de la profondeur égal à 1, 0.9 et 0.8.

PT	PT0	PT20	PT40	PT60	PT75	PM	PM0	PM20	PM40	PM60
9	7	11	8	9	7	10	7	12	10	9
16	13	18	15	17	15	17	13	19	17	18
PM75	**PF**	**PF0**	**PF20**	**PF40**	**PF60**	**PF75**	**LOG**	**RP**	**RS**	**LV**
8	8	7	10	7	8	7	10	10	13	13
16	15	13	18	14	16	15	18	17	14	21
LO	**PLOG**	**158**	**#158**	**521**	**#521**	**522**	**#522**	**523**	**#523**	**524**
12	9	20	13	9	-2	14	3	20	5	4
16	16	22	12	12	-1	17	3	18	5	5
#524	**525**	**#525**	**527**	**#527**	**551**	**#551**	**552**	**#552**	**553**	**#553**
0	17	16	21	4	19	5	6	4	21	17
-1	9	10	23	7	18	4	6	3	20	12
554	**#554**	**555**	**#555**	**633**	**#633**	**651**	**#651**	**714**	**#714**	**721**
15	13	20	9	19	7	8	5	13	6	19
15	11	18	5	18	7	6	4	18	8	25
#721	**722**	**#722**	**723**	**#723**	**724**	**#724**	**725**	**#725**	**726**	**#726**
17	19	8	19	12	10	3	11	13	?	?
13	22	7	23	14	10	4	13	13	?	?
731	**#731**	**732**	**#732**	**741**	**#741**	**742**	**#742**	**743**	**#743**	**744**
1	3	1	-2	8	4	18	18	4	-1	21
1	3	0	-2	8	1	19	10	9	2	24
#744	**745**	**#745**	**801**	**#801**	**802**	**#802**	**803**	**#803**	**804**	**#804**
11	24	18	15	4	10	0	0	4	11	5
13	23	18	21	1	14	1	-2	3	11	3
851	**#851**	**853**	**#853**	**921**	**#921**	**922**	**#922**	**923**	**#923**	**925**
4	1	7	5	7	3	7	1	20	10	11
6	1	14	10	7	2	7	1	16	6	7

#925	926	#926	927	#927	930	#930
11	15	3	11	-3	23	15
8	15	4	13	-4	26	14

Tableau C.13 – Corrélations sur une échelle de -100 à 100 pour chacun des indicateurs INSEE vis à vis de la moyenne et de l'écart-type, calculés sur l'ensemble des cellules pour les données Cigale sur la journée complète. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à l'indicateur INSEE, à la moyenne et à l'écart-type.

H
100
100
100

PT	PT0	PT20	PT40	PT60	PT75	158	#158	521	#521	522
16	10	0	4	3	3	28	24	15	27	6
13	9	0	3	2	2	20	16	12	21	4
11	8	1	2	1	1	13	10	9	16	2
#522	**523**	**#523**	**524**	**#524**	**525**	**#525**	**527**	**#527**	**551**	**#551**
19	13	8	3	22	21	10	2	3	6	3
19	10	7	2	15	18	7	1	2	6	2
20	8	6	1	10	15	4	1	1	6	2
552	**#552**	**553**	**#553**	**554**	**#554**	**555**	**#555**	**633**	**#633**	**651**
0	0	7	4	2	2	2	3	5	0	5
0	0	5	2	1	1	1	2	4	0	3
0	0	3	1	1	1	1	1	3	0	2
#651	**714**	**#714**	**721**	**#721**	**722**	**#722**	**723**	**#723**	**724**	**#724**
4	1	0	1	2	2	2	2	0	0	4
3	0	0	1	1	1	1	2	0	0	3
2	0	0	0	1	1	1	1	0	0	2
725	**#725**	**726**	**#726**	**731**	**#731**	**732**	**#732**	**741**	**#741**	**742**
0	1	0	0	0	1	3	0	5	1	3
0	1	0	0	0	0	2	0	5	0	2
0	0	0	0	0	0	1	0	4	0	1
#742	**743**	**#743**	**744**	**#744**	**745**	**#745**	**801**	**#801**	**802**	**#802**
0	0	0	0	0	3	2	0	0	0	0
0	0	0	0	0	3	2	0	0	0	0
0	0	0	0	0	2	2	0	0	0	0
803	**#803**	**804**	**#804**	**851**	**#851**	**853**	**#853**	**921**	**#921**	**922**
0	0	1	2	5	3	1	0	0	0	0
0	0	1	1	5	2	1	0	0	0	0
0	0	0	1	4	1	0	0	0	0	0
#922	**923**	**#923**	**925**	**#925**	**926**	**#926**	**927**	**#927**	**930**	**#930**
2	0	4	0	0	0	0	1	0	9	0
1	0	3	0	0	0	0	1	0	10	0
1	0	3	0	0	0	0	0	0	10	0

Tableau C.14 – Importances Δ_δ des variables INSEE échelonnées entre 0 et 100 et calculées lors de la construction d'un unique arbre de régression **CART** pour les données Cigale sur une journée. L'arbre en question est présenté dans la Figure 4.17. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à un coefficient δ de pénalisation de la profondeur égal à 1, 0.9 et 0.8. On souligne que l'indication de plage horaire **H** est la plus importante quel que soit δ.

H
100
100
100

PT	PT0	PT20	PT40	PT60	PT75	158	#158	521	#521	522
12	3	2	2	2	3	26	12	12	15	5
12	2	2	2	2	3	26	12	12	14	5
11	2	2	2	1	3	25	12	12	14	5
#522	**523**	**#523**	**524**	**#524**	**525**	**#525**	**527**	**#527**	**551**	**#551**
18	14	9	2	5	4	1	2	0	5	5
18	13	9	2	5	4	1	2	0	5	5
19	13	9	2	5	4	1	2	0	5	5
552	**#552**	**553**	**#553**	**554**	**#554**	**555**	**#555**	**633**	**#633**	**651**
0	0	6	2	0	2	4	1	2	2	5
0	0	6	2	0	2	4	1	2	2	5
0	0	6	2	0	2	4	1	2	2	5
#651	**714**	**#714**	**721**	**#721**	**722**	**#722**	**723**	**#723**	**724**	**#724**
1	0	1	3	2	8	1	0	1	1	0
1	0	1	3	2	8	1	0	1	1	0
1	0	1	3	2	9	1	0	1	0	0
725	**#725**	**726**	**#726**	**731**	**#731**	**732**	**#732**	**741**	**#741**	**742**
0	0	0	0	1	1	0	0	5	0	1
0	0	0	0	1	1	0	0	5	0	1
0	0	0	0	1	1	0	0	5	0	1
#742	**743**	**#743**	**744**	**#744**	**745**	**#745**	**801**	**#801**	**802**	**#802**
0	0	0	1	1	3	0	1	0	0	0
0	0	0	1	1	3	0	1	0	0	0
0	0	0	1	1	3	0	1	0	0	0
803	**#803**	**804**	**#804**	**851**	**#851**	**853**	**#853**	**921**	**#921**	**922**
0	0	1	0	5	1	0	0	0	0	0
0	0	1	0	5	1	0	0	0	0	0
0	0	0	0	4	1	0	0	0	0	0
#922	**923**	**#923**	**925**	**#925**	**926**	**#926**	**927**	**#927**	**930**	**#930**
0	1	2	0	0	0	0	0	0	4	1
0	1	2	0	0	0	0	0	0	4	1
0	1	2	0	0	0	0	0	0	4	1

Tableau C.15 – Importances Δ_δ des variables INSEE échelonnées entre 0 et 100 et calculées lors de la procédure de **Bagging** appliquées aux données Cigale sur une journée. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à un coefficient δ de pénalisation de la profondeur égal à 1, 0.9 et 0.8. On souligne que l'indication de plage horaire **H** est la plus importante quel que soit δ.

PT	PT0	PT20	PT40	PT60	PT75	158	#158	521	#521	522
11	8	12	10	10	11	24	16	13	6	22
1	-1	2	0	2	3	3	5	3	1	4
#522	**523**	**#523**	**524**	**#524**	**525**	**#525**	**527**	**#527**	**551**	**#551**
15	25	9	15	0	15	12	26	11	25	9
4	7	2	4	1	5	4	5	3	9	4
552	**#552**	**553**	**#553**	**554**	**#554**	**555**	**#555**	**633**	**#633**	**651**
2	1	24	15	21	17	20	9	24	16	15
-2	-1	8	9	7	7	8	7	8	7	4
#651	**714**	**#714**	**721**	**#721**	**722**	**#722**	**723**	**#723**	**724**	**#724**
5	19	10	30	15	27	11	26	9	23	11
0	3	1	7	5	8	4	10	2	6	3
725	**#725**	**726**	**#726**	**731**	**#731**	**732**	**#732**	**741**	**#741**	**742**
16	4	?	?	8	5	7	-1	14	4	21
2	-2	?	?	1	0	1	2	4	-1	8
#742	**743**	**#743**	**744**	**#744**	**745**	**#745**	**801**	**#801**	**802**	**#802**
11	11	0	21	9	21	12	18	1	17	6
5	0	-1	4	1	10	2	0	1	4	2
803	**#803**	**804**	**#804**	**851**	**#851**	**853**	**#853**	**921**	**#921**	**922**
1	5	16	5	11	2	12	9	11	8	10
2	3	5	4	0	0	1	4	4	3	2
#922	**923**	**#923**	**925**	**#925**	**926**	**#926**	**927**	**#927**	**930**	**#930**
1	20	7	9	7	15	7	15	3	31	22
6	9	1	0	-3	7	8	2	1	9	5

Tableau C.16 – Corrélations sur une échelle de -100 à 100 pour chacun des indicateurs INSEE vis à vis de la moyenne et de l'écart-type, calculés sur l'ensemble des cellules pour les données HC2. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à l'indicateur INSEE, à la moyenne et à l'écart-type.

PT	PT0	PT20	PT40	PT60	PT75	158	#158	521	#521	522
37	2	5	0	0	4	77	100	62	37	62
40	2	7	0	0	4	68	89	47	31	56
30	2	7	0	0	3	40	55	23	18	34
#522	**523**	**#523**	**524**	**#524**	**525**	**#525**	**527**	**#527**	**551**	**#551**
36	77	37	23	21	43	31	4	19	22	26
36	100	29	23	20	37	29	3	15	25	28
25	100	15	15	14	22	20	1	8	24	22
552	**#552**	**553**	**#553**	**554**	**#554**	**555**	**#555**	**633**	**#633**	**651**
0	0	26	22	5	10	5	17	5	5	7
0	0	33	30	3	8	5	16	8	4	5
0	0	32	30	1	4	3	10	9	2	2
#651	**714**	**#714**	**721**	**#721**	**722**	**#722**	**723**	**#723**	**724**	**#724**
0	14	10	2	2	1	2	13	9	9	0
0	16	14	2	2	1	2	14	8	7	0
0	13	13	1	1	0	1	11	4	3	0
725	**#725**	**726**	**#726**	**731**	**#731**	**732**	**#732**	**741**	**#741**	**742**
4	2	0	0	0	10	2	0	10	0	5
4	2	0	0	0	10	1	0	10	0	4
3	2	0	0	0	7	1	0	7	0	2
#742	**743**	**#743**	**744**	**#744**	**745**	**#745**	**801**	**#801**	**802**	**#802**
10	0	0	2	6	0	12	0	0	0	0
8	0	0	2	7	0	9	0	0	0	0
4	0	0	2	5	0	4	0	0	0	0
803	**#803**	**804**	**#804**	**851**	**#851**	**853**	**#853**	**921**	**#921**	**922**
0	2	6	1	0	10	0	0	1	0	10
0	1	5	1	0	7	0	0	0	0	7
0	0	3	0	0	3	0	0	0	0	3
#922	**923**	**#923**	**925**	**#925**	**926**	**#926**	**927**	**#927**	**930**	**#930**
2	4	0	0	0	1	2	0	0	27	0
2	3	0	0	0	1	1	0	0	40	0
1	1	0	0	0	1	1	0	0	44	0

Tableau C.17 – Importances Δ_δ des variables INSEE échelonnées entre 0 et 100 et calculées lors de la construction d'un unique arbre de régression **CART** pour les données HC2. L'arbre en question est présenté dans la Figure 4.20. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à un coefficient δ de pénalisation de la profondeur égal à 1, 0.9 et 0.8.

PT	PT0	PT20	PT40	PT60	PT75	158	#158	521	#521	522
27	7	6	4	16	4	100	85	61	52	41
28	7	6	4	17	4	100	84	61	52	42
29	7	7	5	17	4	100	83	61	52	42
#522	**523**	**#523**	**524**	**#524**	**525**	**#525**	**527**	**#527**	**551**	**#551**
43	48	28	25	26	29	14	17	5	42	28
43	48	28	25	26	29	14	17	5	43	28
44	49	27	25	26	29	14	17	4	46	29
552	**#552**	**553**	**#553**	**554**	**#554**	**555**	**#555**	**633**	**#633**	**651**
0	0	32	20	9	11	12	8	18	7	8
0	0	33	20	9	11	12	8	19	7	8
0	0	34	21	9	11	12	8	20	7	8
#651	**714**	**#714**	**721**	**#721**	**722**	**#722**	**723**	**#723**	**724**	**#724**
6	5	4	15	6	10	6	7	6	7	3
6	5	4	15	6	10	6	7	6	7	3
6	5	4	16	6	11	6	7	6	7	3
725	**#725**	**726**	**#726**	**731**	**#731**	**732**	**#732**	**741**	**#741**	**742**
1	1	0	0	7	4	5	2	14	5	4
1	1	0	0	7	4	5	2	14	5	4
1	1	0	0	7	4	4	2	14	5	4
#742	**743**	**#743**	**744**	**#744**	**745**	**#745**	**801**	**#801**	**802**	**#802**
4	1	1	5	5	4	3	3	0	2	0
4	1	1	5	5	4	3	3	0	2	0
4	1	1	5	5	5	3	3	0	2	0
803	**#803**	**804**	**#804**	**851**	**#851**	**853**	**#853**	**921**	**#921**	**922**
3	2	5	3	5	4	2	1	1	1	1
3	2	5	3	5	4	2	1	1	1	1
3	2	5	3	5	4	2	1	1	1	1
#922	**923**	**#923**	**925**	**#925**	**926**	**#926**	**927**	**#927**	**930**	**#930**
1	2	1	1	1	1	0	1	1	16	2
1	2	1	1	1	1	0	1	1	16	2
1	2	1	1	1	1	0	1	1	17	2

Tableau C.18 – Importances Δ_δ des variables INSEE échelonnées entre 0 et 100 et calculées lors de la procédure de **Bagging** appliquées aux données HC2. Les lignes de chacun des sous-tableaux correspondent, de haut en bas, à un coefficient δ de pénalisation de la profondeur égal à 1, 0.9 et 0.8.

Références

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control AC-19*, 716–723. System identification and time-series analysis.

Antoniadis, A. and Berruyer, J. (1986). On estimating the number of components in a finite mixture of power series distributions. *Comput. Statist. Data Anal. 4*(4), 229–241.

Antoniadis, A. and Gijbels, I. (2002). Detecting abrupt changes by wavelet methods. *J. Nonparametr. Stat. 14*(1-2), 7–29.

Antoniadis, A., Gijbels, I., and MacGibbon, B. (2000). Non-parametric estimation for the location of a change-point in an otherwise smooth hazard function under random censoring. *Scand. J. Statist. 27*(3), 501–519.

Bahadur, R. R. (1967). An optimal property of the likelihood ratio statistic. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, pp. 13–26. Berkeley, Calif.: Univ. California Press.

Bahadur, R. R. (1971). *Some limit theorems in statistics.* Philadelphia, Pa.: Society for Industrial and Applied Mathematics. Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 4.

Bahadur, R. R., Zabell, S. L., and Gupta, J. C. (1980). Large deviations, tests, and estimates. In *Asymptotic theory of statistical tests and estimation (Proc. Adv. Internat. Sympos., Univ. North Carolina, Chapel Hill, N.C., 1979)*, pp. 33–64. New York: Academic Press.

Bai, Z. D., Rao, C. R., and Wu, Y. (1999). Model selection with data-oriented penalty. *J. Statist. Plann. Inference 77*(1), 103–117.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields 113*(3), 301–413.

Basseville, M. and Nikiforov, I. V. (1993). *Detection of abrupt changes: theory and application.* Prentice Hall Inc.

Blanchard, G. (2001). *Méthodes de mélange et d'agrégation d'estimateurs en reconnaissance de formes. Application aux arbres de décision.* Ph. D. thesis, Université Paris XIII - Paris-Nord. Available at `http://www.math.u-psud.fr/~blanchard/publi/these.ps.gz`.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research 2*, 499–526.

Breiman, L. (1996a). Bagging predictors. *Machine Learning 24*(2), 123–140.

Références.

Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *Ann. Statist. 24*(6), 2350–2383.

Breiman, L. (1998). Arcing classifiers. *Ann. Statist. 26*(3), 801–849. With discussion and a rejoinder by the author.

Breiman, L. (2001). Statistical modeling: the two cultures. *Statist. Sci. 16*(3), 199–231. With comments and a rejoinder by the author.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and regression trees.* Chapman & Hall.

Brillinger (1990). Spatial-temporal modeling of spatially aggregate birth data. *Survey Methodology Journal 16*, 255–269.

Brodsky, B. E. and Darkhovsky, B. S. (1993). *Nonparametric methods in change-point problems.* Kluwer Academic Publishers Group.

Bühlmann, P. and Yu, B. (2002a). Analyzing bagging. *Ann. Statist. 30*(4), 927–961.

Bühlmann, P. and Yu, B. (2002b). Boosting with the L_2-loss: regression and classification. Preprint.

Burkholder, D. L. (1973). Distribution function inequalities for martingales. *Ann. Probability 1*, 19–42.

Carlstein, E., Müller, H.-G., and Siegmund, D. (Eds.) (1994). *Change-point problems.* Hayward, CA: Institute of Mathematical Statistics. Papers from the AMS-IMS-SIAM Summer Research Conference held at Mt. Holyoke College, South Hadley, MA, July 11–16, 1992.

Čencov, N. N. (1982). *Statistical decision rules and optimal inference.* Providence, R.I.: American Mathematical Society. Translation from the Russian edited by Lev J. Leifman.

Chambaz, A. (2002). Detecting abrupt changes in random fields. *ESAIM P&S 6*, 289–209.

Chernoff, H. (1956). Large sample theory: parametric case. *Ann. Math. Statist. 27*, 1–22.

Chou, P. A., Lookabaugh, T., and Gray, R. M. (1989). Optimal pruning with applications to tree-structured source coding and modeling. *IEEE Trans. Inform. Theory 35*(2), 299–315.

Cressie, N. A. C. (1993). *Statistics for spatial data.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons Inc.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probability 3*, 146–158.

Dacunha-Castelle, D. and Gassiat, E. (1997). The estimation of the order of a mixture model. *Bernoulli 3*(3), 279–299.

Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Ann. Statist. 27*(4), 1178–1209.

de Acosta, A. (1994). Projective systems in large deviation theory. II. Some applications. In *Probability in Banach spaces, 9 (Sandjberg, 1993)*, pp. 241–250. Birkhäuser Boston.

Dedecker, J. (2001). Exponential inequalities and functional central limit theorem for random fields. *ESAIM P&S 5.*

Dembo, A. and Zeitouni, O. (1998). *Large deviations techniques and applications.* New York: Springer-Verlag.

Donoho, D. L. (1997). CART and best-ortho-basis: a connection. *Ann. Statist. 25*(5), 1870–1911.

Doukhan, P. (1994). *Mixing.* New York: Springer-Verlag. Properties and examples.

Drucker, H. (1997). Improving regressors using boosting techniques. In *Proc. 14th International Conference on Machine Learning*, pp. 107–115. Morgan Kaufmann.

Dudley, R. M. and Philipp, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Z. Wahrsch. Verw. Gebiete 62*(4), 509–552.

Dupuis, P. and Ellis, R. S. (1997). *A weak convergence approach to the theory of large deviations.* Wiley Series in Probability and Statistics: Probability and Statistics. New York: John Wiley & Sons Inc.

Everitt, B. S. and Hand, D. J. (1981). *Finite mixture distributions.* London: Chapman & Hall. Monographs on Applied Probability and Statistics.

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* New York: John Wiley & Sons Inc.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, pp. 148–156.

Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. To appear in Ann. Inst. H. Poincaré Probab. Statist.

Gassiat, E. and Boucheron, S. (2001). Optimal error exponents in Hidden Markov Models order estimation. Preprint, submitted.

Gey, S. and Nedelec, E. (2001). Model selection for CART regression trees. Preprint.

Gey, S. and Poggi, J.-M. (2002). Boosting cart regression trees. Preprint.

Ghattas, B. (1999). Prévision par arbres de classification. *Mathématiques, Informatique et Sciences Humaines 146*, 31–50.

Ghosh, J. K. and Sen, P. K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Belmont, CA, pp. 789–806. Wadsworth.

Guyon, X. and Yao, J. (1999). On the underfitting and overfitting sets of models chosen by order selection criteria. *J. Multivariate Anal. 70*(2), 221–249.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning.* New York: Springer-Verlag. Data mining, inference, and prediction.

Haughton, D. (1989). Size of the error in the choice of a model to fit data from an exponential family. *Sankhyā Ser. A 51*(1), 45–58.

Haughton, D. and Keribin, C. (2001). Asymptotic probabilities of overestimating and underestimating the order of a model in general regular families. Submitted.

Haughton, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *Ann. Statist. 16*(1), 342–355.

Henna, J. (1985). On estimating of the number of constituents of a finite mixture of continuous distributions. *Ann. Inst. Statist. Math. 37*(2), 235–240.

Huber, P. J. (1967). 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*, pp. 221–233. Berkeley, Calif.: Univ. California Press.

James, L. F., Priebe, C. E., and Marchette, D. J. (2001). Consistent estimation of mixture complexity. *Ann. Statist. 29*(5), 1281–1296.

Kallenberg, W. C. M. and Kourouklis, S. (1992). Hodges-Lehmann optimality of tests. *Statist. Probab. Lett. 14*(1), 31–38.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A 62*(1), 49–66.

Korostelëv, A. P. and Tsybakov, A. B. (1993). *Minimax theory of image reconstruction*, Volume 82 of *Lecture Notes in Statistics*. New York: Springer-Verlag.

Kourouklis, S. (1991). Bahadur efficiency of likelihood ratio and related tests in nonregular models. *Austral. J. Statist. 33*(3), 291–298.

Lagrange, X., Godlewski, P., and Tabbane, S. (1999). *Réseaux GSM-DCS, des principes à la norme*. Hermes Sciences Publications.

Lavielle, M. On the use of penalized contrasts for solving inverse problems. Application to the DDC (Detection of Divers Changes) problem. Submitted.

Lavielle, M. (1999). Detection of multiple changes in a sequence of dependent variables. *Stochastic Process. Appl. 83*(1), 79–102.

Lavielle, M. and Lebarbier, E. (2001). An application of MCMC methods for the multiple change-points problem. *Signal Processing 81*, 39–53.

Lavielle, M. and Ludeña, C. (2000). The multiple change-points problem for the spectral distribution. *Bernoulli 6*(5), 845–869.

Lavielle, M. and Moulines, E. (2000). Least-squares estimation of an unknown number of shifts in a time series. *J. Time Ser. Anal. 21*(1), 33–59.

Ledoux, M. (1992). Sur les déviations modérées des sommes de variables aléatoires vectorielles indépendantes de même loi. *Ann. Inst. H. Poincaré Probab. Statist. 28*(2), 267–280.

Léonard, C. (2000). Minimizers of energy functionals under not very integrable constraints. Preprint.

Léonard, C. and Najim, J. (2000). An extension of Sanov's theorem. Application to the Gibbs conditioning principle. Preprint.

Leonardi, G. P. and Tamanini, I. (2002). Metric spaces of partitions. Preprint.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *Ann. Statist. 20*(3), 1350–1360.

Lindsay, B. G. (1989). Moment matrices: applications in mixtures. *Ann. Statist. 17*(2), 722–740.

Lindsay, B. G. and Lesperance, M. L. (1995). A review of semiparametric mixture models. *J. Statist. Plann. Inference 47*(1-2), 29–39. Statistical modelling (Leuven, 1993).

Lugosi, G. (2000). Lectures on statistical learning theory. Presented at the Garchy Seminar on Mathematical Statistics and Applications, available at `http://www.econ.upf.es/~lugosi`.

Mallows (1973). Some comments on C_p. *Technometrics 15*, 661–675.

Mammen, E. and Tsybakov, A. B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist. 23*(2), 502–524.

Massart, P. (2000). Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math. (6) 9*(2), 245–303.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models.* New York: Marcel Dekker Inc. Inference and applications to clustering.

Móricz, F. (1983). A general moment inequality for the maximum of the rectangular partial sums of multiple series. *Acta Math. Hungar. 41*(3-4), 337–346.

Móricz, F. A., Serfling, R. J., and Stout, W. F. (1982). Moment and probability bounds with quasisuperadditive structure for the maximum partial sum. *Ann. Probab. 10*(4), 1032–1040.

Müller, H.-G., Stadtmüller, U., and Tabnak, F. (1997). Spatial smoothing of geographically aggregated data, with application to the construction of incidence maps. *J. Amer. Statist. Assoc. 92*(437), 61–71.

Petrov, V. V. (1975). *Sums of independent random variables.* New York: Springer-Verlag.

Petrov, V. V. (1995). *Limit theorems of probability theory.* New York: The Clarendon Press Oxford University Press. Sequences of independent random variables, Oxford Science Publications.

Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory 1*, 295–314.

Rao, M. M. and Ren, Z. D. (1991). *Theory of Orlicz spaces.* New York: Marcel Dekker Inc.

Rio, E. (2000). *Théorie asymptotique des processus aléatoires faiblement dépendants.* Springer.

Rissanen, J. (1978). Modelling by shortest data description. *Automatica 14*, 465–471.

Rockafellar, R. T. (1970). *Convex analysis.* Princeton University Press.

Schied, A. (1998). Cramer's condition and Sanov's theorem. *Statist. Probab. Lett. 39*, 55–60.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist. 6*(2), 461–464.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc. 82*(398), 605–610.

Senoussi, R. (1990). Statistique asymptotique presque sûre de modèles statistiques convexes. *Ann. Inst. H. Poincaré Probab. Statist. 26*(1), 19–44.

Serfling, R. J. (1968). Contributions to central limit theory for dependent variables. *Ann. Math. Statist. 39*, 1158–1175.

Talagrand, M. (1996a). New concentration inequalities in product spaces. *Invent. Math. 126*(3), 505–563.

Talagrand, M. (1996b). A new look at independence. *Ann. Probab. 24*(1), 1–34.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* Chichester: John Wiley & Sons Ltd.

van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes.* New York: Springer-Verlag. With applications to statistics.

Vapnik, V. N. (1998). *Statistical learning theory.* New York: John Wiley & Sons Inc.

Wu, L. (1994). Large deviations, moderate deviations and LIL for empirical processes. *Ann. Probab. 22*(1), 17–27.

Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz's criterion. *Statist. Probab. Lett. 6*(3), 181–189.

Ziemer, W. P. (1989). *Weakly differentiable functions.* New York: Springer-Verlag. Sobolev spaces and functions of bounded variation.