Estimation de densité en dimension élevée et classification de courbes

Laurent Rouviere

Université Montpellier II

–SCIENCES ET TECHNIQUES DU LANGUEDOC–

T H È S E

pour obtenir le grade de

Docteur de l'Université Montpellier II

Discipline	:	Mathématiques appliquées
Ecole Doctorale	:	**I**nformation, **S**tructures, **S**ystèmes
Formation Doctorale	:	Biostatistique

Estimation de Densité en Dimension Élevée et Classification de Courbes

par

Laurent ROUVIÈRE

Présentée et soutenue publiquement le **18 novembre 2005** devant le jury composé de :

MM.	A. BERLINET	Professeur, Université Montpellier II	Directeur de Thèse
	P. BESSE	Professeur, Université Toulouse III	Rapporteur
	G. BIAU	Professeur, Université Montpellier II	Directeur de Thèse
	D. BOSQ	Professeur, Université Paris VI	Président
	L. CAVALIER	Professeur, Université Aix-Marseille I	Examinateur
	B. RÉMILLARD	Professeur, HEC Montréal	Rapporteur
	M. WEGKAMP	Professeur, Florida State University	Examinateur

Remerciements

Ce fut pour moi un plaisir de travailler sous la direction d'Alain Berlinet et de Gérard Biau. Alain Berlinet m'a initié à la recherche au cours de mon année de DEA. Un an plus tard, Alain Berlinet proposait à Gérard Biau de participer à l'encadrement de mes travaux de recherche. Ce dernier accepta sans hésitation et fit preuve d'une grande disponibilité même lorsqu'il était en poste à Paris. Alain Berlinet et Gérard Biau m'ont appris à la fois la rigueur scientifique et l'ouverture d'esprit. Aujourd'hui encore, je reste impressionné par l'étendue de leurs connaissances, tant théoriques qu'appliquées. Leurs conseils et leur soutien ont été particulièrement précieux dans l'accomplissement de mon travail. Je leur adresse de chaleureux remerciements pour tout ce qu'ils m'ont appris et j'espère que notre collaboration ne s'arrêtera pas avec cette thèse.

Merci infiniment à Philippe Besse et Bruno Rémillard qui ont accepté, malgré un emploi du temps que je devine accablant, la tâche ingrate de rapporteur. Merci également à Denis Bosq qui a accepté d'être Président du jury ainsi qu'à Laurent Cavalier et Marten Wegkamp qui ont donné spontanément leur accord pour faire partie du jury.

J'ai une pensée reconnaissante pour les personnes avec qui j'ai eu le plaisir de collaborer dans le cadre de mon monitorat. Je pense particulièrement à Gilles Ducharme, Djamal Echikr, Marc Johannides, Irène Larramendy et Saïd Mounime qui m'ont accordé leur confiance et auprès de qui j'ai beaucoup appris sur le métier d'enseignant.

Je voudrais également remercier tous les membres de l'Equipe de Probabilités et Statistique de l'Université Montpellier II. J'ai aussi une pensée particulière pour tous les doctorants et jeunes docteurs qui ont rendu le quotidien de cette thèse plus agréable. Je pense notamment à Marie-José, Sandie et Vivien pour nos "nombreuses" conversations, mais aussi à Abdou, Ahmad, Azaël, Ghislain et bien entendu à Thomas.

J'adresse mes plus sincères remerciements à Gilles Caraux, Pierre Cartigny et Jean-Pierre Vila qui ont bien voulu m'accueillir au sein de l'Unité de Biométrie sur le campus ENSAM-INRA de Montpellier. J'ai eu la chance de travailler dans des conditions privilégiées et de tisser des liens d'amitié que je crois solides. Plusieurs noms me viennent à l'esprit : Alain, Anne, Brigitte, Cécile, Christophe,

Frédéric, Gilberte, Isabelle, Luc, Martine, Nadine, Nicolas, Pascal, Patrice, Philippe, Véronique, Vincent. Merci à tous.

Par ailleurs, sur un plan plus personnel, je tiens à avoir une pensée pour mes proches, amis ou membres de la famille. Leur soutien, leurs encouragements et leur bonne humeur me sont allés droit au coeur et m'ont permis de sortir du contexte de cette thèse lorsque cela était nécessaire. Je terminerai en remerciant mes parents, Annie et Alain, pour leur gentillesse, leur compréhension et pour m'avoir accordé une liberté totale tout au long de mes études.

Table des matières

Introduction

1.1 Présentation de la thèse

L'*apprentissage statistique* désigne un vaste ensemble de méthodes et d'algorithmes permettant, dans un sens général, d'extraire l'information pertinente de données ou d'apprendre des comportements à partir d'exemples. Les applications de ce paradigme sont très nombreuses, allant de la recherche d'informations dans de grands ensembles de données (fouille de textes ou d'images) à la biologie (reconstruction des réseaux génétiques, puces ADN, etc). L'apprentissage statistique est aujourd'hui confronté à des données dont la nature est de plus en plus *complexe* (courbes, images, etc) et qui prennent des valeurs dans des espaces dont la dimension est toujours plus élevée. Les méthodes statistiques traditionnelles doivent alors être adaptées pour permettre de traiter ces données d'un genre nouveau. Dans cet esprit, l'objectif de la thèse consiste à étudier et approfondir des techniques d'*estimation de la densité* et de *classification* dans des espaces de dimension élevée. Nous avons choisi de structurer notre travail en trois parties.

La première partie, intitulée **compléments sur les histogrammes modifiés**, est composée de deux chapitres consacrés à l'étude d'une famille d'estimateurs non paramétriques de la densité, les *histogrammes modifiés*, connus pour posséder de bonnes propriétés de convergence au sens des critères de la théorie de l'information. Cette partie s'inscrit dans la continuité des travaux de thèse de Brunel [9] et Biau [7]. Dans le premier chapitre, les histogrammes modifiés sont envisagés comme des systèmes dynamiques à espace d'états de dimension infinie. Le second chapitre est consacré à l'étude de ces estimateurs pour des dimensions supérieures à un.

La deuxième partie de la thèse, intitulée **méthodes combinatoires en estimation de la densité**, se divise en deux chapitres. Nous nous intéressons dans cette partie aux performances *à distance finie* d'estimateurs de la densité sélectionnés à l'intérieur d'une famille d'estimateurs candidats dont le cardinal n'est pas nécessairement fini. Dans le premier chapitre, nous étudions les performances

de ces méthodes dans le cadre de la sélection des différents paramètres des histogrammes modifiés. Nous poursuivons, dans le second chapitre, par la sélection d'estimateurs à noyau dont le paramètre de lissage s'adapte localement au point d'estimation et aux données.

La troisième et dernière partie, plus appliquée et indépendante des précédentes, présente une méthode permettant de classer des observations fonctionnelles à partir d'une décomposition des données dans des bases d'ondelettes.

1.2 Compléments sur les histogrammes modifiés

La reconstruction d'une courbe ou d'une surface à partir d'observations qui n'apportent qu'une information partielle est un thème de la statistique moderne sous-jacent à de nombreux problèmes pratiques. Dans cette première partie, nous étudions le problème de l'estimation d'une densité de probabilité. Plaçons-nous sur l'espace mesurable $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ et rappelons qu'une variable aléatoire X à valeurs dans $\mathbb{R}^d$ admet f comme densité si, pour tout borélien A de $\mathbb{R}^d$, on a $\int_A f(x)\,\mathrm{d}x = \mathbf{P}(X \in A)$. Ainsi, une densité de probabilité permet de calculer les probabilités d'ensembles pour des variables aléatoires continues. La représentation graphique de la fonction f permet également de visualiser la distribution de la variable aléatoire X.

Cependant, dans de nombreuses applications, la densité f est inconnue et on ne dispose que d'un n-échantillon i.i.d. $X_1, \ldots, X_n$ issu d'une variable aléatoire X admettant f comme densité. Le problème du statisticien consiste alors à utiliser cet échantillon pour construire un estimateur f_n qui soit le plus "proche" possible de la densité f. Cette notion de proximité n'est pas absolue et requiert la spécification d'un critère d'erreur. Deux types de critères peuvent alors être envisagés : mesure d'erreur globale ou mesure d'erreur ponctuelle. Dans ce travail de thèse, nous nous placerons dans le cadre d'applications nécessitant l'estimation de toute la courbe représentative de f (plutôt qu'une valeur en un point particulier) et nous ne considérerons donc que des critères globaux.

Parmi les mesures d'erreurs globales les plus utilisées, citons les normes L_p

$$||f - f_n||_p = \begin{cases} \left(\int |f - f_n|^p \right)^{1/p} & \text{si } 1 \leq p < \infty \\ \\ \sup |f - g| & \text{si } p = \infty\,. \end{cases}$$

En particulier, la décomposition classique du carré de l'erreur L_2 moyenne en un terme de biais et de variance lui confère une interprétabilité statistique et permet

des calculs "relativement" aisés. L'erreur L_1 est également un des critères privilégiés des statisticiens, et ceci pour au moins quatre raisons. D'abord, ce critère est calculable sans aucune hypothèse supplémentaire sur les densités (alors que les densités doivent être de carré intégrable pour l'erreur L_2). Ensuite, le critère L_1 possède une signification claire en termes de probabilités grâce à l'égalité dite de Scheffé [19]

$$\int |f - f_n| = 2 \sup_{B \in \mathcal{B}(\mathbb{R}^d)} \left| \int_B f - \int_B f_n \right|, \tag{1.1}$$

le membre de droite étant égal à deux fois la *distance en variation totale* entre les mesures associées à f et f_n. On déduit donc de l'égalité (1.1) que la distance en variation totale tend vers 0 selon un mode stochastique lorsque la distance L_1 entre f et f_n tend vers 0 selon le même mode. En outre, si nous savons par exemple que $\int |f - f_n| < 0.06$, alors les différences entre les probabilités d'ensembles seront au plus de 0.03. Le théorème de Scheffé entraîne également que la distance L_1 reste invariante par transformation bijective de l'espace $\mathbb{R}^d$. Cette propriété peut être simplement exploitée en univarié, lorsque l'on désire visualiser sur l'écran l'erreur L_1 commise dans la queue de distribution entre une densité à support infini et son estimateur. Il suffit alors de transformer la partie de la queue de distribution intéressante d'une façon continue et monotone en un intervalle borné. Alors que les formes des densités changent, la distance L_1 reste invariante. Enfin, l'erreur L_1 commise entre f et f_n peut être aisément visualisée : elle correspond à l'aire comprise entre les courbes représentatives des deux fonctions. Pour toutes ces raisons, la recherche d'estimateurs possédant de bonnes propriétés pour le critère L_1 est un des thèmes privilégiés des statisticiens. Nous utiliserons principalement ce critère dans ce travail de thèse.

Cela étant, dans certains domaines de la statistique tels que la compression de données, les réseaux de télécommunications, les problèmes de classification ou encore les réseaux de neurones (voir Berlinet, Vajda et van der Meulen [6]), la convergence définie par la norme L_1 peut se révéler insuffisante. On lui préfère alors la convergence définie par des *pseudo-distances* comme par exemple l'*entropie relative* (ou *information de Kullback-Leibler* ou I-*divergence*). Etant donné deux densités f et f_n, l'*entropie relative* de f par rapport à f_n, notée $D(f, f_n)$, est définie par

$$D(f, f_n) = \begin{cases} \displaystyle\int f \log \frac{f}{f_n} & \text{si } f \ll f_n \\ +\infty & \text{sinon.} \end{cases}$$

L'inégalité de *Pinsker* (voir par exemple Kullback [14])

$$||f - f_n||_1^2 \leq 2D(f, f_n)$$

implique que l'entropie relative induit une topologie plus forte sur l'espace des densités de probabilité que la topologie associée à la norme L_1. Cette remarque motive la mise au point d'un estimateur non paramétrique de la densité f convergeant en entropie relative (*resp.* en entropie relative moyenne), c'est-à-dire tel que $D(f, f_n)$ tende vers 0 presque sûrement (*resp.* $\mathbf{E}D(f, f_n)$ tende vers 0). Remarquons que ceci peut soulever quelques difficultés puisque, par exemple, dans le cas de l'estimateur histogramme usuel $\hat{f}_n$, la quantité $D(f, \hat{f}_n)$ peut être infinie avec une probabilité non nulle !

Barron, Györfi et van der Meulen [4] ont construit une famille d'estimateurs convergeant en entropie relative et en entropie relative moyenne. La seule condition requise sur la densité f est l'existence d'une densité g telle que $D(f, g) < \infty$. Soulignons au passage que cette dernière condition implique l'absolue continuité de f par rapport à g. Parmi les estimateurs présentés par Barron, Györfi et van der Meulen [4], l'*histogramme modifié* (initialement proposé par Barron [3]) est défini à partir de n observations $X_1, \ldots, X_n$ indépendantes et distribuées suivant la loi de f de la manière suivante :

- Soit g une densité connue (dite *densité de référence*) associée à la loi de probabilité ν_g (dite *mesure de référence*) ;
- Soit ℓ un entier tel que $1 \leq \ell$ et soit $h = 1/\ell$;
- Considérons une partition de $\mathbb{R}^d$, $P = \{A_1, \ldots, A_\ell\}$ telle que $\nu_g(A_i) = h$, $i = 1, \ldots, \ell$;
- Alors, en notant $a_n = 1/(nh+1)$, on définit l'histogramme modifié f_n par :

$$f_n(x) = \left[(1 - a_n)\frac{\mu_n\big(A(x)\big)}{h} + a_n\right] g(x) = \frac{n\mu_n\big(A(x)\big) + 1}{nh + 1}\, g(x)\,, \qquad (1.2)$$

  où μ_n désigne la mesure empirique associée à l'échantillon $X_1, \ldots, X_n$ et $A(x) = A_i$ si $x \in A_i$.

La première expression de f_n nous présente cet estimateur comme un mélange entre un estimateur de type histogramme et la densité de référence, d'où le nom d'histogramme modifié. La seconde écriture nous montre que cet estimateur est en fait construit comme une déformation par morceaux de la densité de référence g. Sur chaque cellule A_i de la partition, le coefficient multiplicateur de $g(x)$ vaut

$$\frac{n\mu_n\big(A_i\big) + 1}{nh + 1} = \frac{n\mu_n\big(A_i\big) + 1}{n\nu_g(A_i) + 1}\,.$$

Ainsi, sur les classes où la mesure empirique est supérieure à la mesure de référence, on corrige la densité g en la "déformant vers le haut" ; à l'inverse, lorsque la mesure empirique est inférieure à la mesure de référence, on déforme g vers le bas (voir Figure 1.1).

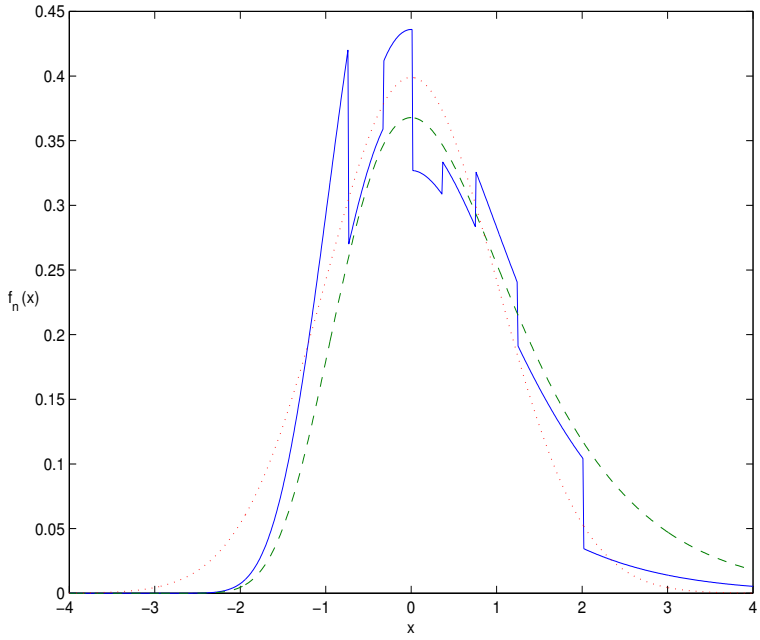

FIG. 1.1 – Histogramme modifié (trait continu) d'une densité gaussienne $\mathcal{N}(0,1)$ (pointillés), la densité de référence est une densité de Gumbel (tirets), $n = 100$, $\ell = 8$.

Lorsque le nombre de classes ℓ et l'échantillon sont fixés, cet estimateur peut être vu comme une fonctionnelle admettant en entrée la densité de référence g et fournissant en sortie une nouvelle densité de probabilité (estimateur de f) que l'on peut noter $B_\ell g$. Puisque $B_\ell g$ est une densité de probabilité, nous pouvons alors la considérer comme densité de référence de l'histogramme modifié : cela conduit à un nouvel estimateur de f, noté $B_\ell^2 g$, qui peut à son tour devenir densité de référence de l'histogramme modifié et ainsi de suite... Ce processus itératif amène naturellement à considérer le *système dynamique* B_ℓ à espace d'états de dimension infinie, défini de la façon suivante :

$$\begin{aligned} B_\ell \ : \ & \mathcal{D} \to \mathcal{D} \\ & g \mapsto B_\ell g \,, \end{aligned} \tag{1.3}$$

où $\mathcal{D}$ représente l'ensemble des densités de probabilité sur $\mathbb{R}^d$. Berlinet et Biau [5] ont étudié ce système dynamique en dimension un. Ces auteurs, sous certaines hypothèses, ont montré la stationnarité des trajectoires de densités $\{B_\ell^p g\}_{p\geq 0}$ après un nombre fini d'itérations. Dans ce cas, la famille d'estimateurs $\{B_\ell^p g : p \geq 0\}$ est de cardinal fini. Dans le **premier chapitre** de cette première partie, nous étudierons différentes méthodes permettant de sélectionner un estimateur à

l'intérieur de cette famille. Ces méthodes seront ensuite appliquées pour tenter d'améliorer les performances d'estimateurs à noyau de la densité.

La construction de l'histogramme modifié repose sur une partition dont les éléments sont de même mesure de référence. Dans le cas réel, le choix d'une partition basée sur les quantiles de la densité de référence s'impose de façon naturelle. Plus précisément, étant donné un entier positif ℓ (nombre de classes), la partition $P = \{A_1, \ldots, A_\ell\}$ est définie par

$$\left\{A_1, A_2, \ldots, A_\ell\right\} = \left\{]-\infty, q_1],]q_1, q_2], \ldots,]q_{\ell-1}, \infty[\right\},$$

où, pour $i = 1, \ldots, \ell - 1$, q_i désigne le quantile d'ordre ih de la densité g. Cependant, le concept de quantile univarié étant fortement lié à la relation d'ordre de $\mathbb{R}$, le passage au domaine multivarié n'est pas immédiat. Par conséquent, pour des dimensions plus grandes que un, il faut s'attacher à mettre en évidence une partition dont les classes soient de mesure de référence identique. Dans le **second chapitre** de cette partie, nous présentons différentes méthodes permettant de construire de telles partitions. Nous étudions la convergence en entropie relative et en entropie relative moyenne des histogrammes modifiés associés à ces partitions.

1.3 Méthodes combinatoires en estimation de la densité

Les estimateurs non paramétriques de la densité dépendent en général d'un ou plusieurs paramètres dont le choix se révèle crucial, aussi bien pour la précision locale que pour la précision globale de l'estimateur. Par exemple dans le cas de l'estimateur à noyau $f_{n,h}$ (le noyau K est fixé et h désigne la fenêtre), il est facile de voir que la loi de densité $f_{n,h}$ converge vers la mesure empirique lorsque h tends vers 0 alors que $f_{n,h}$ tend uniformément vers la fonction nulle lorsque h tend vers l'infini...

Les paramètres inconnus des estimateurs non paramétriques de la densité sont souvent sélectionnés en minimisant les termes dominants dans le développement asymptotique du critère d'erreur considéré. Bien que ces approches soient performantes dans de nombreuses situations, elles présentent au moins deux inconvénients. D'abord, le calcul de ces développements asymptotiques nécessite en général des hypothèses sur la régularité de la densité à estimer. Ceci peut être frustrant dans la mesure où ces hypothèses sont difficiles à vérifier dans la réalité ! Ensuite, ces procédures sont de nature asymptotique, ce qui signifie qu'elles ne sont, théoriquement, valables que lorsque n est infini...

1.3 Méthodes combinatoires en estimation de la densité

Dans un ouvrage récent, Devroye et Lugosi [13] explorent de nouvelle méthodes permettant de sélectionner automatiquement les paramètres d'estimateurs de la densité. Plus précisément, étant donné $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$ une famille de densités candidates paramétrée par Θ, les méthodes développées par Devroye et Lugosi permettent de sélectionner automatiquement un estimateur f_n dans $\mathcal{F}$ tel que :

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \leq 3 \inf_{\theta \in \Theta} \mathbf{E}\left\{ \int |f_\theta - f| \right\} + D_n \,. \tag{1.4}$$

Dit autrement, l'erreur L_1 moyenne de l'estimateur sélectionné ne dépasse pas trois fois la plus petite erreur possible, plus un terme résiduel D_n qu'il va falloir s'attacher à contrôler. Ce terme D_n dépend de la richesse combinatoire de la classe $\mathcal{F}$ et peut être contrôlé via un détour par la théorie de Vapnik et Chervonenkis [21] sur la convergence uniforme de la mesure empirique.

On peut voir le premier terme du membre de droite de l'inégalité (1.4) comme une *erreur d'approximation* : l'erreur commise par le "meilleur" estimateur présent dans la classe. Le second terme peut être associé à une *erreur d'estimation* : l'erreur commise par l'estimateur sélectionné par rapport au meilleur estimateur présent dans la classe (voir Figure 1.2). Ces deux termes varient en général en sens inverse : plus la famille $\mathcal{F}$ est riche et plus le terme d'approximation risque de diminuer. En contrepartie, il sera plus difficile de trouver le "meilleur" estimateur dans $\mathcal{F}$, et vice-versa.

Lorsque les densités candidates dépendent des données (c'est par exemple le cas lorsque l'on cherche à sélectionner la fenêtre de l'estimateur à noyau ou le pas de l'histogramme), la famille de densités candidates s'écrit $\mathcal{F} = \{f_{n,\theta}, \theta \in \Theta\}$ et l'inégalité (1.4) devient

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \leq 3\big(1 + C_n\big) \inf_{\theta \in \Theta} \mathbf{E}\left\{ \int |f_{n,\theta} - f| \right\} + D_n$$

où le terme C_n est en général "petit", typiquement de l'ordre de $1/\sqrt{n}$.

L'algorithme de sélection d'un tel estimateur f_n est basé sur la minimisation d'un critère empirique du genre

$$\sup_{A \in \mathcal{A}} \Big| \int_A f_\theta - \mu_n(A) \Big| \tag{1.5}$$

où μ_n désigne la mesure empirique associée à l'échantillon $X_1, \ldots, X_n$ et $\mathcal{A}$ représente une classe d'ensembles judicieusement choisie. Remarquons d'emblée que si $\mathcal{A}$ désigne la tribu borélienne de $\mathbb{R}^d$, alors la quantité (1.5) à optimiser est

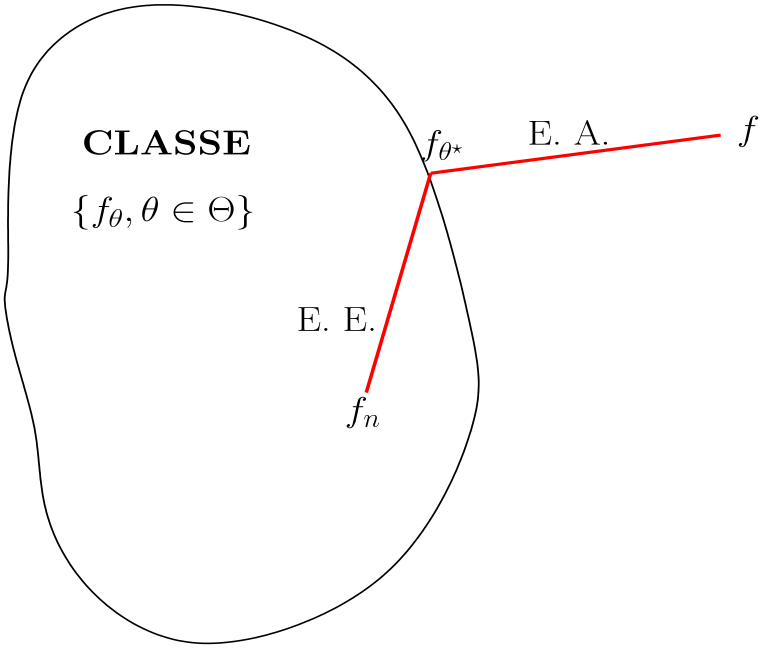

FIG. 1.2 – Schéma illustrant l'inégalité (1.4), E. E. représente l'erreur d'estimation, E. A. celle d'approximation, $\theta^\star$ est le paramètre associé au "meilleur" estimateur dans classe.

constamment égale à 1. A l'inverse, si la classe d'ensemble $\mathcal{A}$ est trop petite, alors la proximité entre $\int_A f_\theta$ et $\int_A f$ ne va pas forcément impliquer que f_θ soit "proche" de f.

Devroye et Lugosi [13], en s'inspirant des travaux de Yatracos [22], ont montré dans leur ouvrage qu'il suffit de considérer la classe d'ensembles (appelée *classe de Yatracos*)

$$\mathcal{A} = \left\{ \{x : f_\theta(x) > f_{\theta'}(x)\} : (\theta, \theta') \in \Theta^2 \right\}.$$

Dans la majorité des cas, bien que complexe d'un point de vue combinatoire, cette classe reste assez facile à manipuler. Si on considère par exemple le cas où $\mathcal{F}$ représente l'ensemble de toutes les densités gaussiennes univariées,

$$\mathcal{F} = \left\{ f_{m,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} : m \in \mathbb{R}, \sigma > 0 \right\},$$

il est facile de voir que la classe de Yatracos associée à $\mathcal{F}$ est composée d'intervalles fermés et d'unions de deux demi-intervalles (voir Figure 1.3).

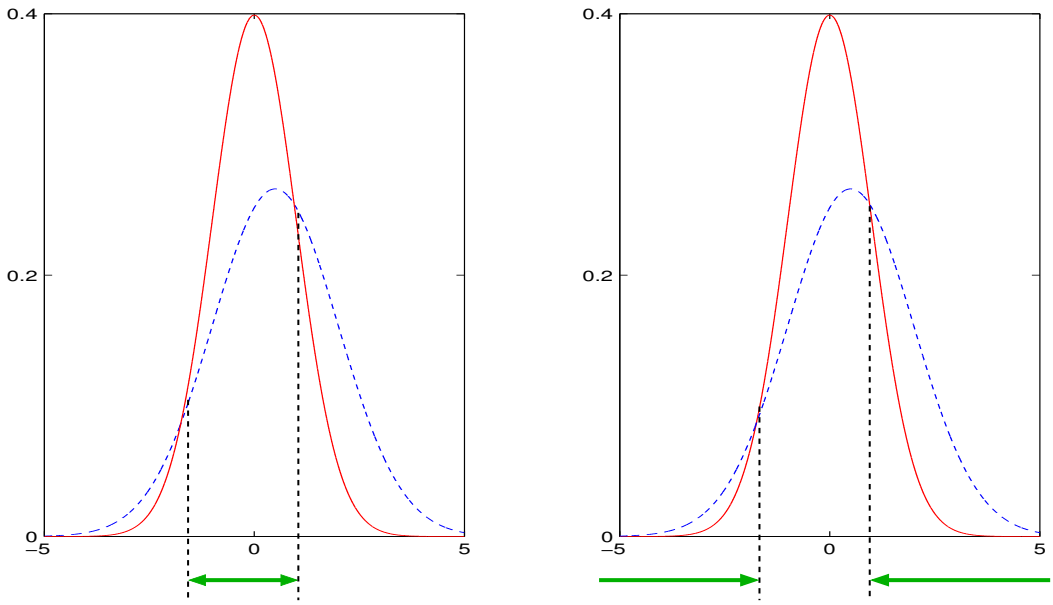

FIG. 1.3 – Deux ensembles de la classe de Yatracos (flèches) associés aux densités f_1 (lois $\mathcal{N}(0,1)$, trait plein) et f_2 (loi $\mathcal{N}(0.5,1.5)$, tirets) : $\{x : f_1(x) > f_2(x)\}$ (gauche), $\{x : f_2(x) > f_1(x)\}$ (droite).

Dans le **premier chapitre** de cette deuxième partie, nous étudions cette méthode dans le cadre de la sélection de la partition P et de la densité de référence g des histogrammes modifiés. L'originalité du travail réside dans le fait qu'à notre connaissance, aucune méthode permettant de sélectionner le paramètre fonctionnel g n'a encore été proposée.

Parmi les nombreux estimateurs non paramétriques de la densité, l'estimateur à noyau (Parzen [15], Rosenblatt [16]) est probablement le plus utilisé. Rappelons que cet estimateur est défini par :

$$f_{n,h}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big), \quad x \in \mathbb{R}^d\,, \tag{1.6}$$

où le *noyau* $K : \mathbb{R}^d \to \mathbb{R}$ vérifie $\int K = 1$ et $h > 0$ est la *fenêtre* (ou *paramètre de lissage*). Cet estimateur est flexible, dans la mesure où il laisse à l'utilisateur une grande liberté non seulement dans le choix du noyau K, mais encore dans le choix du réel h. Lorsqu'on se limite aux noyaux K positifs, les vitesses de convergence varient peu en fonction de K et les critères essentiels du choix du noyau sont alors la simplicité et la vitesse de calcul d'une part, la régularité de la courbe à obtenir

d'autre part. En revanche, le choix du paramètre h se révèle crucial, aussi bien pour la précision locale que globale de l'estimateur.

L'estimateur (1.6) est connu pour posséder de bonnes propriétés lorsque la densité f est suffisamment régulière. Cependant, ses performances s'amoindrissent en présence de cibles plus complexes (comme par exemple des densités multimodales, voir Sain et Scott [18]). En outre, la dimension de l'espace des observations peut passablement affecter la qualité de l'estimation (Sain [17]). Cet ensemble de défauts peut être corrigé, dans une certaine mesure, en permettant au paramètre h de s'adapter localement au point d'estimation et aux données. On obtient alors une nouvelle famille d'estimateurs, appelés *estimateurs à noyau variables*, et définis de manière générale par

$$f_{n,h}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(x,X_i)} K\left(\frac{x-X_i}{h(x,X_i)}\right), \quad x \in \mathbb{R}^d\,, \tag{1.7}$$

où $h : \mathbb{R}^d \times \mathbb{R}^d \to (0,\infty)$ est une fonction mesurable. De nombreux auteurs se sont intéressés aux propriétés de tels estimateurs. Abramson [2] a montré que pour un bon choix de $h(x,X_i)$, le terme de biais de l'erreur L_2 peut atteindre des vitesses de convergence habituellement réservées à des estimateurs à noyau (fixes) utilisant des noyaux négatifs d'ordre 4. Terrell et Scott [20] ont également montré que de tels estimateurs permettent d'améliorer de manière significative les performances des estimateurs à noyau (fixes) en grande dimension. Dans le cas univarié, pour certaines familles de densités, Devroye et Lugosi [12] ont obtenu, pour des estimateurs du type (1.7), des vitesses de convergence pour le critère L_1 plus rapides que les vitesses habituelles des estimateurs à noyau. Considérons par exemple le *modèle-jouet* $f(x) = 2x$ sur $[0,1]$. En s'appuyant sur les résultats de Devroye et Lugosi [12] (Lemme 1), on peut montrer que pour le choix

$$h_0(x) = \frac{1}{2x}\mathbf{1}_{\{x\le 1/2\}} + 2(1-x)\mathbf{1}_{\{x>1/2\}}\,,$$

on a

$$\mathbf{E}\left\{\int |f_{n,h_0(x)} - f|\right\} \le \sqrt{\frac{8}{n}}\,. \tag{1.8}$$

Sur la Figure 1.4, nous comparons les performances de l'estimateur $f_{n,h_0(x)}$ avec celles du meilleur estimateur à noyau fixe $f_{n,h^\star}$, c'est-à-dire

$$h^\star \in \operatorname*{argmin}_{h>0} \int |f_{n,h} - f|\,.$$

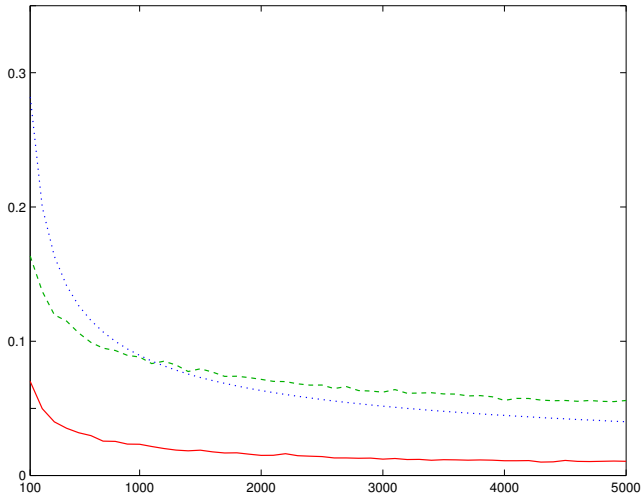

FIG. 1.4 – Erreurs L_1 commises par l'estimateur $f_{n,h_0(x)}$ (trait plein) et par l'estimateur $f_{n,h^\star}$ (tirets) pour les tailles d'échantillon $n = 100, \ldots, 5000$. En pointillés, la borne $\sqrt{8/n}$ de l'équation (1.8). Les résultats sont moyennés sur 150 répétitions.

Nous voyons clairement que les erreurs L_1 commises par l'estimateur à noyau variable sont inférieures à celles de l'estimateur à noyau fixe. Par exemple, pour des tailles d'échantillon proches de 1000, faire varier la fenêtre permet de réduire l'erreur L_1 de 75% environ.

Dans le **second chapitre** de cette partie, nous nous intéressons à la sélection d'un estimateur à noyau à fenêtre variable. Plus précisément, à partir des méthodes combinatoires décrites plus haut, nous montrerons comment sélectionner une fenêtre variable de manière à ce que l'erreur L_1 moyenne commise par l'estimateur sélectionné ne dépasse pas la plus petite erreur possible, à un facteur constant près, plus un terme résiduel qui tend vers zéro sous de faibles conditions.

1.4 Classification de courbes

Dans de nombreux domaines de la statistique contemporaine, les individus prennent la forme de *courbes (ou surfaces) aléatoires*. Ces courbes peuvent, par exemple, représenter la température en un point du globe, le cours d'une action en bourse, le tracé d'un électrocardiogramme ou encore la consommation d'électricité d'une grande ville... Les appareils de mesure ne captant que la valeur de la courbe à certains instants, les données disponibles ne sont en fait que des ver-

sions discrétisées de la courbe. Toutefois, le statisticien a intérêt à prendre en compte le caractère *continu* de ces observations. Il devient dès lors pertinent de ne plus considérer ces individus comme des vecteurs de grande dimension, mais plutôt de les appréhender comme des "fonctions", c'est-à-dire comme des objets uniques évoluant dans des espaces de dimension infinie. Il faut alors adapter les méthodes statistiques classiques à ces nouveaux individus. On regroupe sous le vocable général de *statistique fonctionnelle* l'ensemble des techniques statistiques permettant de traiter efficacement ces données d'un genre nouveau.

Dans le contexte de la *classification fonctionnelle supervisée*, un *label* (c'est-à-dire une variable qualitative admettant plusieurs modalités) est associé à chaque individu, et on cherche à prédire le label inconnu d'une nouvelle observation. Pour ce faire, nous disposons d'un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ de variables aléatoires indépendantes et identiquement distribuées, où X_i est une fonction aléatoire à valeurs dans un espace fonctionnel $\mathcal{F}$, et Y_i est une variable aléatoire à valeurs dans $\{0, 1\}$. L'objectif est de construire, à l'aide de cet échantillon d'apprentissage, une *règle de classification*

$$g_n : \mathcal{F} \to \{0, 1\}$$

qui, à une nouvelle fonction X indépendante des X_i, associe le label 0 ou 1. On parle alors, avec un léger abus de langage, de *classification de courbes*.

Une telle procédure de classification trouve de nombreuses applications dans l'industrie, où les courbes peuvent, par exemple, représenter un même phénomène au cours du temps : évolution de la température, du pH ou encore de l'intensité d'un signal sonore.

La performance d'une règle g_n est mesurée par sa *probabilité d'erreur* :

$$L(g_n) = \mathbf{P}\big(g_n(X) \neq Y | (X_1, Y_1), \ldots, (X_n, Y_n)\big) .$$

Théoriquement, le problème est résolu puisque la *règle de Bayes*

$$g^\star(x) = \begin{cases} 0 & \text{si } \mathbf{P}(Y = 0 | X = x) \geq \mathbf{P}(Y = 1 | X = x) \\ 1 & \text{sinon} \end{cases}$$

est optimale, au sens où :

$$L^\star = \mathbf{P}\big(g^\star(X) \neq Y\big) = \inf_{g : \mathcal{F} \to \{0,1\}} \mathbf{P}\big(g(X) \neq Y\big)$$

(voir Devroye, Györfi et Lugosi [11], page 10). Bien entendu, la règle de Bayes dépend de la distribution (inconnue) de (X, Y) et ne peut donc être calculée à

partir des données. On cherchera donc à construire une règle g_n dont la probabilité d'erreur $L(g_n)$ soit aussi proche que possible de la probabilité d'erreur optimale $L^\star$. De manière plus précise, on dira que la règle g_n est convergente lorsque

$$\lim_{n\to\infty} \mathbf{E}L(g_n) = L^\star \,. \tag{1.9}$$

A ce jour, de nombreuses méthodes de classification ont été étudiées, testées et comparées pour des observations évoluant en dimension finie (voir Devroye, Györfi et Lugosi [11]). Bien que ces techniques puissent, sous certaines hypothèses, s'étendre au domaine fonctionnel, elles se retrouvent en général confrontées au fléau de la dimension (voir par exemple Abraham, Biau et Cadre [1] pour une discussion détaillée concernant la règle du noyau en dimension infinie). Afin de pallier cette difficulté, il est crucial d'effectuer une étape préliminaire de réduction de la dimension ou de sélection de modèle.

En estimation fonctionnelle, les méthodes de *projection* sont souvent employées pour réduire la dimension ou pour débruiter un signal. Ces techniques consistent à réduire la dimension infinie des observations en ne considérant que certains coefficients des données décomposées dans une base appropriée. Dans un travail récent, Biau, Bunea et Wegkamp [8] se sont appuyés sur de telles méthodes pour classer des courbes. L'approche proposée par ces auteurs consiste à appliquer la règle des plus proches voisins sur les d premiers coefficients de Fourier des observations fonctionnelles.

Il est bien connu que la transformée de Fourier permet une bonne analyse *fréquentielle* d'un signal. La transformée en ondelettes (voir par exemple Cohen [10]) présente, en outre, l'avantage d'effectuer une analyse en *temps* en plus de l'analyse en fréquence. Par exemple, la Figure 1.5 représente un signal constitué de deux sinusoïdes de fréquence 10 et 50 Hz (penser à deux notes de musique). La transformée de Fourier permet de retrouver ces deux fréquences, mais ne nous informe pas sur la localisation temporelle du changement de régime du signal. En revanche, ce changement est parfaitement visible lorsque l'on décompose le signal dans une base d'ondelettes. Nous voyons en effet sur la Figure 1.6, que les premiers niveaux de résolution (correspondant aux fréquences faibles) possèdent des coefficients *forts* dans la première "moitié" de la courbe, tandis que pour les derniers niveaux de résolution (fréquences élevées), les coefficients *forts* se trouvent dans la deuxième moitié de la courbe.

Il nous semble dès lors opportun de tenter d'utiliser des bases d'ondelettes en classification fonctionnelle, dans l'espoir d'obtenir de bons résultats lorsque l'information permettant de discriminer est localisée dans une partie de la fonction (ce qui est souvent le cas en pratique). Dans ce cas, seul un petit nombre de coefficients

d'ondelettes sera nécessaire pour résumer cette information alors qu'elle sera répandue sur l'ensemble des coefficients de Fourier. Dans le **dernier chapitre** de cette thèse, nous généralisons la procédure de classification de Biau, Bunea et Wegkamp [8] à des bases d'ondelettes ainsi qu'à des classes de règles plus générales que les plus proches voisins. Nous établissons les propriétés asymptotiques et à distance finie de la méthode envisagée et nous illustrons les performances de cette approche sur des jeux de données réelles et simulées.

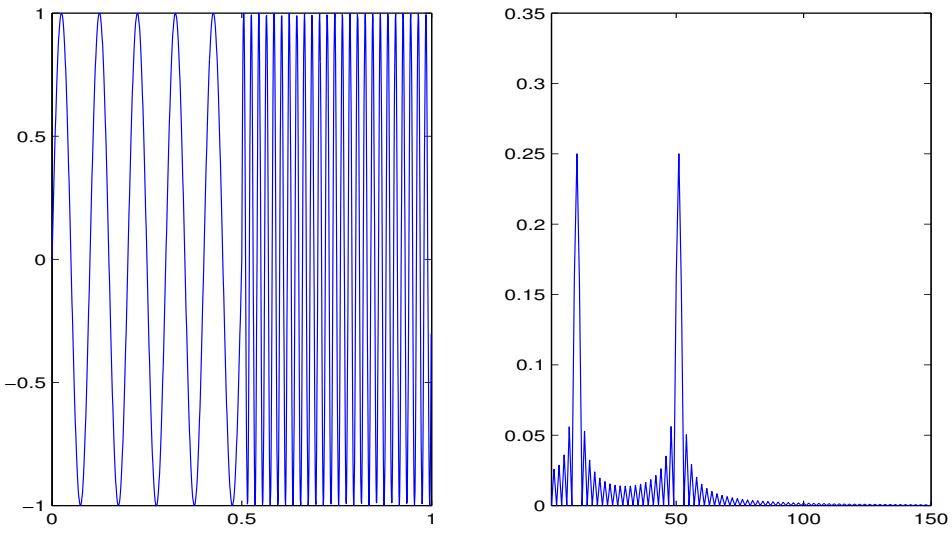

FIG. 1.5 – Gauche : sinus à 10Hz suivi d'un sinus à 50Hz. Droite : sa transformée de Fourier pour les fréquences positives.

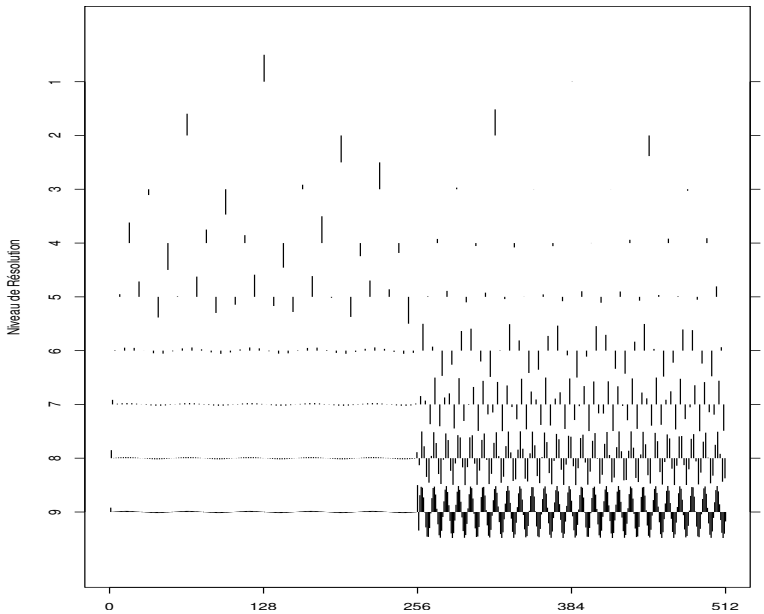

FIG. 1.6 – Transformée du signal de la Figure 1.5 dans une base d'ondelettes.

Bibliographie

[1] C. Abraham, G. Biau, and B. Cadre. On the kernel rule for function classification. *Annals of the Institute of Statistical Mathematics*, 2005. A paraître.

[2] I. Abramson. On bandwidth variation in kernel estimates. *The Annals of Statistics*, 10 :1217–1223, 1982.

[3] A.R. Barron. The convergence in information of probability density estimators. In *Proceedings of the International Symposium of IEEE on Information Theory*, Kobe : Japan, June 19-24 1988.

[4] A.R. Barron, L. Györfi, and E.C. van der Meulen. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Transaction on Information Theory*, 38 :1437–1454, 1992.

[5] A. Berlinet and G. Biau. Iterated modified histograms as dynamical systems. *Journal of Nonparametric Statistics*, 16 :385–401, 2004.

[6] A. Berlinet, I. Vajda, and E.C. van der Meulen. About the asymptotic accuracy of barron density estimates. *IEEE Transactions on Information Theory*, 38 :1437–1454, 1998.

[7] G. Biau. *Méthodes itératives en estimation fonctionnelle et systèmes dynamiques*. PhD thesis, Université Montpellier II, 2000.

[8] G. Biau, F. Bunea, and M. Wegkamp. Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory*, 51 :2163–2172, 2005.

[9] E. Brunel. *Sur l'estimation de la densité et de la fonction de hasard : Estimateurs à noyaux et de Barron, critère de Kullback, applications*. PhD thesis, Université Montpellier II, 1999.

[10] A. Cohen. *Numerical Analysis of Wavelet Methods*. Studies in mathematics and its applications. Elsevier, Amsterdam, 2003.

[11] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer–Verlag, New-York, 1996.

[12] L. Devroye and G. Lugosi. Variable kernel estimates : On the impossibility of tunning the parameters. In E. Giné and D. Mason, editors, *High-Dimensional Probability II*, pages 405–424. Springer–Verlag, New York, 2000.

[13] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation.* Springer–Verlag, New York, 2001.

[14] S. Kullback. A lower bound for discrimination in terms of variation. *IEEE Transactions on Information Theory*, 13 :126–127, 1967.

[15] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33 :1065–1076, 1962.

[16] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27 :832–837, 1956.

[17] S. R. Sain. Multivariate locally adaptive density estimators. *Computational Statistics and Data Analysis*, 39 :165–186, 2002.

[18] S. R. Sain and D. W. Scott. On locally adaptive density estimation. *Journal of the American Statistical Association*, 436 :1525–1534, 1996.

[19] H. Scheffé. A useful convergence theorem for probability distributions. *Annals of Mathematical Statistics*, 18 :434–458, 1947.

[20] G. R. Terrell and D. W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20 :1236–1265, 1992.

[21] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16 :264–280, 1971.

[22] Y. G. Yatracos. Rates of convergence of minimun distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13 :768–774, 1985.

Première partie

Compléments sur les Histogrammes Modifiés

Chapitre 1

Sélection d'Histogrammes Modifiés Itérés

Résumé

Les histogrammes modifiés sont des estimateurs de la densité connus pour posséder de bonnes propriétés de convergence au sens des critères de la théorie de l'information. Ces estimateurs sont construits à partir d'une densité de référence g. Dans ce travail, nous envisageons les histogrammes modifiés comme des systèmes dynamiques fonctionnels qui, à une densité g donnée, associent une trajectoire d'estimateurs $\{B^p g\}_{p\geq 0}$. Berlinet et Biau [3] ont montré, sous certaines hypothèses, que cette trajectoire devient presque sûrement stationnaire. La famille d'estimateurs $\{B^p g, p \geq 0\}$ est alors de cardinal fini. Nous présentons deux méthodes permettant de sélectionner automatiquement un estimateur à l'intérieur de cette famille. Ces deux procédures sont ensuite utilisées pour tenter d'améliorer les performances d'estimateurs à noyau de la densité.

1.1 Introduction

Plaçons-nous sur l'espace mesurable $(\mathbb{R}, \mathcal{B})$, où $\mathcal{B}$ représente la tribu borélienne de $\mathbb{R}$, et désignons par f et g deux densités de probabilité définies par rapport à la mesure de Lebesgue. On rappelle que la distance L_1 et la distance de *Kullback-*

Leibler (ou *entropie relative*) entre f et g sont respectivement définies par :

$$||f-g||_1 = \int |f-g| \qquad \text{et} \qquad D(f,g) = \begin{cases} \displaystyle\int f \log \frac{f}{g} & \text{si } f \ll g \\ \infty & \text{sinon}\,. \end{cases}$$

Il est bien connu (voir par exemple Kullback [11]) que ces deux distances sont liées par l'inégalité dite de *Pinsker* :

$$||f-g||_1 \leq 2D(f,g)\,.$$

Cette relation implique que la distance de Kullback-Leibler induit une topologie plus forte sur l'espace des densités de probabilité que celle associée à la distance L_1.

Dans de nombreux domaines de la statistique (voir Berlinet, Vajda et van der Meulen [6] pour des exemples) la convergence L_1 peut se révéler insuffisante et on lui préfère alors la convergence définie par la distance de Kullback-Leibler. Cependant, trouver des estimateurs de la densité convergeant au sens de l'entropie relative peut soulever quelques difficultés. Remarquons, par exemple, que dans le cas de l'estimateur histogramme usuel $\hat{f}_n$, la quantité $D(f,\hat{f}_n)$ peut être infinie avec une probabilité non nulle.

Afin de pallier cette difficulté, Barron, Györfi et van der Meulen [2] ont montré qu'il était possible de construire un estimateur f_n de f convergeant presque sûrement en entropie relative et en entropie relative moyenne. Cet estimateur, initialement proposé par Barron [1], est appelé *histogramme modifié*. Il est défini à partir de n-échantillon i.i.d. $X_1,\ldots,X_n$ d'une variable aléatoire X de $\mathbb{R}$ possédant f (inconnue) comme densité commune de la manière suivante :

- Soit g une densité connue (dite *densité de référence*) associée à la loi de probabilité ν_g (dite *mesure de référence*).
- Soit ℓ un entier tel que $1 \leq \ell$ et soit $h = 1/\ell$.
- Considérons une partition de $\mathbb{R}$, $P = \{A_1,\ldots,A_\ell\}$ telle que $\nu_g(A_i) = h$, $i = 1,\ldots,\ell$.
- Alors, en notant $a_n = 1/(nh+1)$, on définit l'histogramme modifié f_n par :

$$f_n(x) = \left[(1-a_n)\frac{\mu_n\big(A(x)\big)}{h} + a_n\right] g(x) = \frac{n\mu_n\big(A(x)\big)+1}{nh+1}\, g(x)\,, \qquad (1.1)$$

  où μ_n désigne la mesure empirique associée à l'échantillon $X_1,\ldots,X_n$ et $A(x) = A_i$ si $x \in A_i$.

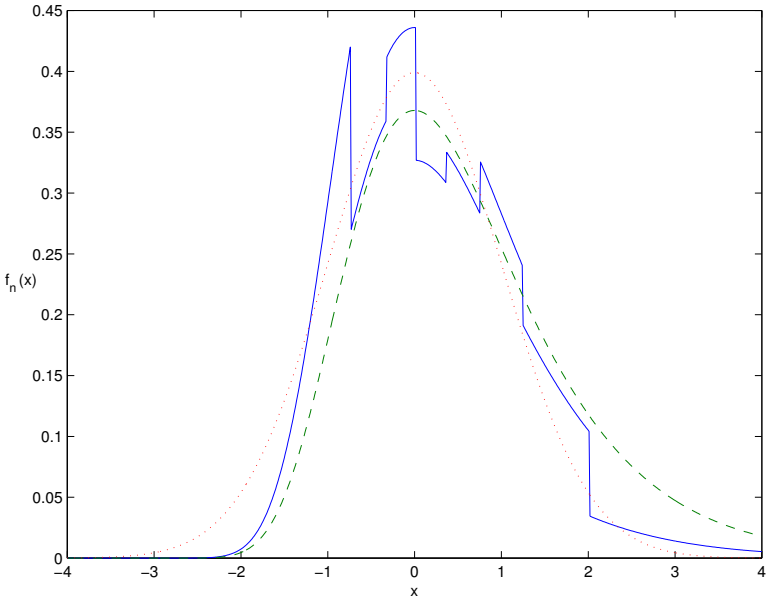

FIG. 1.1 – Histogramme modifié (trait continu) d'une densité gaussienne $\mathcal{N}(0,1)$ (pointillés). La densité de référence est une densité de Gumbel (tirets), $n = 100$, $\ell = 8$.

La première expression de f_n nous présente cet estimateur comme un mélange entre un estimateur de type histogramme et la densité de référence, d'où le nom d'histogramme modifié. La seconde écriture nous montre que cet estimateur est en fait construit comme une déformation par morceaux de la densité de référence g. Sur chaque cellule A_i de la partition, le coefficient multiplicateur de $g(x)$ vaut

$$\frac{n\mu_n(A_i)+1}{nh+1} = \frac{n\mu_n(A_i)+1}{n\nu_g(A_i)+1}\,.$$

Ainsi, sur les classes où la mesure empirique est supérieure à la mesure de référence, on corrige la densité g en la "déformant vers le haut" ; à l'inverse lorsque la mesure empirique est inférieure à la mesure de référence, on déforme g vers le bas (voir Figure 1.1).

Lorsque le nombre de classes ℓ et l'échantillon sont fixés, l' estimateur (1.1) peut être vu comme une fonctionnelle admettant en entrée la densité de référence g et fournissant en sortie une nouvelle densité de probabilité (estimateur de f) que l'on peut noter $B_\ell g$. Puisque $B_\ell g$ est une densité de probabilité, nous pouvons alors la considérer comme densité de référence de l'histogramme modifié : cela conduit à un nouvel estimateur de f, noté $B_\ell^2 g$, qui peut à son tour devenir densité de référence de l'histogramme modifié et ainsi de suite... Ce processus itératif nous

amène à considérer le *système dynamique* B_ℓ défini de la façon suivante :

$$\begin{aligned} B_\ell \ : \ & \mathcal{D} \to \mathcal{D} \\ & g \mapsto B_\ell g \,, \end{aligned} \tag{1.2}$$

où $\mathcal{D}$, l'espace d'états, représente l'ensemble des densités de probabilité sur $\mathbb{R}$.

On dira qu'une densité g de $\mathcal{D}$ est *stationnaire* pour B_ℓ si elle vérifie l'équation $B_\ell g = g$. Une trajectoire $\{B_\ell^p g\}_{p\geq 0}$ contenant une telle densité est dite stationnaire. Dans ce cas, le système n'évolue plus après un nombre *fini* d'itérations. Le *système dynamique* (1.2) a été étudié par Berlinet et Biau [3] qui ont montré la stationnarité de la trajectoire $\{B_\ell^p g\}_{p\geq 0}$ sous certaines hypothèses. Par conséquent, sous ces mêmes hypothèses, la famille d'estimateurs

$$\{B_\ell^p g : p \geq 0\} \tag{1.3}$$

est de cardinal fini. Dans ce travail, nous prolongeons les idées de Berlinet et Biau [3] en étudiant deux méthodes (une pour le critère L_1, l'autre pour le critère de Kullback-Leibler) permettant de sélectionner automatiquement un estimateur à l'intérieur de la famille (1.3).

La suite de ce chapitre se divise en trois parties. Dans la première partie (Section 2) nous rappelons les principaux résultats obtenus par Berlinet et Biau [3] relatifs au système dynamique (1.2). Puis, dans une deuxième partie (Section 3), nous présentons les deux procédures de sélection. Enfin, dans la dernière partie, nous utilisons les méthodes présentées précédemment pour tenter d'améliorer les performances d'un estimateur à noyau de la densité (Section 4).

1.2 Etude du système dynamique

Dans cette partie, nous rappelons les principaux résultats relatifs au système dynamique (1.2). Pour plus de détails et pour les preuves de ces résultats, nous renvoyons le lecteur à Berlinet et Biau [3].

Nous commençons par une remarque élémentaire : une densité g est stationnaire pour B_ℓ si et seulement si $\mu_n(A_i) = h$ pour $i = 1, \ldots, \ell$. Ceci ne peut se produire que si ℓ (le nombre de classes) est un diviseur de n (le nombre d'observations), ce que nous supposerons dans la suite. Pour tout $p \in \mathbb{N}$, on note :

- $q_1^p, q_2^p, \ldots, q_{\ell-1}^p$ les quantiles d'ordre ih ($i = 1, \ldots, \ell - 1$) de la densité $B_\ell^p g$ et on désigne par $\{A_1^p, \ldots, A_\ell^p\}$ la partition associée, *i.e.*,

$$\left\{A_1^p, \ldots, A_\ell^p\right\} = \left\{]-\infty, q_1^p],]q_1^p, q_2^p], \ldots,]q_{\ell-1}^p, \infty[\right\} ;$$

- $\alpha_1^p, \ldots, \alpha_\ell^p$ les coefficients associés à la partition $\{A_1^p, \ldots, A_\ell^p\}$, *i.e.*,

$$\alpha_i^p = \frac{n\mu_n(A_i^p) + 1}{nh + 1}, \quad i = 1, \ldots, \ell\,.$$

Avec ces notations, les densités B_ℓ^{p+1} et B_ℓ^p sont liées par la relation :

$$B_\ell^{p+1} g = \sum_{i=0}^{\ell-1} \alpha_{i+1}^p \mathbf{1}_{]q_i^p, q_{i+1}^p]} B_\ell^p g\,,$$

où $\mathbf{1}_A$ désigne la fonction indicatrice de l'ensemble A. Par conséquent, l'étude de la trajectoire $\{B_\ell^p g\}_{p\geq 0}$ se réduit à l'étude des suites $\{(q_1^p, \ldots, q_{\ell-1}^p)\}_{p\geq 0}$ et $\{(\alpha_1^p, \ldots, \alpha_{\ell-1}^p)\}_{p\geq 0}$. Effectuons, lorsque $\ell \geq 3$, les hypothèses suivantes :

HYPOTHÈSES (HL_i, $i = 2, \ldots, \ell - 1$) : pour p assez grand,

$$q_i^p > X_{\left(\frac{in}{\ell} - 1\right)}\,,$$

où $X_{(1)}, \ldots, X_{(n)}$ désigne le vecteur des statistiques d'ordre des observations $X_1, \ldots, X_n$.

Berlinet et Biau [3] ont alors démontré le théorème suivant :

Théorème 1.2.1 *Supposons que ℓ divise n et que l'hypothèse $(H) = (HL_2) + \ldots + (HL_{\ell-1})$ soit vérifiée dès que $\ell \geq 3$. Alors chaque suite $\{q_i^p\}_{p\geq 0}$ ($i = 1, \ldots, \ell - 1$) devient presque sûrement stationnaire après un nombre fini d'itérations.*

On en déduit facilement le corollaire suivant :

Corollaire 1.2.1 *Sous les hypothèses du Théorème 1.2.1 :*
- *Chaque suite $\{\alpha_i^p\}_{p\geq 0}$ ($i = 1, \ldots, \ell$) atteint la valeur 1 après un nombre fini d'itérations.*
- *La suite de densités $\{B_\ell^p g\}_{p\geq 0}$ est presque sûrement stationnaire.*

Parmi les nombreuses simulations que nous avons effectuées, nous n'avons jamais rencontré une situation où la convergence n'ait pas lieu, ce qui laisse sous-entendre que la condition (H) est vérifiée pour une grande famille de densités g. Ainsi, lorsque le nombre de classes ℓ divise le nombre d'observations n, les suites de quantiles $\{q_i^p\}_{p\geq 0}$ et de coefficients $\{\alpha_i^p\}_{p\geq 0}$ deviennent stationnaires après un nombre fini d'itération (voir Figure 1.2). Le système dynamique B_ℓ engendre alors naturellement une famille d'estimateurs $\{B_\ell^p g : p \geq 0\}$ de cardinal *fini* que l'on notera P_ℓ.

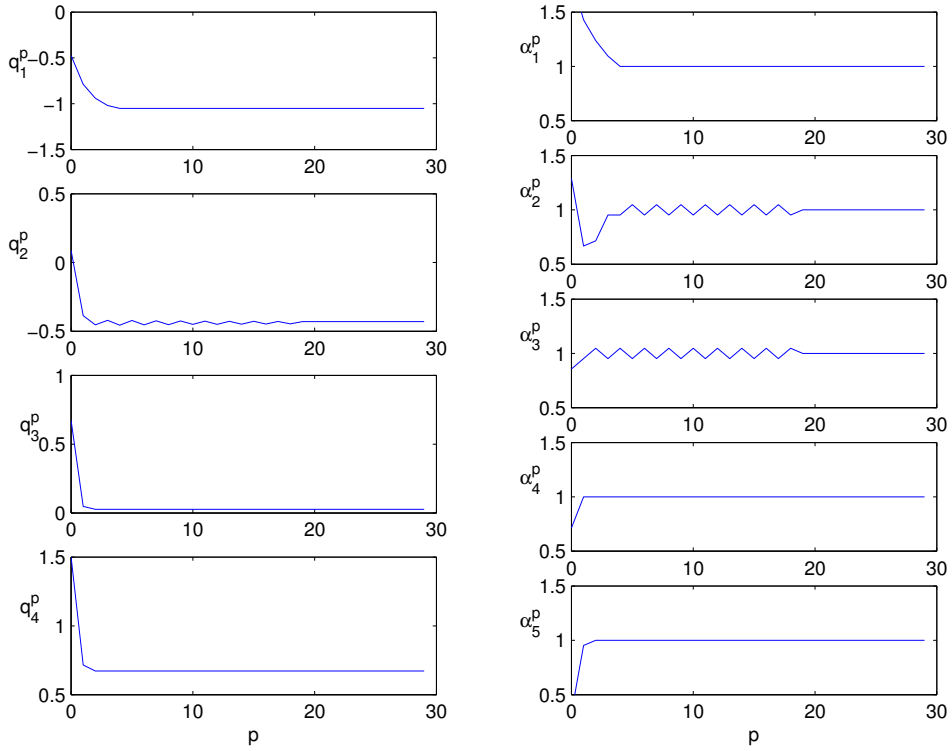

FIG. 1.2 – Evolution des quantiles (gauche) et des coefficients (droite) obtenus avec une densité initiale g de Gumbel. La densité à estimer est une gaussienne $\mathcal{N}(0,1)$, $n = 100$, $\ell = 5$, $n/\ell = 20$.

Bien que le choix de la densité de référence n'affecte le comportement asymptotique de l'histogramme modifié qu'au travers des constantes (voir par exemple Berlinet et Brunel [4]), ce choix se révèle, en revanche, crucial à distance finie. Sur l'exemple de la Figure 1.4 nous voyons que les erreurs L_1 commises par les premiers itérés sont très élevées (proches de 2) en comparaison aux derniers itérés (proches de 0.5). On peut alors se poser le problème du choix d'un estimateur à l'intérieur de cette famille. Une première idée consiste à choisir comme estimateur de f la densité stationnaire de cette famille. Comme nous le montre la Figure 1.3, ce choix est en effet particulièrement efficace lorsque la densité de référence est "éloignée" de la densité à estimer. En revanche, toujours sur ce même exemple, l'erreur L_1 minimale n'est pas commise par la densité stationnaire mais par le cinquième itéré (voir Figure 1.4).

Dans ce chapitre, nous nous posons alors le problème de sélectionner dans la famille finie $\{B_\ell^p g, p \geq 0\}$, un estimateur particulier $B_\ell^{p_0} g$ tel que

$$p_0 \in \underset{p\geq 0}{\operatorname{argmin}} \left\{ ||B_\ell^p g - f||_1 \right\} \qquad \text{ou} \qquad p_0 \in \underset{p\geq 0}{\operatorname{argmin}} \left\{ D(f, B_\ell^p g) \right\}.$$

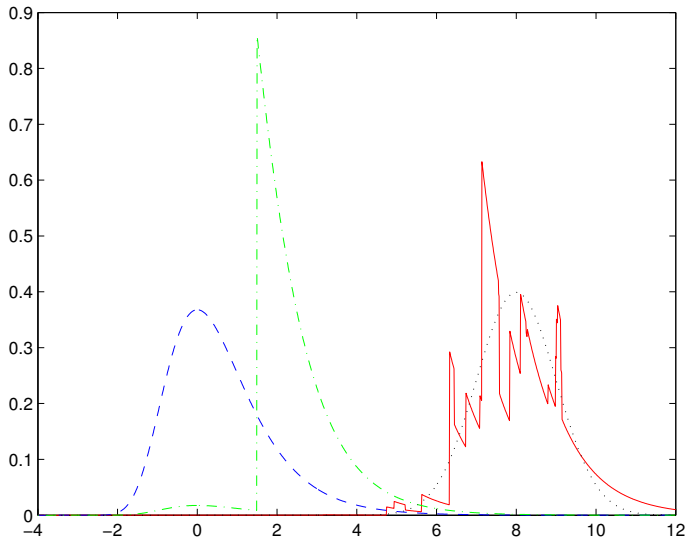

FIG. 1.3 – Comparaison du premier itéré (points-tirets) avec l'estimateur stationnaire (trait plein) d'une densité gaussienne $\mathcal{N}(8,1)$ (pointillés). La densité de référence est une densité de Gumbel (tirets), $n = 100$, $\ell = 5$, $n/\ell = 20$.

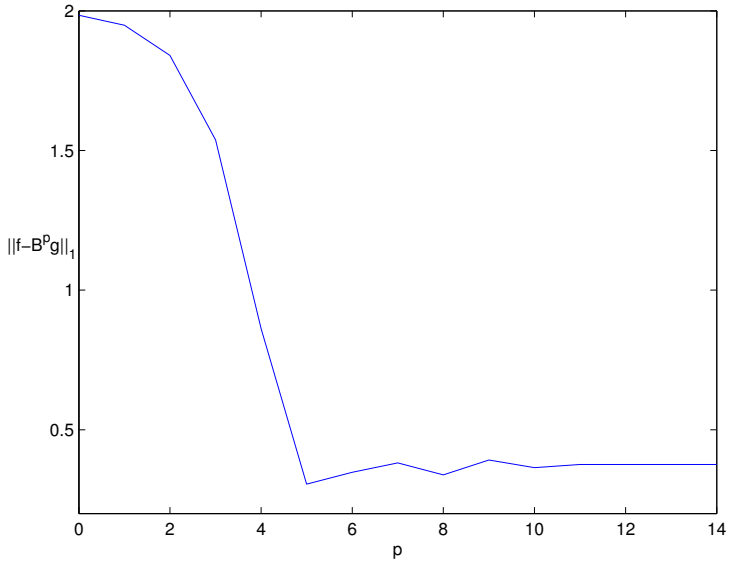

FIG. 1.4 – Erreurs L_1 commises par les 15 premiers itérés de la Figure 1.3.

Le problème auquel nous sommes confrontés est qu'en pratique la densité f est inconnue. Par conséquent nous ne sommes pas à même de calculer les erreurs $||B_\ell^p g - f||_1$ et $D(f, B_\ell^p g)$, et nous devons donc définir une stratégie permettant de choisir automatiquement un estimateur dans la famille $\{B_\ell^p g : p \geq 0\}$ à partir de l'échantillon $X_1, \ldots, X_n$.

1.3 Les procédures de sélection

Nous présentons dans cette section deux méthodes (une pour le critère L_1 et l'autre pour le critère de Kullback-Leibler) permettant de sélectionner automatiquement un estimateur de la densité dans une famille d'estimateurs candidats. Le modèle mathématique se présente de la manière suivante. Soit g une densité de référence fixée. Nous n'effectuons aucune hypothèse sur cette densité. En outre, ce peut être un estimateur de la densité cible f dépendant des observations $X_1, \ldots, X_n$. Pour chaque entier ℓ et p, on note $f_{n,\theta}$ le p-ième histogramme modifié itéré (défini par (1.1) et (1.2)) construit à partir de ℓ classes, $\theta = (\ell, p)$. Etant donné l'ensemble

$$\Theta = \left\{(\ell, p), \ell \text{ divise } n \text{ et } p \geq 0\right\}, \tag{1.4}$$

nous proposons maintenant deux procédures permettant de sélectionner un estimateur particulier dans la famille (finie) $\mathcal{F}_\Theta = \{f_{n,\theta}, \theta \in \Theta\}$.

1.3.1 Critère L_1

La méthode que nous utilisons ici est basée sur les outils combinatoires développés par Devroye et Lugosi [10]. Nous la présentons d'abord dans un contexte général avant de l'appliquer à notre problème. Soit $\mathcal{F}_\Theta = \{f_{n,\theta}\}$ une famille d'estimateurs de la densité pouvant dépendre des n données $X_1, \ldots, X_n$ et paramétrée par $\theta \in \Theta$. Soit $m < n$ un entier qui partage l'échantillon en deux sous-échantillons. On introduit la classe de variables aléatoires

$$\mathcal{A}_\Theta = \left\{\left\{x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x)\right\} : (\theta, \theta') \in \Theta^2\right\}$$

(appelée *classe de Yatracos* associée à Θ en référence aux travaux de Yatracos [17]), et on définit

$$\Delta(f_{n-m,\theta}) = \sup_{A \in \mathcal{A}_\Theta} \left| \int_A f_{n-m,\theta} - \mu_m(A) \right|,$$

où μ_m désigne la mesure empirique associée à l'échantillon $X_{n-m+1}, \ldots, X_n$.

L'*estimateur de la distance minimum* f_n est alors défini comme n'importe quel estimateur $f_{n-m,\theta}$ vérifiant

$$\Delta(f_{n-m,\theta}) \leq \inf_{\theta^* \in \Theta} \Delta(f_{n-m,\theta^*}) + \frac{1}{n}\,,$$

le terme $1/n$ visant simplement à assurer l'existence d'une telle densité.

Rappelons à ce stade que la *distance en variation totale* pour deux mesures de probabilité μ_1 et μ_2 est définie par

$$T(\mu_1, \mu_2) = \sup_{B \in \mathcal{B}} \big|\mu_1(B) - \mu_2(B)\big|\,,$$

où $\mathcal{B}$ représente la tribu borélienne de $\mathbb{R}$. Il est facile de voir que $T(\mu_1, \mu_2) \leq 1$. Lorsque μ_1 possède une densité et $\mu_2 = \mu_n$ (la mesure empirique), on a même $T(\mu_1, \mu_n) = 1$! En particulier, toute tentative de sélectionner une densité $f_{n-m,\theta}$ en minimisant un critère du genre

$$\sup_{B \in \mathcal{B}} \Big| \int_B f_{n-m,\theta} - \mu_m(B) \Big|$$

est sans espoir, puisque la quantité à optimiser est alors constamment égale à 1. Bien entendu, cela ne se produit plus lorsque l'on remplace le supremum sur $\mathcal{B}$ par le supremum sur une classe plus petite, comme par exemple la classe de Yatracos. L'estimateur de la distance minimum est donc construit en minimisant un critère empirique *plus petit* que la variation totale.

Devroye et Lugosi [10] ont montré que l'*estimateur de la distance minimum* f_n vérifie l'inégalité oracle suivante :

$$\int |f_n - f| \leq 3 \inf_{\theta \in \Theta} \int |f_{n-m,\theta} - f| + 4\Delta(f) + \frac{3}{n}\,. \tag{1.5}$$

Le terme $\inf_{\theta \in \Theta} \int |f_{n-m,\theta} - f|$ représente la plus petite erreur qui puisse être commise lorsque l'on approche f par un élément de $\mathcal{F}_\Theta$. Evidemment, la valeur de ce terme d'erreur optimale, qui dépend de la cible f, nous est inconnue. Heuristiquement, l'inégalité (1.5) signifie donc que l'erreur commise par l'estimateur f_n ne dépasse pas trois fois l'erreur minimum sur la classe plus un terme résiduel, $\Delta(f)$, qu'il va falloir s'attacher à contrôler. Ce contrôle peut être effectué via un détour par la théorie de Vapnik et Chervonenkis [16] sur la convergence uniforme de la mesure empirique.

Rappelons que le coefficient de pulvérisation $\mathbf{S}_{\mathcal{A}_\Theta}(m)$ d'un ensemble de m points par la classe d'ensembles $\mathcal{A}_\Theta$ est défini par

$$\mathbf{S}_{\mathcal{A}_\Theta}(m) = \max_{x_1,\ldots,x_m \in \mathbb{R}} \mathrm{Card}\big\{\{x_1, \ldots, x_m\} \cap A : A \in \mathcal{A}_\Theta\big\}\,.$$

Dit autrement, le coefficient de pulvérisation n'est autre que le nombre maximum de sous-ensembles de m points pouvant être obtenus à l'aide de recouvrements par des ensembles de $\mathcal{A}_\Theta$. Des arguments de nature combinatoire (voir Vapnik et Chervonenkis [16]) montrent que

$$\mathbf{E}\big\{\Delta(f)\big\} \leq 2\mathbf{E}\left\{\sqrt{\frac{\log 2\mathbf{S}_{\mathcal{A}_\Theta}(m)}{m}}\right\}.$$

Cette majoration, combinée avec l'inégalité (1.5), nous conduit à l'inégalité suivante

$$\mathbf{E}\left\{\int |f_n - f|\right\} \leq 3 \inf_{\theta\in\Theta} \mathbf{E}\left\{\int |f_{n-m,\theta} - f|\right\} + 8\mathbf{E}\left\{\sqrt{\frac{\log 2\mathbf{S}_{\mathcal{A}_\Theta}(m)}{m}}\right\} + \frac{3}{n} \quad (1.6)$$

qui est centrale dans les travaux de Devroye et Lugosi [10]. On remarquera que ce résultat est *non-asymptotique* (nul besoin d'un passage à la limite pour avoir de l'information) et qu'il ne nécessite *aucune hypothèse de régularité* sur la densité cible f. Il nous semble que ces faits sont suffisamment rares pour mériter d'être soulignés. La seule difficulté, qui réside dans le calcul de $\mathbf{S}_{\mathcal{A}_\Theta}(m)$, est désormais de nature combinatoire.

Cette procédure de sélection sera à nouveau utilisée (et rappelée de manière plus brève) dans la deuxième partie de cette thèse. En particulier, nous nous attacherons à borner le coefficient de pulvérisation $\mathbf{S}_{\mathcal{A}_\Theta}(m)$ pour différentes familles d'estimateurs $\mathcal{F}_\Theta$. Dans le contexte qui est le nôtre, nous rappelons que $f_{n-m,\theta}$ désigne le p-ième histogramme modifié itéré construit à partir de ℓ classes et des $n-m$ premières données. Afin que la famille $\mathcal{F}_\Theta = \{f_{n-m,\theta} : \theta \in \Theta\}$ soit finie, nous cherchons à sélectionner θ dans l'ensemble

$$\Theta = \big\{\theta = (\ell, p) : \ell \text{ divise } n-m \text{ et } p \geq 0\big\}.$$

Nous rappelons que P_ℓ désigne le cardinal de la famille d'estimateurs itérés construit à partir de ℓ classes. Ainsi, en notant $P_{n,m} = \max_{\ell:\ell \text{ divise } n-m} P_\ell$, on a

$$\mathrm{Card}\{\mathcal{F}_\Theta\} = \sum_{\ell:\ell \text{ divise } n-m} P_\ell \leq P_{n,m}(n-m).$$

Il est également facile de voir que lorsque la famille d'estimateurs $\mathcal{F}_\Theta$ est finie, on a

$$\mathbf{S}_{\mathcal{A}_\Theta}(m) \leq \mathrm{Card}\{\mathcal{A}_\Theta\} \leq \mathrm{Card}\{\mathcal{F}_\Theta\}\big(\mathrm{Card}\{\mathcal{F}_\Theta\} - 1\big) \leq P_{n,m}^2(n-m)^2.$$

On déduit alors de (1.6) l'inégalité suivante pour l'estimateur de la distance minimum

$$\mathbf{E}\Big\{ \int |f_n - f| \Big\} \leq 3 \inf_{\theta\in\Theta} \mathbf{E}\Big\{ \int |f_{n-m,\theta} - f| \Big\} + 8\mathbf{E}\Big\{ \sqrt{\frac{\log 2 + 2\log\big(P_{n,m}(n-m)\big)}{m}} \Big\} + \frac{3}{n}\,.$$

Dans de nombreux modèles (voir par exemple les chapitres 1 et 2 de la deuxième partie), il est possible d'obtenir des bornes pour la variable aléatoire $\mathbf{S}_{\mathcal{A}_\Theta}(m)$ dépendant *seulement* de n et m, pas de l'échantillon. En revanche, dans le cadre de travail que nous nous sommes fixés, il semble plus difficile de contrôler la variable aléatoire $P_{n,m}$ indépendamment des données.

1.3.2 Critère de Kullback-Leibler

Nous utilisons ici le critère de validation croisée au sens de l'information de Kullback-Leibler pour sélectionner θ dans

$$\Theta = \big\{\theta = (\ell, p) : \ell \text{ divise } n \text{ et } p \geq 0\big\}\,.$$

Nous rappelons que la distance de Kullback-Leibler entre la densité à estimer f et l'estimateur $f_{n,\theta}$ s'écrit

$$\begin{aligned} D(f, f_{n,\theta}) &= \int f\,\log f \;- \int f\,\log f_{n,\theta} \\ &= \mathbf{E}\Big\{\log f(X)\Big\} - \mathbf{E}\Big\{\log f_{n,\theta}(X)\,|\,X_1,\ldots,X_n\Big\}\,. \end{aligned}$$

La première quantité est l'opposée de l'entropie de f que nous supposons finie. Bien entendu, elle ne dépend pas de θ et n'intervient pas dans la minimisation de l'erreur. Quant au second terme, qui dépend de le densité f inconnue, il peut être approché par :

$$-\frac{1}{n}\sum_{i=1}^{n} \log\, f^i_{n,\theta}(X_i)$$

où $f^i_{n,\theta}$ désigne l'estimateur $f_{n,\theta}$ amputé de la i-ème observation, *i.e.*,

$$f^i_{n,\theta}(x) = \frac{n\mu^i_n\big(A(x)\big)+1}{nh+1} g(x) \qquad \text{avec} \qquad \mu^i_n\big(A(x)\big) = \frac{1}{n-1}\sum_{j\neq i} \mathbf{1}_{\{X_j\in A(x)\}}\,.$$

Le critère de validation croisée consiste alors à choisir θ dans Θ qui minimise

$$-\frac{1}{n}\sum_{i=1}^{n} \log\, f^i_{n,\theta}(X_i)\,. \tag{1.7}$$

Remarquons que ce critère peut être aussi interprété comme l'estimateur de validation croisée du maximum de vraisemblance. En effet, choisir θ dans Θ qui minimise (1.7) équivaut à choisir θ qui maximise

$$\prod_{i=1}^{n} f_{n,\theta}^{i}(X_i)\,,$$

qui apparaît comme un estimateur de la vraisemblance.

La méthode de validation croisée au sens de l'information de Kullback-Leibler révèle un certain nombre de faiblesses pour les estimateurs non paramétriques classiques tels que les histogrammes ou les estimateurs à noyau (Brunel [8]). Les résultats relatifs à ces estimateurs nécessitent des hypothèses précises sur la densité à estimer (notamment sur le comportement des queues de distribution de f). En revanche, de récents travaux de Berlinet et Brunel [4] ont mis en évidence l'intérêt de cette méthode dans le contexte des histogrammes modifiés. Ces auteurs ont, en particulier, montré que ce critère de validation croisée est asymptotiquement optimal pour sélectionner le nombre de classes ℓ de ces estimateurs.

1.4 Application : "améliorer" un estimateur à noyau

Dans cette partie, nous appliquons les méthodes de sélection présentées précédemment au problème suivant. Soit un n-échantillon i.i.d. $X_1, \ldots, X_n$ issu d'une variable aléatoire X admettant f comme densité commune. Etant donné K une fonction réelle, positive et d'intégrale 1 sur $\mathbb{R}$ et h un entier strictement positif, on définit l'estimateur g_n associé au *noyau* K et à la *fenêtre* (ou *paramètre de lissage*) h par

$$g_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big)\,.$$

Nous proposons, dans cette section, d'utiliser cet estimateur comme densité initiale du système dynamique (1.2). Pour chaque entier ℓ et p, on notera $f_{n,\theta}$ le p-ième histogramme modifié construit à partir de ℓ classes ($\theta = (\ell, p)$). Nous adaptons, dans un premier temps, les deux procédures de sélection présentées dans la section précédente pour choisir un estimateur $f_{n,\theta}$ particulier. Puis, dans un second temps, nous comparons les performances de l'estimateur sélectionné à celles de l'estimateur à noyau.

L'estimateur à noyau g_n est flexible, dans la mesure où il laisse à l'utilisateur une grande latitude non seulement dans le choix du noyau K, mais encore dans le

choix du paramètre réel h. Lorsqu'on se limite aux noyaux K positifs, les vitesses de convergence varient peu en fonction de K et les critères essentiels du choix du noyau sont alors la régularité de la courbe à obtenir d'une part, la simplicité et la vitesse de calcul d'autre part. C'est pour cette dernière raison que, dans cette étude, nous nous limiterons à un noyau gaussien.

En revanche, le choix du paramètre de lissage h se révèle crucial aussi bien pour la précision locale que globale de l'estimateur g_n. A ce jour, de nombreuses méthodes de sélection automatique ont été proposées, testées et comparées (voir par exemple Marron [12], Berlinet et Devroye [5]). Nous reprenons ici deux de ces méthodes que nous résumons briêvement dans le paragraphe suivant.

1.4.1 Les estimateurs à noyau considérés

Plug-in L_2 : si $h \to 0$ et $nh \to \infty$ lorsque $n \to \infty$, alors sous des hypothèses standard concernant la régularité de la densité f, la fenêtre qui minimise l'erreur quadratique intégrée moyenne est de la forme :

$$h_{pi} = \frac{\int K^2}{(\int t^2 K)^2} \Big(\int f''^2 \Big)^{-1/5} n^{-1/5} \tag{1.8}$$

(voir Bosq et Lecoutre [7] et Tsybakov [15]). Outre sa nature asymptotique, la largeur de la fenêtre optimale dépend de la densité cible f au travers du paramètre $\int f''^2$ et ne peut donc être utilisée telle quelle dans les calculs. Une façon classique de remédier à ce dernier problème consiste à remplacer la quantité $\int f''^2$ par un estimateur approprié. Cette approche conduit à un ensemble de méthodes que l'on a coutume de regrouper sous le vocable général de *méthodes plug-in*, et qui ont fait l'objet d'une recherche active (voir par exemple Deheuvels [9], Nadaraya [13] ou encore Sheater et Jones [14]). L'approche *plug-in* que nous utilisons ici est celle décrite par Bosq et Lecoutre [7] (Chapitre 6) qui suggèrent de choisir la valeur de $\int f''^2$ associée à la densité normale

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2\sigma^2} \Big),$$

où σ^2 sera estimé par la variance empirique $S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$, avec $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Pour ce choix, lorsque K est le noyau gaussien, la fenêtre (1.8) s'écrit alors

$$h_{pi} \simeq 1.059 \frac{S_n}{n^{1/5}}.$$

On notera $g_{nh_{pi}}$ l'estimateur à noyau associé à cette fenêtre. Bien entendu, même si n est suffisamment grand, rien d'un point de vue théorique ne nous assure de

bonnes performances de l'estimateur associé à la fenêtre (1.8) pour les critères L_1 et de Kullback-Leibler. Cependant, les méthodes *plug-in* L_2 étant souvent utilisées, il n'est pas incohérent d'étudier leurs performances pour des critères autres que le critère L_2.

La méthode du double noyau : dans la méthode du double noyau (Berlinet et Devroye [5]), on considère deux noyaux différents K et L dont les fonctions caractéristiques associées ne coïncident sur aucun voisinage ouvert de l'origine. Le noyau K sera le noyau gaussien et on notera g^1_{nh} l'estimateur associé à ce noyau et à la fenêtre h. Comme suggéré par Berlinet et Devroye [5], nous considérons comme second noyau le noyau polynomial L défini par

$$L(x) = \begin{cases} \frac{7-31x^2}{4} & \text{si } |x| \leq 1/2 \\ \frac{x^2-1}{4} & \text{si } 1/2 \leq |x| \leq 1 \\ 0 & \text{si } 1 \leq |x| \end{cases}$$

et nous désignons par g^2_{nh} l'estimateur associé à ce nouveau noyau L et à la fenêtre h. Le paramètre de lissage h_{dn} est sélectionné en minimisant la distance L_1 entre g^1_{nh} et g^2_{nh} (voir Figure 1.5), c'est-à-dire

$$h_{dn} = \underset{h>0}{\operatorname{argmin}} \int |g^1_{nh} - g^2_{nh}| \,.$$

Berlinet et Devroye [5] ont étudié les propriétés de l'estimateur sélectionné, en montrant notamment que le résultat de convergence suivant est vérifié :

$$\lim_{n\to\infty} \mathbf{E}\left\{ \int |g^1_{nh_{dn}} - f| \right\} = 0 \,.$$

1.4.2 Les densités tests

Afin d'illustrer les performances des méthodes de sélection présentées, nous utilisons 9 densités tests :

D1. La densité uniforme $U_{[0,1]}$ sur $[0,1]$;

D2. La densité gaussienne $\mathcal{N}(0,1)$;

D3. La densité $f(x) = 1/(2\sqrt{x})$ sur $[0,1]$;

D4. La densité d'un mélange de deux lois uniformes $0.5U_{[0,1]} + 0.5\mathcal{N}(1,1)$;

D5. La densité d'un mélange de deux lois gaussiennes $\frac{1}{3}\mathcal{N}(-20, \frac{1}{4}) + \frac{2}{3}\mathcal{N}(0,1)$;

D6. La densité d'un mélange de trois lois uniformes $0.5U_{[0,1]} + 0.25U_{[0,0.1]} + 0.25U_{[0.9,1]}$;

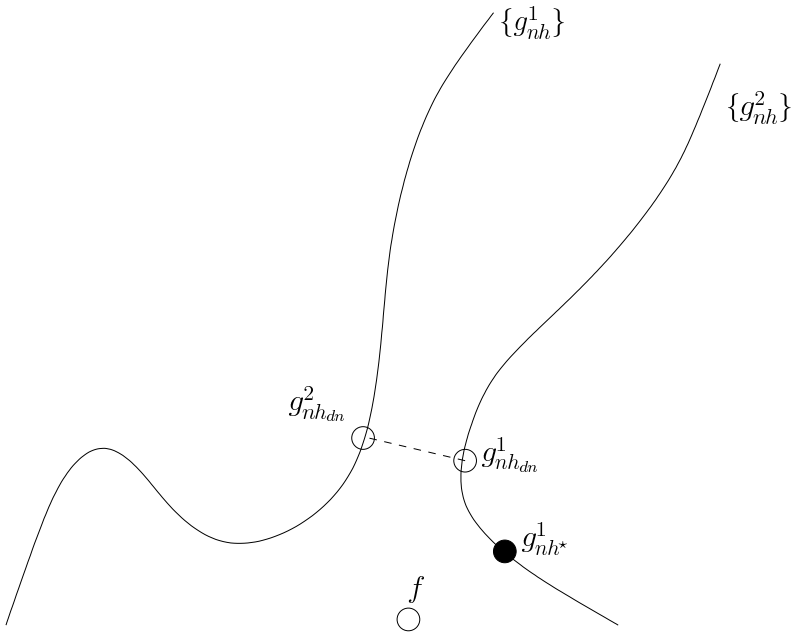

FIG. 1.5 – Deux familles d'estimateurs de la densité dans l'ensemble de toutes les densités. Le paramètre de lissage sélectionné minimise la distance L_1 entre g^1_{nh} et g^2_{nh}.

D7. La densité de $S(X + 0.1)$ où S est la variable aléatoire *signe* qui vaut 1 avec probabilité 0.5 ou -1 avec probabilité 0.5 et X a pour densité $f(x) = 4(1 - x^{1/3})$ sur $[0, 1]$.

D8. La densité d'un mélange de six lois gaussiennes

$$\frac{32}{63}\mathcal{N}\Big(-\frac{31}{21}, \frac{32}{63}\Big) + \frac{32}{63}\mathcal{N}\Big(\frac{17}{21}, \frac{16}{63}\Big) + \frac{8}{63}\mathcal{N}\Big(\frac{41}{21}, \frac{8}{63}\Big) + \frac{4}{63}\mathcal{N}\Big(\frac{53}{21}, \frac{4}{63}\Big) + \frac{2}{63}\mathcal{N}\Big(\frac{59}{21}, \frac{2}{63}\Big) + \frac{1}{63}\mathcal{N}\Big(\frac{62}{21}, \frac{1}{63}\Big);$$

D9. La densité d'un autre mélange de six lois gaussiennes

$$\frac{1}{2}\mathcal{N}(0, 1) + \frac{1}{10}\mathcal{N}(-1, 0.1) + \frac{1}{10}\mathcal{N}(-0.5, 0.1) + \frac{1}{10}\mathcal{N}(0, 0.1) + \frac{1}{10}\mathcal{N}(0.5, 0.1) + \frac{1}{10}\mathcal{N}(1, 0.1)\,.$$

Ces neuf densités sont représentées sur la Figure 1.6. Afin de rendre l'étude la plus pertinente possible, nous avons choisi des densités présentant différents aspects. Ainsi, nous remarquons que :

- les densités D2, D5, D8 et D9 sont lisses ;

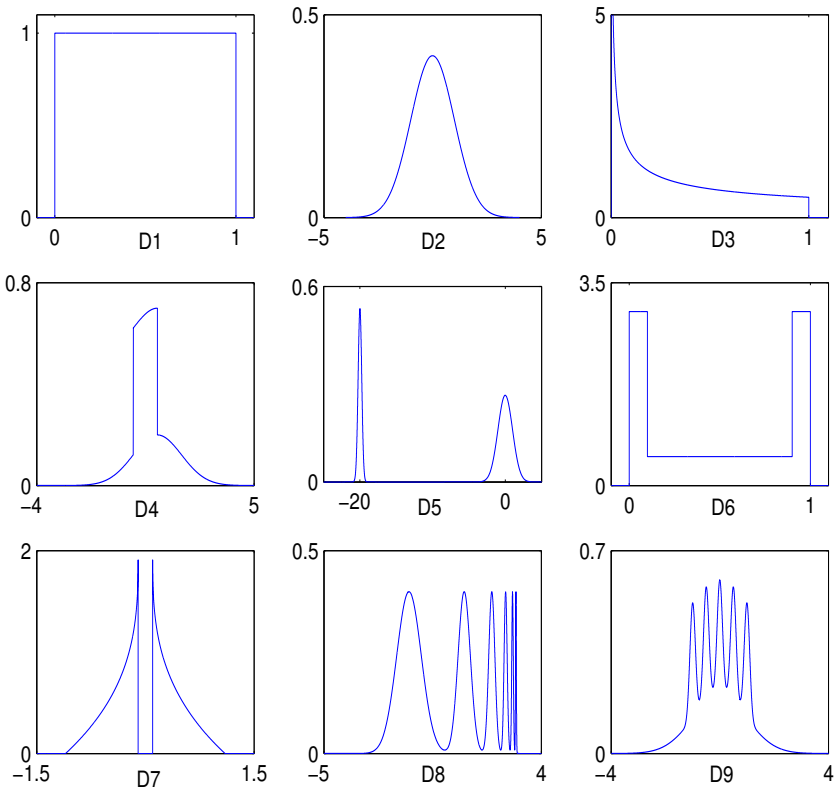

FIG. 1.6 – Les 9 densités tests.

- la densité D3 possède un pic "infini" à l'origine ;
- les densités D4, D6 et D7 possèdent des discontinuités ;
- les densités D5, D6, D7, D8 et D9 sont multimodales.

Pour chaque densité, nous avons simulé 50 échantillons et nous avons appliqués les deux procédures de sélection décrites précédemment.

Dans le cas de la méthode combinatoire (critère L_1), les échantillons sont de taille $n = 500$. Nous considérons les 300 premières observations pour construire les estimateurs et les 200 dernières pour sélectionner les paramètres, c'est-à-dire que $m = 200$. Nous utilisons les estimateurs à noyau $g_n = g_{nh_{pi}}$ et $g_n = g_{nh_{dn}}$ comme densité de référence des histogrammes modifiés. On notera $f_{n-m,\theta}$ le p-ième itéré construit à partir de ℓ classes, $\theta = (\ell, p)$.

On utilise la méthode combinatoire présentée en Section 1.3.1, pour sélectionner un estimateur dans la famille

$$\left\{f_{n-m,\theta}, \theta \in \Theta\right\}$$

avec

$$\Theta = \{\theta = (\ell, p), \ell \in \{5, 10, 15\} \text{ et } p \geq 0\}\,. \tag{1.9}$$

On désignera par f_n l'estimateur sélectionné (*estimateur de la distance minimum*) et par $L_1(f_n)$ l'erreur L_1 commise par cet estimateur. Nous avons représenté dans le Tableau 1.1 :

- les erreurs L_1 moyennes commises par les estimateurs à noyau $g_{nh_{pi}}$ (colonne 2) et $g_{nh_{dn}}$ (colonne 5) ;
- les erreurs L_1 moyennes commises par les estimateurs de la distance minimum lorsque la densité initiale est l'estimateur à noyau $g_{nh_{pi}}$ (colonne 3) et lorsque la densité initiale est l'estimateur à noyau $g_{nh_{dn}}$ (colonne 6) ;
- les écarts relatifs **E.R.** entre les erreurs L_1 des estimateurs à noyau et des histogrammes modifiés, *i.e.*,

$$\textbf{E.R.} = \frac{L_1(g_n) - L_1(f_n)}{L_1(g_n)}\,.$$

Pour la méthode de validation croisée (critère de Kullback-Leibler) présentée en Section 1.3.2, les échantillons sont de tailles $n = 150$. On notera f_n^{cv} l'estimateur sélectionné à l'intérieur de la famille

$$\left\{f_{n,\theta}, \theta \in \Theta\right\}$$

où Θ désigne l'espace des paramètres (1.9). On notera également $KL(f_n^{cv})$ l'erreur au sens de la distance de Kullback-Leibler commise par l'estimateur f_n^{cv}. Les résultats sont présentés dans le Tableau 1.2.

Nous observons tout d'abord que les performances des estimateurs à noyau varient suivant la densité à estimer. Ces estimateurs se comportent en effet très bien pour les densités uniformes et gaussiennes (D1 et D2). Ils sont en revanche moins performants pour les cibles plus complexes (D3, D5, D6 et D8 par exemple) Les résultats de ces simulations mettent également en relief le problème bien connu du choix de la fenêtre. En effet, la performance de la procédure de sélection de la fenêtre de l'estimateur à noyau varie en fonction de la densité à estimer : la méthode plug-in donne de meilleurs résultats pour les densités D2, D7 et D9 tandis que la méthode du double noyau est plus performante pour les modèles D5, D6 et D8. Ces disparités peuvent être corrigées, dans une certaine mesure,

f	$L_1(g_{n,h_{pi}})$	$L_1(f_n)$	**E.R.**	$L_1(g_{n,h_{dn}})$	$L_1(f_n)$	**E.R.**
D1	0.1799	0.2057	**-0.1434**	0.2120	0.2172	**-0.0245**
D2	0.0747	0.0868	**-0.1620**	0.1717	0.1482	0.1369
D3	0.3651	0.2956	0.1904	0.3381	0.2884	0.1470
D4	0.2226	0.2052	0.0782	0.2222	0.1986	0.1062
D5	1.2047	0.3538	0.7063	0.4625	0.3164	0.3159
D6	0.7508	0.3064	0.5919	0.3605	0.3182	0.1173
D7	0.3136	0.2783	0.1126	0.3980	0.2550	0.3593
D8	0.5465	0.3671	0.3283	0.2757	0.2732	0.0098
D9	0.3393	0.2755	0.1880	0.4601	0.2660	0.4219

TAB. 1.1 – La méthode combinatoire pour la sélection de θ.

f	$KL(g_{n,h_{pi}})$	$KL(f_n^{cv})$	**E.R.**	$KL(g_{n,h_{dn}})$	$KL(f_n^{cv})$	**E.R.**
D1	0.1125	0.1218	**-0.0827**	0.1163	0.1844	**-0.5856**
D2	0.0187	0.0273	**-0.4599**	0.0613	0.0491	0.1990
D3	0.3600	0.1851	0.4858	0.3789	0.1803	0.5241
D4	0.0762	0.0813	**-0.0669**	0.0921	0.0815	0.1151
D5	1.3177	0.1382	0.8951	0.3827	0.1066	0.7215
D6	0.5582	0.1683	0.6985	0.2289	0.1514	0.3386
D7	0.2033	0.1275	0.3728	0.4190	0.1494	0.6434
D8	0.3231	0.2132	0.3401	0.1585	0.1600	**-0.0095**
D9	0.0936	0.0976	**-0.0427**	0.1684	0.1017	0.3961

TAB. 1.2 – Le critère de validation croisée pour la sélection de θ.

en sélectionnant un histogramme modifié itéré. Nous remarquons en effet que les erreurs (L_1 ou de Kullback-Leibler) commises par les histogrammes modifiés sélectionnés ne diffèrent guère suivant la procédure de sélection du paramètre de lissage de l'estimateur à noyau. De plus, à l'exception des modèles D1 et D2 (pour lesquels l'estimateur à noyau se comporte très bien), les performances des histogrammes modifiés sélectionnés sont meilleures que celles des estimateurs à noyau. Cette amélioration devient même très significative pour les modèles D5, D6, D7, D8 et D9 (voir Figure 1.7 et Figure 1.8). D'une certaine manière, ces

remarques nous amènent à considérer les histogrammes modifiés comme un outil potentiel permettant d'améliorer, ou tout au moins de corriger dans certaines situations, les performances des estimateurs à noyau.

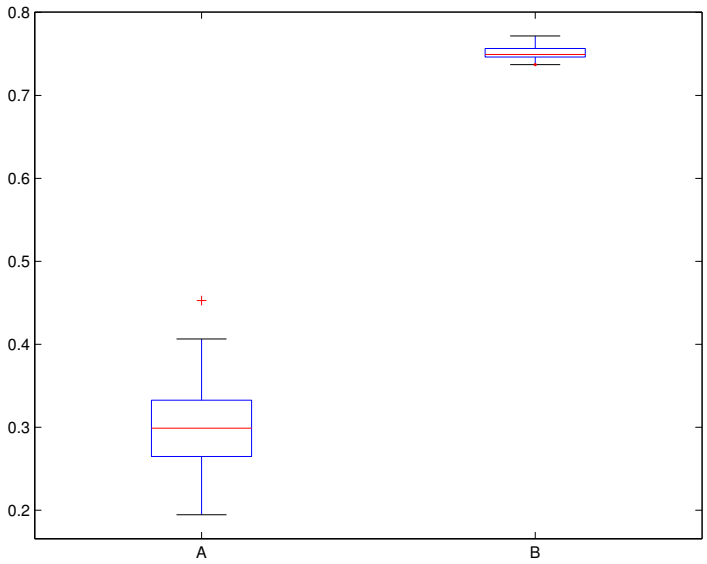

FIG. 1.7 – Boxplots (pour les 50 répétitions) des erreur L_1 commises par les histogrammes modifiés sélectionnés par la méthode combinatoire (A) et les estimateurs à noyau $g_{n,h_{pi}}$ (B) pour le modèle D6.

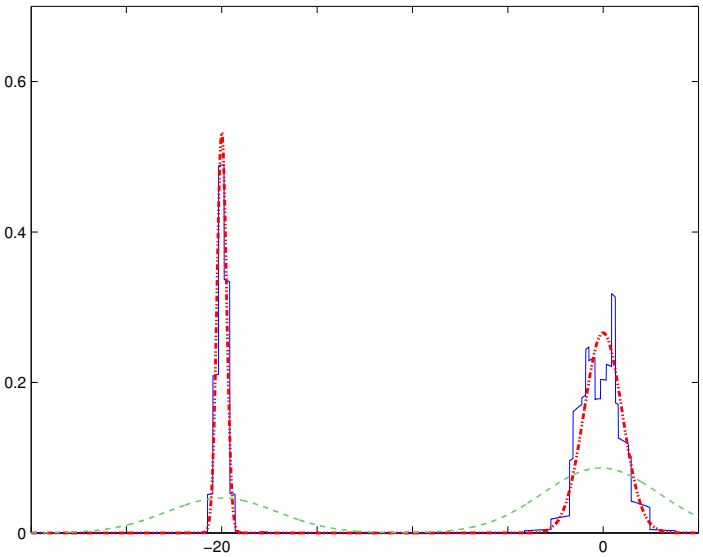

FIG. 1.8 – Densité D5 (pointillés), estimateur à noyau $g_{n,h_{pi}}$ (tirets), histogramme modifié sélectionné par la méthode combinatoire (trait plein).

Bibliographie

[1] A. R. Barron. The convergence in information of probability density estimators. In *Proceedings of the International Symposium of IEEE on Information Theory*, Kobe : Japan, June 19-24 1988.

[2] A. R. Barron, L. Györfi, and E. C. van der Meulen. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Transactions on Information Theory*, 38 :1437–1454, 1992.

[3] A. Berlinet and G. Biau. Iterated modified histograms as dynamical systems. *Journal of Nonparametric Statistics*, 16 :385–401, 2004.

[4] A. Berlinet and E. Brunel. Cross-validated density estimates based on Kullback-Leibler information. *Journal of Nonparametric Statistics*, 16 :493–513, 2003.

[5] A. Berlinet and L. Devroye. A comparison of kernel density estimates. *Publications de l'Institut de Statistique de l'Université de Paris*, 38 :3–59, 1994.

[6] A. Berlinet, I. Vajda, and E.C. van der Meulen. About the asymptotic accuracy of Barron density estimates. *IEEE Transactions on Information Theory*, 44 :999–1009, 1998.

[7] D. Bosq and J.P. Lecoutre. *Théorie de l'Estimation Fonctionnelle.* Economica, Paris, 1987.

[8] E. Brunel. *Sur l'estimation de la densité et de la fonction de hasard : Estimateurs à noyaux et de Barron, critère de Kullback, applications.* PhD thesis, Université Montpellier II, 1999.

[9] P. Deheuvels. Conditions nécessaires et suffisantes de convergence ponctuelle presque sûre et uniforme presque sûre des estimateurs de la densité. *Comptes Rendus Mathématiques de l'Académie des Sciences de Paris*, 278 :1217–1220, 1974.

[10] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation.* Springer–Verlag, New York, 2001.

[11] S. Kullback. A lower bound for discrimination in terms of variation. *IEEE Transactions on Information Theory*, 13 :126–127, 1967.

[12] J.S. Marron. Automatic smoothing parameter selection : a survey. *Empirical Economics*, 13 :187–208, 1988.

[13] E.A. Nadaraya. On the integral mean square error of some nonparametric estimates for the density function. *Theory of Probability and its Applications*, 19 :133–141, 1974.

[14] S.J. Sheater and M.C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, B53 :683–690, 1991.

[15] A.B. Tsybakov. *Introduction a l'estimation non-paramétrique.* Springer, 2004.

[16] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16 :264–280, 1971.

[17] Y.G. Yatracos. Rates of convergence of minimun distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13 :768–774, 1985.

Chapter 2

Effective Construction of Modified Histograms in Higher Dimensions*

Abstract

Density estimation raises delicate problems in higher dimensions especially when strong convergence is required and data marginals can be highly correlated. Modified histograms have been introduced to circumvent the problem of low bin counts when convergence is considered in the sense of information divergence. These estimates are defined from some reference probability density and an associated partition which is defined in the univariate case from the quantiles of the reference density. Therefore, in the multivariate case, the definition of the partition causes an additional problem related to the lack of total order. In this paper, we present a method for constructing modified multivariate histograms such that the corresponding partition is well adapted to the observed data. The approach is based on a data-driven coordinate system selected by cross-validation. We discuss the performance of our estimate with the help of a finite sample simulation study.

2.1 Introduction

We consider the problem of estimating an unknown probability density f defined on $\mathbb{R}^d$ based on independent, identically distributed observations $X_1, \ldots, X_n$ from f. Here the quality of estimation will be evaluated by a nonnegative divergence

*Article écrit en collaboration avec Alain Berlinet et publié dans le livre *Statistical Modeling and Analysis for Complex Data Problems*, ed P. Duchesne et B. Rémillard.

$F(f, f_n)$. Of interest are estimators f_n consistent in the sense

$$\lim_{n\to\infty} F(f, f_n) = 0 \; a.s. \quad \text{or} \quad \lim_{n\to\infty} \mathbf{E}F(f, f_n) = 0$$

where $\mathbf{E}$ denotes the expectation with respect to the random vector $(X_1, \ldots, X_n)$ figuring in the estimate f_n. The two most important divergences in mathematical statistics and information theory are the total variation V and the information divergence D. They are defined by

$$V(f, g) = \frac{1}{2} \int |f - g| = \frac{1}{2} \|f - g\|_{L_1}$$

$$D(f, g) = \begin{cases} \displaystyle\int f \log \frac{f}{g} & \text{if } f \ll g \\ \infty & \text{otherwise,} \end{cases}$$

where $\ll$ denotes absolute continuity. It is well known (cf. Csizár [13], Kemperman [20], and Kullback [22]) that for all densities f and g, $V(f, g)$ and $D(f, g)$ are linked by the following inequality, called *Pinsker* inequality:

$$2V^2(f, g) \leq D(f, g)\,,$$

which entails that the information divergence is topologically stronger than the total variation.

In numerous application fields of statistics (data compression, telecommunication networks, classification, pattern recognition or neural networks...), the consistency defined by total variation may prove inadequate. This is the case when precise estimation of tail probabilities or convergence of integrals of various functionals are required (see Berlinet, Vajda and van der Meulen [7] for discussion). Another concern with convergence in total variation is that, given any sequence of density estimates, the rate of convergence of the expected L_1 error can be arbitrary slow (Devroye [15]). Therefore stronger topologies such as information divergence are often preferred.

Classical nonparametric density estimates such as kernel estimates and histograms are not universally consistent in information divergence (see Hall [19]). The modified histograms introduced by Barron [2] and Barron, Györfi and van der Meulen [3] circumvent this problem. They are defined as follows.

Suppose that we observe independent $\mathbb{R}^d$-valued random variables $X_1, \ldots, X_n$ with common unknown density f.

- Denote by g a known density on $\mathbb{R}^d$ and by ν_g the associated probability measure;
- Define a sequence of integers $\{\ell_n\}_{n\geq 1}$ such that $1 \leq \ell_n \leq n$ and let $h_n = 1/\ell_n$;
- Introduce a sequence of partitions $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, \ldots, A_{n,\ell_n}\}$, $n \geq 1$, such that $\nu_g(A_{n,i}) = h_n$, $i = 1, \ldots, \ell_n$;
- Finally consider, for $a_n = 1/(nh_n + 1)$ the following estimator f_n

$$f_n(x) = \left[(1 - a_n)\frac{\mu_n\big(A_n(x)\big)}{h_n} + a_n\right] g(x) = \frac{n\mu_n\big(A_n(x)\big) + 1}{nh_n + 1} g(x)\,. \qquad (2.1)$$

  where μ_n stands for the empirical measure associated with the sample $X_1, \ldots, X_n$ and $A_n(x)$ stands for $A_{n,i}$ if $x \in A_{n,i}$.

The estimate (2.1) is a mixture of a histogram-type density estimate and the known density g. It can also be regarded as a piecewise transformation of g itself, which is thus often called in this context the *reference density*.

Under the conditions

$$D(f, g) < \infty, \qquad \lim_{n\to\infty} h_n = 0 \text{ and } \lim_{n\to\infty} nh_n = \infty\,,$$

almost sure consistency in information divergence and consistency in expected information divergence have been proved by Barron, Györfi and van der Meulen [3]. For further results on modified histograms we refer the reader to Berlinet and Brunel [5], Berlinet, Györfi and van der Meulen [6], Berlinet and Biau [4] and Györfi, Liese, Vajda and van der Meulen [18].

When $d = 1$, the quantiles of the reference density are used to partition $\mathbb{R}$. Formally, denoting by G the distribution function associated with the probability density g (g is defined on $(a; b)$, a and b may be infinite), we set

$$A_{n,i} = \Big(G^{-1}\Big(\frac{i-1}{\ell_n}\Big), G^{-1}\Big(\frac{i}{\ell_n}\Big)\Big],\ i = 1, \ldots, \ell_n\,,$$

where the interval $(.,.]$ is understood as open on the left and closed on the right only when its upper bound is finite and where G^{-1} is the quantile function defined by

$$\begin{cases} G^{-1}(\alpha) = \inf\{x : G(x) \geq u\} & \text{if } 0 < \alpha < 1 \\ G^{-1}(\alpha) = a & \text{if } \alpha = 0 \\ G^{-1}(\alpha) = b & \text{if } \alpha = 1\,. \end{cases} \qquad (2.2)$$

Thus, univariate modified histograms result from the comparison of the quantiles of g with the empirical quantiles. Under mild conditions the choice of g does not affect dramatically the asymptotics. Practically, however, g should not be "too far" from f, so that the comparison between the empirical measure and the reference density over the partition makes sense.

For $d \geq 2$, the choice of such a partition is much more delicate because the lack of total order does not allow to define multivariate quantiles having the same properties as univariate ones. The aim of this paper is to propose a method for constructing multivariate modified histograms. In Section 2, we give two algorithms to construct this estimate. The first one uses rectangles to partition $\mathbb{R}^d$ (as for the standard multivariate regular histogram estimate). However, the performance of this estimate becomes poor in the presence of high correlation among components of the data vector. This leads us to a more effective method which results from a transformation of these rectangles. We use the data-driven coordinate system introduced by Chaudhuri and Sengupta [12]. In Section 3, we select this coordinate system by cross-validation and we end with some simulations showing the very good performance of the second estimate.

2.2 Construction of the estimator

Not any sequence of partitions of $\mathbb{R}^d$ has good properties to build consistent estimates. The following concept, introduced by Csizár [14] has a great importance in the definition of suitable partitions.

Definition 2.2.1 *A sequence of partitions $\{\mathcal{P}_n\}$ of $\mathbb{R}^d$ is said to be ν-approximating for a given probability measure ν if, for every measurable set A and for every $\epsilon > 0$, there is for all n sufficiently large a set A_n equal to a union of sets in $\{\mathcal{P}_n\}$ such that*

$$\nu(A_n \triangle A) < \epsilon\,,$$

where $A_n \triangle A$ denotes the symmetric difference of A_n and A.

As proved by Barron, Györfi and van der Meulen [3] this notion is basic in the proof of consistency of modified histograms.

The partition of a univariate modified histogram is computed from the quantiles of the reference density. Several authors have proposed extensions of quantiles to multidimensional spaces. Chaudhuri [11] proposed the notion of geometric quantile which generalizes the spatial median studied earlier (see Brown [8], Kermperman [21]). Chakraborty [10] transformed these geometric quantiles in order to obtain affine equivariant multivariate quantiles. Liu, Parelius and Singh [23]

proposed to define affine equivariant multivariate quantiles using depth analysis. They generalized half-space depth quantiles introduced by Tuckey [25].

Given a measure ν, using *quantile contour plots* of Chakraborty [10] or *center outward quantiles surfaces* of Liu, Parelius and Singh [23], one can construct a sequence of partitions $\mathcal{P}_n = \{A_{n,1}, \ldots, A_{n,\ell_n}\}$ such that $\nu(A_{n,i}) = h_n$ $(i = 1, \ldots, \ell_n)$. These sequences are nested in the sense that for all n there exists a sequence

$$B_{n,1} \subset B_{n,2} \subset \ldots \subset B_{n,\ell_n}$$

such that

$$\forall i = 1, \ldots, \ell_n, \; A_{n,i} = B_{n,i} - \bigcup_{j=1}^{i-1} B_{n,j} \, . \tag{2.3}$$

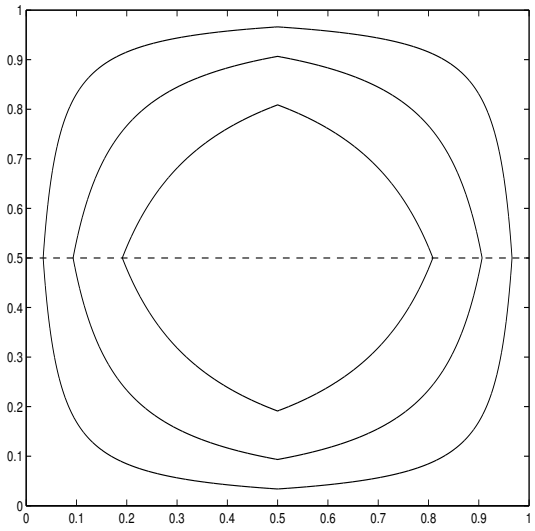

Figure 2.1: Half-space depth center-outward quantile surface of order 0.25, 0.5 and 0.75 for the uniform distribution on the square $[0,1]^2$.

Such a sequence of partitions is not ν-approximating for any measure ν. For example, let ν be the uniform distribution on the square $[0,1]^2$ and consider a sequence of partitions built from halfspace depth quantiles (see Liu, Parelius and Singh [23]). Formally, for $i = 1, \ldots, \ell_n$, $B_{n,i}$ is *a half-space depth center-outward quantile surface* of order i/ℓ_n and $A_{n,i}$ is defined by (2.3) (see Figure 2.1). Consider the horizontal line which passes through the center of the square (dashed line in Figure 2.1). This line splits the square into two rectangles. If A denotes one of these rectangles, it is easily seen that for all sets A_n equal to a

union of sets in $\mathcal{P}_n$, we have

$$\nu(A_n \triangle A) = 0.5\,,$$

which entails that $\mathcal{P}_n$ is not ν-approximating.

Other authors have defined quantiles in multidimensional spaces (see Brown and Hettmansperger [9], Eddy [16] and [17]), but as far as we know none permits the construction of modified histograms for any reference density. This leads us to restrict our attention to a certain class of reference densities.

2.2.1 Regular modified histograms

The standard regular (unmodified) histogram is defined by a partition of $\mathbb{R}^d$ into rectangular cells of widths $h_1, \ldots, h_d$. The goal of this paragraph is the adaptation of this partition to modified histograms. In this regard we only consider reference densities g such that

$$g(x_1, \ldots, x_d) = g_1(x_1) \ldots g_d(x_d)\,, \tag{2.4}$$

where $g_1, \ldots, g_d$ are univariate densities. For $j = 1, \ldots, d$, we denote by G_j the distribution function associated with the probability density g_j and by G_j^{-1} the quantile function as in (2.2).

Given i.i.d. observations $X_1, \ldots, X_n$ from a density f on $\mathbb{R}^d$ and given a reference density g such as (2.4), modified multivariate histograms are built as follows:

- Set $\ell = \ell_1 \ldots \ell_d$ with $\ell_1, \ldots, \ell_d$ positive integers and let $h_j = 1/\ell_j$ for $j = 1, \ldots, d$;
- For $j = 1, \ldots, d$ and $i_j = 1, \ldots, \ell_j - 1$, compute univariate quantiles of order $i_j h_j$ of g_j. Denote by q_{j,i_j} these quantiles, *i.e.*,

  $$q_{j,i_j} = G_j^{-1}(i_j h_j)$$

  with the convention $q_{j,0} = -\infty$ and $q_{j,\ell_j} = \infty$;
- Consider the grid defined by the above family $\{q_{j,i_j}\}$; this grid leads to a partition of $\mathbb{R}^d$ into ℓ hyperrectangles (see Figure 2.2), say

  $$A_{i_1,\ldots,i_d} = \prod_{j=1}^{d} (q_{j,i_j-1}, q_{j,i_j}]\,; \tag{2.5}$$

- For each of these cells, compute the empirical measure:

$$\mu_n(A_{i_1,\ldots,i_d}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \in A_{i_1,\ldots,i_d}\}} ;$$

- The *regular modified multivariate histogram density estimate* f_n is defined by:

$$f_n(x) = \frac{n\mu_n\big(A(x)\big) + 1}{nh + 1} g(x) \tag{2.6}$$

where $h = h_1 \ldots h_d$ and $A(x)$ stands for $A_{i_1,\ldots,i_d}$ if $x \in A_{i_1,\ldots,i_d}$.

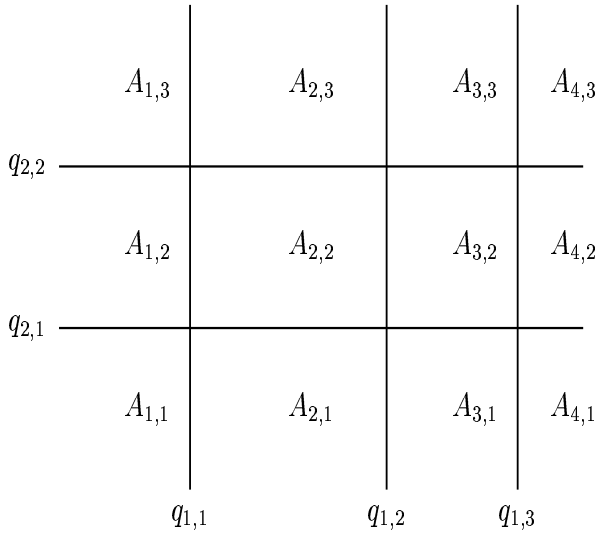

Figure 2.2: Example of partition in $\mathbb{R}^2$: $\ell_1 = 4$, $\ell_2 = 3$.

Denote by ν_g the probability measure associated with the reference density g. It is easily seen that, for any set $A_{i_1,\ldots,i_d}$,

$$\nu_g(A_{i_1,\ldots,i_d}) = h\,.$$

Consistency in information divergence and expected information divergence is established in our next theorem.

Theorem 2.2.1 *Let f_n be the regular modified histogram defined in* (2.6). *Assume that $D(f,g) < \infty$.*

(i) If $h_j = h_{j,n}$ $(j = 1, \ldots, d)$ *and* $\lim_{n\to\infty} \max_{1\le j\le d} h_{j,n} = 0$ *then the sequence of partition*

$$\{\mathcal{P}_n\} = \{A_{n,i_1,\ldots,i_d}\}_{\substack{1\le j\le d \\ 1\le i_j\le \ell_{j,n}}}$$

defined in (2.5) *is* ν_g*-approximating.*

(ii) Moreover assume that $\lim_{n\to\infty} nh_n = \infty$ $(h_n = h_{1,n} \ldots h_{d,n})$*, then*

$$\lim_{n\to\infty} D(f, f_n) = 0 \ a.s \ \textit{and} \ \lim_{n\to\infty} \mathbf{E}D(f, f_n) = 0\,.$$

Proof We first prove (i). Let S denote the support of ν and $\overline{S}$ its complement in $\mathbb{R}^d$. With a slight abuse of notation, we denote

$$\{\mathcal{P}_n\} = \{A_{n,1}, \ldots, A_{n,\ell_n}\}\,.$$

For $j = 1, \ldots, d$, let $a_j = \inf\{x \in \mathbb{R} : g_j(x) \neq 0\}$ and $b_j = \sup\{x \in \mathbb{R} : g_j(x) \neq 0\}$ (a_j and b_j may be infinite). Let S_j (resp. $\overline{S}_j$) be the projection of S (resp. $\overline{S}$) on (a_j, b_j). $\overline{S}_j$ is the union of k_j distinct intervals of length $\tau_j(i)$ $(i = 1, \ldots, k_j)$. For $x = (x_1, \ldots, x_d) \in S$, let $p_j(x_j)$ denote the number of intervals of $\overline{S}_j$ before x_j and consider for $j = 1, \ldots, d$

$$\begin{aligned} T_j : \quad S_j &\mapsto \mathbb{R} \\ x_j &\to x_j - \sum_{i=1}^{p_j(x_j)} \tau_j(i) \end{aligned}$$

and

$$\begin{aligned} T : \qquad S &\mapsto \mathbb{R}^d \\ (x_1, \ldots, x_d) &\to (T_1(x_1), \ldots, T_d(x_d))\,. \end{aligned}$$

The application T allows to remove the hyperrectangles R of $\mathbb{R}^d$ such that $\nu_g(R) = 0$. Fix a measurable set A. If A_n is equal to a union of sets in $\mathcal{P}_n$ then

$$\nu_g(A_n \,\triangle\, A) = \nu_g^T(T(A_n) \,\triangle\, T(A))\,,$$

where

$$\begin{aligned} \nu_g^T : \quad T(S) &\mapsto [0, 1] \\ A &\to \nu_g\big(T^{-1}(A)\big)\,. \end{aligned}$$

Therefore, it suffices to prove that the partition

$$\{\mathcal{P}_n^T\} = \{T(A_{n,1}), \ldots, T(A_{n,\ell_n})\}$$

is ν_g^T-approximating. Note that $T(A_{n,i})$ $(i = 1, \ldots, \ell_n)$ are hyperrectangles of $\mathbb{R}^d$ such that $\nu_g^T(T(A_{n,i})) = h_{1,n} \ldots h_{d,n}$. Denoting by $T_j(A_{n,i})$ the projection of $T(A_{n,i})$ over the j-th component of $\mathbb{R}^d$, it is easily seen that the length of $T_j(A_{n,i})$ tends to zero as $\lim_{n\to\infty} h_{j,n} = 0$. For each ball B centered at some point x_0, we deduce that

$$\lim_{n\to\infty} \max_{\{i:T(A_{n,i})\cap B\neq\emptyset\}} \operatorname{diam}(T(A_{n,i})) = 0\,,$$

where $\operatorname{diam}(E) = \sup_{x,y\in E} d(x,y)$ and $d(x,y)$ denotes the distance in $\mathbb{R}^d$. It follows from Csiszár (1973, p. 168) that the partition $\{\mathcal{P}_n^T\}$ is ν_g^T-approximating. Combining (i) with Theorem 2 in Barron, Györfi and van der Meulen [3] gives (ii). ■

2.2.2 Influence of correlation

Through an example, we study the influence of the shape of the data vector on the performance of the density estimate defined in (2.6). Table 2.1 gives the total variation $V(f, f_n)$ and the information divergence $D(f, f_n)$ between binormals and their standard modified histogram estimates. Simulated binormals have 0 mean, unit standard deviation and varying correlation (from 0 to 0.95), the size of the samples is $n = 250$. To construct the estimate, we choose $\ell_1 = \ell_2 = 5$ and the reference density g is a product of Gumbel densities:

$$g(x,y) = \exp\big(-x - \exp(-x)\big) \exp\big(-y - \exp(-y)\big)\,. \tag{2.7}$$

ρ	0	0.25	0.5	0.75	0.95
$V(f, f_n)$	0.19	0.19	0.21	0.22	0.34
$D(f, f_n)$	0.32	0.33	0.35	0.43	0.74

Table 2.1: Total variation and information divergence according to the correlation.

Results are clearly better in the presence of weak correlation. One can explain it as follows. On Figure 2.3, we have represented a sample of size $n = 250$ simulated from a binormal with 0 mean and identity variance matrix (LEFT) and the image of this sample by the affine transformation (RIGHT):

$$T(x) = \Sigma^{1/2}x + a$$

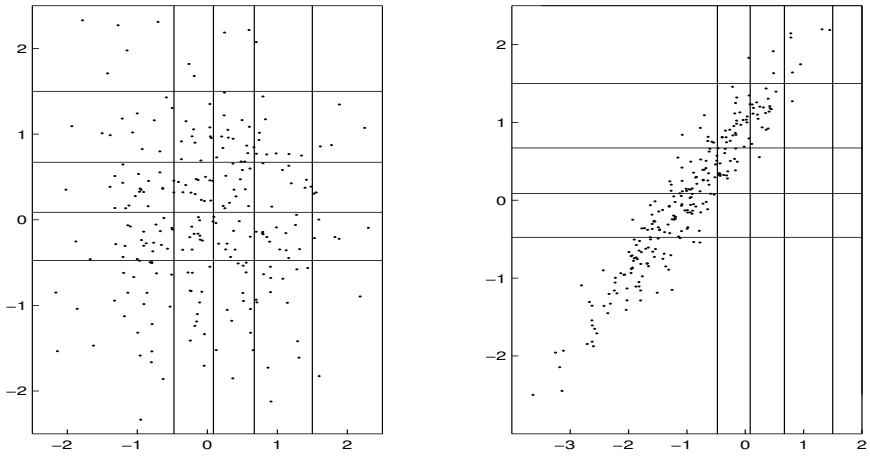

Figure 2.3: See text in Subsection 2.2.2.

where

$$\Sigma = \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix} \text{ and } a = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

Note that the transformed sample can be seen as a sample simulated from a binormal $\mathcal{N}(a, \Sigma)$. We represent on these graphics the partition used to construct regular modified histograms with a reference density of Gumbel (see (2.7)) and $\ell_1 = \ell_2 = 5$. For the transformed sample, only few classes possesses observations, the partition is not well adapted to the data cloud. Therefore the comparison between the empirical measure and the reference density over the partition does not make much sense.

To correct this, we now propose to construct data dependent modified histograms for which keeping the parameters g and ℓ_j $(j = 1, \ldots, d)$ fixed, the corresponding partition is equivariant under affine transformation of data vectors. Our method is inspired by the affine equivariant *quantile contour plots* defined by Chakraborty [10].

2.2.3 Data-driven modified histograms

Statistical practice suggests that histograms based on data-dependent partitions will provide better performance than those based on a fixed sequence of partitions. Theoretical evidence for this superiority was put forward by Stone [24].

In this paragraph, we construct a modified histogram based on a data-dependent partition equivariant under affine transformation of the data vector (for fixed g

and ℓ_j, $j = 1, \ldots, d$). The approach is based on a "data-driven coordinate system" introduced by Chaudhuri and Sengupta [12].

Formally, fix n_0 such that $n_0 > d+1$ and consider n_0+n data points $X_1, \ldots, X_{n_0+n}$ i.i.d. from a density f on $\mathbb{R}^d$. Split the data into a set $X_1, \ldots, X_{n_0}$ used for choosing the "data-driven coordinate system" and a set $X_{n_0+1}, \ldots, X_{n_0+n}$ used for constructing the density estimate. To lighten the notation we will write $X_1^\star, \ldots, X_n^\star$ for $X_{n_0+1}, \ldots, X_{n_0+n}$.

- Set $\ell = \ell_1 \ldots \ell_d$ with $\ell_1, \ldots, \ell_d$ positive integers, and let $h_j = 1/\ell_j$ for $j = 1, \ldots, d$;

- Let $\alpha = \{k_0, k_1, \ldots, k_d\}$ denote a subset of $\{1, 2, \ldots, n_0\}$ of size $(d+1)$. Consider the points $X_{k_0}, \ldots, X_{k_d}$ which will form a "data-driven coordinate system", where X_{k_0} will determine the origin and the lines joining that origin to the remaining d data points $X_{k_1}, \ldots, X_{k_d}$ will form various coordinate axis. Consider the $d \times d$ matrix

$$X(\alpha) = \{X_{k_1} - X_{k_0}, \ldots, X_{k_d} - X_{k_0}\}.$$

  If f is absolutely continuous on $\mathbb{R}^d$, $X(\alpha)$ is an invertible matrix with probability one for any choice of α (see Chaudhuri and Sengupta [12]). Next, transform all the observations in terms of the new coordinate system as

$$\begin{cases} \dot{X}_i = \{X(\alpha)\}^{-1} X_i, & i = 1, \ldots, n_0\,, \\ \\ \dot{X}_i^\star = \{X(\alpha)\}^{-1} X_i^\star, & i = 1, \ldots, n\,. \end{cases}$$

- Let $\tilde{g}$ be a density on $\mathbb{R}^d$ such that

$$\tilde{g}(x_1, \ldots, x_d) = \tilde{g}_1(x_1) \ldots \tilde{g}_d(x_d)\,, \tag{2.8}$$

  where $\tilde{g}_1(x_1), \ldots, \tilde{g}_d(x_d)$ are univariate densities. Define $p = (p_1, \ldots, p_d)$ the coordinatewise median associated with the density $\tilde{g}$, *i.e.*,

$$p_j = \tilde{G}_j^{-1}(0.5)\,, \quad j = 1, \ldots, d\,,$$

  and let $\dot{X}_{([n_0/2])}$ be the empirical coordinatewise median from the sample $\dot{X}_1, \ldots, \dot{X}_{n_0}$, *i.e.*,

$$\dot{X}_{([n_0/2])} = (\dot{X}^{(1)}_{([n_0/2])}, \ldots, \dot{X}^{(d)}_{([n_0/2])})\,.$$

  where $[\ \ ]$ stands for the integer part and $\dot{X}^{(j)}_{(1)}, \ldots, \dot{X}^{(j)}_{(n_0)}$ denotes the order statistics of the j-th components of the data vector $\dot{X}_1, \ldots, \dot{X}_{n_0}$. Consider

the vector $b_X^\alpha = p - \dot{X}_{([n_0/2])}$ and let $\tilde{X}_i^\star$ be the image of $\dot{X}_i^\star$ by the translation of vector b_X^α, *i.e.*,

$$\tilde{X}_i^\star = \dot{X}_i^\star + b_X^\alpha\,, \quad i = 1, \ldots, n\,.$$

As for the regular modified histograms presented above, for $j = 1, \ldots, d$ and $i_j = 1, \ldots, \ell_j - 1$, denote $\tilde{q}_{j,i_j}$ the quantile of order $i_j h_j$ of $\tilde{g}_j$. These quantiles lead to a partition of $\mathbb{R}^d$ into ℓ hyperrectangles say

$$\tilde{A}_{i_1,\ldots,i_d} = \prod_{j=1}^{d} (\tilde{q}_{j,i_j-1}, \tilde{q}_{j,i_j}]\,.$$

Let μ_n (resp. $\tilde{\mu}_n$) be the empirical measure associated with the sample $X_1^\star, \ldots, X_n^\star$ (resp. $\tilde{X}_1^\star, \ldots, \tilde{X}_n^\star$);

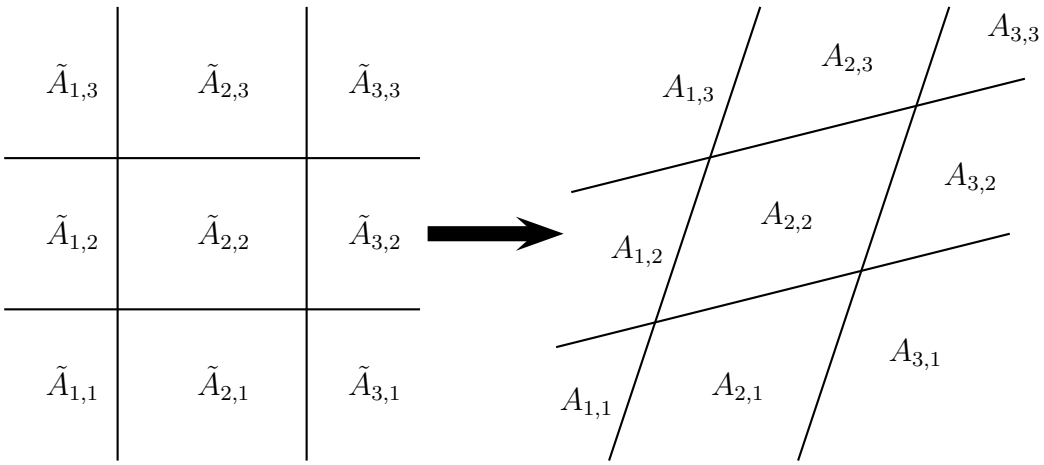

Figure 2.4: Transformation of a partition in $\mathbb{R}^2$.

- Express the $\tilde{A}_{i_1,\ldots,i_d}$'s in terms of the original coordinate system, *i.e.*,

$$A_{i_1,\ldots,i_d} = X(\alpha)\big(\tilde{A}_{i_1,\ldots,i_d} - b_X^\alpha\big)\,.$$

$\tilde{A}_{i_1,\ldots,i_d}$ is the image of the hyperrectangle $A_{i_1,\ldots,i_d}$ by an affine transformation therefore $\tilde{A}_{i_1,\ldots,i_d}$ is an hyperparallelogram (see Figure 2.4). Moreover, it is easily seen that

$$\mu_n(A_{i_1,\ldots,i_d}) = \tilde{\mu}_n(\tilde{A}_{i_1,\ldots,i_d})\,;$$

- Finally, fix

$$g_\alpha(x) = \frac{1}{\left|det\big(X(\alpha)\big)\right|}\, \tilde{g}\big(\{X(\alpha)\}^{-1}x + b_X^\alpha\big)\,, \tag{2.9}$$

then the *data-driven modified histogram density estimate* is defined by

$$f_n(x) = \frac{n\mu_n\big(A(x)\big)+1}{nh+1}\, g_\alpha(x)\,, \tag{2.10}$$

where $h = h_1 \dots h_d$ and $A(x)$ stands for $A_{i_1,\dots,i_d}$ if $x \in A_{i_1,\dots,i_d}$.

Lemma 2.2.1 *The estimate $f_n(x)$ defined in* (2.10) *is a modified histogram in the sense of* (2.1).

Proof It suffices to prove the following assertions:

- g_α is a density (we will denote by ν_{g_α} the measure associated with this density);
- for $j = 1, \dots, d$ and $i_j = 1, \dots, m_j$, $\nu_{g_\alpha}(A_{i_1,\dots,i_d}) = h$.

These assertions are direct consequences of the change of variables theorem. ■

Remark 2.2.1 One can use other translations, however our choice of b_X^α seems to be well adapted to our estimate. Indeed, modified histograms result from the comparison between the reference density and the empirical measure. Thus, our translation is chosen so that the image of $\dot{X}_1, \dots, \dot{X}_{n_0}$ has the same median as the density $\tilde{g}$. This translation can be seen as a "bias correction". We choose the median because of its robustness.

From now on, given random variables $T_1, \dots, T_{n_0+n}$, we write $T_1^\star, \dots, T_n^\star$ for $T_{n_0+1}, \dots, T_{n_0+n}$ and $\mu_n(A; T_1^\star, \dots, T_n^\star)$ for the empirical measure associated with $T_1^\star, \dots, T_n^\star$. Moreover, with a slight abuse of notation, we will denote by $\{A_{i_1,\dots,i_d}\}$ the partition

$$\{A_{i_1,\dots,i_d}\}_{\substack{1\le j\le d\\ 1\le i_j\le \ell_j}}\,.$$

We now prove the equivariance of the partition under arbitrary affine transformations of data vectors.

Theorem 2.2.2 *The partition $\{A_{i_1,\dots,i_d}\}$ is equivariant under arbitrary affine transformations of data vectors. We can formulate it as follows.*
Let the d-dimensional vectors $X_1, \dots, X_{n_0+n}$ be transformed into $Z_1, \dots, Z_{n_0+n}$ with $Z_i = MX_i + c$ where M is a $d \times d$ nonsingular matrix and c is a vector in $\mathbb{R}^d$. Suppose that we use the same density $\tilde{g}$ and the same integers ℓ_j

($j = 1, \ldots, d$) to construct the data-driven modified histogram from $X_1, \ldots, X_{n_0+n}$ and $Z_1, \ldots, Z_{n_0+n}$. If $\{A_{i_1,\ldots,i_d}\}$ (resp. $\{B_{i_1,\ldots,i_d}\}$) denotes the partition computed from the sample $X_1, \ldots, X_{n_0+n}$ (resp. $Z_1, \ldots, Z_{n_0+n}$), then for all integers $i_1, \ldots, i_d$ such that $1 \leq i_j \leq \ell_j$ and $1 \leq j \leq d$ we have

(i) $B_{i_1,\ldots,i_d} = MA_{i_1,\ldots,i_d} + c\,.$

(ii) $\mu_n(A_{i_1,\ldots,i_d}; X_1^\star, \ldots, X_n^\star) = \mu_n(B_{i_1,\ldots,i_d}; Z_1^\star, \ldots, Z_n^\star)\,.$

Proof Let $\alpha = \{k_0, k_1, \ldots, k_d\}$ be a subset of $\{1, \ldots, n_0\}$ of size $d+1$. Consider

$$X(\alpha) = \{X_{k_1} - X_{k_0}, \ldots, X_{k_d} - X_{k_0}\}$$

and

$$Z(\alpha) = \{Z_{k_1} - Z_{k_0}, \ldots, Z_{k_d} - Z_{k_0}\}$$

so that we have $Z(\alpha) = MX(\alpha)$.
Note that for $i = 1, \ldots, n_0$

$$\begin{aligned}\dot{Z}_i =& Z(\alpha)^{-1} Z_i \\ =& X(\alpha)^{-1} X_i + \big(MX(\alpha)\big)^{-1} c \\ =& \dot{X}_i + \big(MX(\alpha)\big)^{-1} c\,.\end{aligned}$$

Therefore $\dot{Z}_{[n_0/2]} = \dot{X}_{[n_0/2]} + (MX(\alpha))^{-1} c$ and $b_Z^\alpha = b_X^\alpha - \big(MX(\alpha)\big)^{-1} c$.

As we use the same density $\tilde{g}$ and the same integers ℓ_j ($j = 1, \ldots, d$), the partitions computed for transformed observations will be the same for the samples $X_1, \ldots, X_{n_0+n}$ and $Z_1, \ldots, Z_{n_0+n}$. We denote $\{\tilde{A}_{i_1,\ldots,i_d}\}$ this partition.

To compute $\{A_{i_1,\ldots,i_d}\}$ and $\{B_{i_1,\ldots,i_d}\}$ we only have to retransform $\{\tilde{A}_{i_1,\ldots,i_d}\}$. For all integers $i_1, \ldots, i_d$ such that $1 \leq i_j \leq \ell_j$ and $1 \leq j \leq d$, it follows that

$$\begin{aligned}B_{i_1,\ldots,i_d} =& Z(\alpha)(\tilde{A}_{i_1,\ldots,i_d} - b_Z^\alpha) \\ =& MX(\alpha)\Big(\tilde{A}_{i_1,\ldots,i_d} - \big(b_X^\alpha - (MX(\alpha))^{-1} c\big)\Big) \\ =& M\Big(X(\alpha)\big(\tilde{A}_{i_1,\ldots,i_d} - b_X^\alpha\big)\Big) + c \\ =& MA_{i_1,\ldots,i_d} + c\,,\end{aligned}$$

which gives *(i)*.

Since $\tilde{X}_i^\star = X(\alpha)^{-1} X_i^\star + b_X^\alpha$ and $\tilde{Z}_i^\star = Z(\alpha)^{-1} Z_i^\star + b_Z^\alpha$ ($i = 1, \ldots, n$), it easily follows that

$$\forall i = 1, \ldots, n,\ \tilde{X}_i^\star = \tilde{Z}_i^\star\,.$$

Therefore

$$\mu_n(\tilde{A}_{i_1,\ldots,i_d}; \tilde{X}_1^\star, \ldots, \tilde{X}_n^\star) = \mu_n(\tilde{A}_{i_1,\ldots,i_d}; \tilde{Z}_1^\star, \ldots, \tilde{Z}_n^\star)$$

and *(ii)* is proved. ■

Note that *(ii)* implies that we only have to compute the empirical measure of the rectangles $\tilde{A}_{i_1,\ldots,i_d}$ associated with the transformed observations, a much simpler device than the empirical measure of the hyperparallelogram $A_{i_1,\ldots,i_d}$.

It is worth pointing out that the actual reference density (in the sense of (2.1)) is g_α which implies that the reference density depends on the data. However, in practice we can have some a priori idea on the density to estimate. In that case, we could be interested in constructing modified histogram for a particular reference density g. It is possible to use this algorithm provided that g may be written in the form

$$g(x) = \frac{1}{|\det(M)|}\tilde{g}(M^{-1}x + a) \tag{2.11}$$

where $x = (x_1, \ldots, x_d)$, M is an invertible matrix $d \times d$, a is a vector of $\mathbb{R}^d$ and $\tilde{g}$ is a product of univariate densities, *i.e.*,

$$\tilde{g}(x) = \tilde{g}_1(x_1) \ldots \tilde{g}_d(x_d)\,.$$

In this situation we no longer split the data (all the observations are used to construct the modified histogram) and we only have to replace $X(\alpha)$ by M and b_X^α by a. Note that multinormal densities $g_{\mu,\Sigma}$ are in the form of (2.11). Nevertheless Theorem 2.2.2 does not hold for such modified histograms.

Summarizing, we have found a partition equivariant under arbitrary affine transformation. Consistency in information divergence of the corresponding estimate is a straightforward consequence of the next lemma (whom proof is straightforward).

Lemma 2.2.2 *Information divergence is invariant under invertible transformation of the data sample.*

Corollary 2.2.1 *Let f_n be the data-driven modified histogram defined in* (2.10). *Assume that $D(f, g_\alpha) < \infty$ a.s. Moreover, assume that for $i = 1, \ldots, d$, $h_i = h_{i,n}$ (therefore $h = h_n$),*

$$\lim_{n\to\infty} \max_{1\le i\le d} h_{i,n} = 0 \text{ and } \lim_{n\to\infty} nh_n = \infty\,,$$

then

$$\lim_{n\to\infty} D(f, f_n) = 0 \text{ a.s.} \quad \text{and} \quad \lim_{n\to\infty} \mathbf{E}^{(n_0)} D(f, f_n) = 0 \text{ a.s.}$$

where $\mathbf{E}^{(n_0)}$ denotes the conditional expectation given the X_i's for which $1 \le i \le n_0$.

Proof Fix $X_1, \ldots, X_{n_0}$ such that $D(f, g_\alpha) < \infty$. Let $\tilde{f}$ (resp. $\tilde{f}_n$) be the target density f (resp. the density estimate f_n) in the transformed coordinate system, *i.e.*,

$$\begin{cases} \tilde{f}(x) = & \left|det\big(X(\alpha)\big)\right| f\big(X(\alpha)(x - b_X^\alpha)\big) \\ \\ \tilde{f}_n(x) = & \left|det\big(X(\alpha)\big)\right| f_n\big(X(\alpha)(x - b_X^\alpha)\big)\,. \end{cases}$$

$\tilde{f}_n$ is the regular modified histogram density estimate of $\tilde{f}$ (see page 54) using $\tilde{g}$ as reference density. From Theorem 2.2.1, it follows that

$$\begin{cases} \lim\limits_{n\to\infty} D(\tilde{f}, \tilde{f}_n) = 0 \; a.s. \\ \\ \lim\limits_{n\to\infty} \mathbf{E}^{(n_0)} D(\tilde{f}, \tilde{f}_n) = 0\,. \end{cases}$$

The conclusion follows from Lemma 2.2.2. ■

2.3 Selection of α

The performance of the data-driven modified histogram clearly depends upon the choice of ℓ_j $(j = 1, \ldots, d)$, $\tilde{g}$ and α. Here we will restrict our attention to the choice of α. Recent univariate results obtained by Berlinet and Brunel [5] show that the Kullback-Leibler cross-validation technique works well for selecting ℓ_1 from the data. We extend this procedure to the selection of α.

Let $X_1, \ldots, X_{n_0+n}$ be i.i.d. observations from a density f and let S_{n_0} denote the collection of all subsets of size $d+1$ of $\{1, \ldots, n_0\}$. Fix ℓ_j $(j = 1, \ldots, d)$ and $\tilde{g}$ (such as (2.8)). For $\alpha \in S_{n_0}$, let us denote by f_n^α the data-driven modified multivariate histogram defined in (2.10).

Expanding the actual information divergence error yields

$$D(f, f_n^\alpha) = \int f \log f - \int f \log f_n^\alpha\,. \tag{2.12}$$

The second integral could be written as $\mathbf{E}[\log f_n^\alpha(X)]$, where the expectation is taken with respect to the evaluating point and not over the sample. The cross-validation device consists in removing one data point among $X_1^\star, \ldots, X_n^\star$ and using the remaining $(n-1)$ points to construct an estimator of $\mathbf{E}(\log f_n^\alpha(X))$. This step is repeated for each $X_i^\star$ $(i = 1, \ldots, n)$. Let $f_n^{\alpha,i}$ be the modified histogram density estimate defined after deleting the i-th observation, *i.e.*,

$$f_n^{\alpha,i}(x) = \frac{n\mu_n^i\big(A(x); X_1^\star, \ldots, X_n^\star\big) + 1}{nh + 1} g_\alpha(x)$$

where g_α is defined by (2.9) and

$$\mu_n^i\big(A(x); X_1^\star, \ldots, X_n^\star\big) = \frac{1}{n-1}\sum_{j\neq i} \mathbf{1}_{\{X_j^\star \in A(x)\}}\,.$$

With this notation, an estimate of $\mathbf{E}[\log f_n^\alpha(X)]$ is given by

$$\frac{1}{n}\sum_{i=1}^n \log f_n^{\alpha,i}(X_i^\star)$$

and since the first integral in (2.12) does not depend on α, we deduce a cross-validation criterion for the choice of α:

choose $\hat{\alpha} \in S_{n_0}$ which minimizes $CV(\alpha) = -\dfrac{1}{n}\displaystyle\sum_{i=1}^n \log\, f_n^{\alpha,i}(X_i^\star)$.

For fixed α, we have proved that the partition is affine equivariant. The next theorem states that this property still holds when α is selected by cross-validation.

Theorem 2.3.1 *The choice of α by cross-validation is invariant under arbitrary affine transformations of data vectors. We can formulate it as follows.*
Let the d-dimensional vectors $X_1, \ldots, X_{n_0+n}$ be transformed into $Z_1, \ldots, Z_{n_0+n}$ with $Z_i = MX_i + c$ where M is a $d\times d$ nonsingular matrix and c is a vector in $\mathbb{R}^d$. Suppose that we use the same density $\tilde{g}$ and the same integers ℓ_j $(j = 1, \ldots, d)$ to construct the data-driven modified histograms $f_{n,X}^\alpha$ (with $X_1, \ldots, X_{n_0+n}$) and $f_{n,Z}^\alpha$ (with $Z_1, \ldots, Z_{n_0+n}$). Then

$$\hat{\alpha} \text{ minimizes } -\frac{1}{n}\sum_{i=1}^n \log f_{n,X}^{\alpha,i}(X_i^\star) \;\Leftrightarrow\; \hat{\alpha} \text{ minimizes } -\frac{1}{n}\sum_{i=1}^n \log f_{n,Z}^{\alpha,i}(Z_i^\star).$$

Proof We will denote by $\{A_{i_1,\ldots,i_d}\}$ (resp. $\{B_{i_1,\ldots,i_d}\}$) the partition associated with the modified histogram constructed from the sample $X_1, \ldots, X_{n_0+n}$ (resp. $Z_1, \ldots, Z_{n_0+n}$).
We have

$$\frac{1}{n}\sum_{i=1}^n \log f_{n,Z}^{\alpha,i}(Z_i^\star) = \frac{1}{n}\sum_{i=1}^n \log \frac{n\mu_n^i\big(B(Z_i^\star); Z_1^\star, \ldots, Z_n^\star\big)+1}{nh+1} g_\alpha(Z_i^\star)$$

where

$$g_\alpha(Z_i^\star) = \frac{1}{|det(Z(\alpha))|}\tilde{g}\big(Z(\alpha)^{-1}Z_i^\star + b_Z^\alpha\big)\,.$$

Theorem 2.2.2 and its proof give

$$\begin{cases} Z(\alpha) &= MX(\alpha) \\ b_Z^\alpha &= b_X^\alpha - \big(MX(\alpha)\big)^{-1} c \\ B_{i_1,\ldots,i_d} &= MA_{i_1,\ldots,i_d} + c\,. \end{cases}$$

Moreover, it is easily seen that $\mu_n^i\big(B(Z_i^\star); Z_1^\star, \ldots, Z_n^\star\big) = \mu_n^i\big(A(X_i^\star); X_1^\star, \ldots, X_n^\star\big)$. Putting all pieces together, we obtain

$$-\frac{1}{n}\sum_{i=1}^{n} \log f_{n,Z}^{\alpha,i}(Z_i^\star) = -\frac{1}{n}\sum_{i=1}^{n} \log f_{n,X}^{\alpha,i}(X_i^\star) + \log(|\det(M)|)\,.$$

Since M does not depend on α, the proof is complete. ■

2.4 Simulations

In this paragraph we are presenting some finite sample simulation results on the efficiency of the data-driven modified histogram f_n^α defined by (2.10) compared with the regular modified histogram f_n defined by (2.6). We use two data sets.

We first simulated 50 samples of size n_0+n ($n_0+n = 150,\ 300,\ 550$) from bivariate normal populations with zero means, unit standard deviations and varying correlation coefficients $\rho = 0,\ 0.25,\ 0.5,\ 0.75$ and 0.95. For each sample, we have computed modified histograms f_n and f_n^α (α is selected by cross-validation). We choose

$$\begin{cases} \tilde{g}(x,y) = \exp\big(-x-\exp(-x)\big)\exp\big(-y-\exp(-y)\big) \\ n_0 = 50 \\ \ell_1 = \ell_2 = 4 \text{ for } n = 100 \\ \ell_1 = \ell_2 = 5 \text{ for } n = 250 \\ \ell_1 = \ell_2 = 6 \text{ for } n = 500\,. \end{cases}$$

We display in Table 2.2 the average of $D(f, f_n)$ and $D(f, f_n^\alpha)$ and the gain Ga in information divergence

$$Ga = \frac{D(f,f_n) - D(f,f_n^\alpha)}{D(f,f_n)}.$$

For the second data set, points are generated from multivariate symmetric Laplace distributions (see Anderson [1]) with density

$$f(x) = \frac{2}{(2\pi)^{d/2}|\Sigma|^{1/2}}(x^t\Sigma^{-1}x/2)^{v/2}K_v\Big(\sqrt{2x^t\Sigma^{-1}x}\Big)\,,$$

n	ρ	$D(f, f_n)$	$D(f, f_n^\alpha)$	Ga
100	0	0.36	0.23	0.36
250		0.32	0.15	0.53
500		0.29	0.11	0.62
100	0.25	0.36	0.24	0.33
250		0.32	0.15	0.53
500		0.30	0.11	0.63
100	0.5	0.39	0.23	0.41
250		0.35	0.15	0.57
500		0.31	0.11	0.65
100	0.75	0.50	0.23	0.54
250		0.41	0.15	0.63
500		0.36	0.10	0.72
100	0.95	0.85	0.24	0.72
250		0.73	0.14	0.81
500		0.64	0.11	0.83

Table 2.2: Regular modified histograms versus data-driven modified histograms.

where $v = (2-d)/2$, Σ is a $d \times d$ non-negative definite symmetric matrix and $K_v(u)$ is the modified Bessel function of the third kind given by

$$K_v(u) = \frac{1}{2}\Big(\frac{u}{2}\Big)^v \int_0^\infty t^{-v-1} \exp\Big(-t - \frac{u^2}{4t}\Big)\, dt\,,\ u > 0\,.$$

We set several dimensions $d = 2, 4, 8, 10$ and several sample sizes $n_0 + n$. For each $(d, n_0 + n)$, we simulated 50 samples from a symmetric Laplace distribution with

$$\Sigma = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \dots & \rho & 1 \end{pmatrix} \qquad \rho = 0; 0.5; 0.95\,.$$

The density $\tilde{g}$ is a multivariate standard normal distribution and we choose $\ell_j = 3$ $(j = 1, \dots, d)$. For $d = 2$ and 4 we again take $n_0 = 50$ to select α by cross-validation.

However, for higher dimensions the optimization problem is very heavy and takes too much time to reach an adequate solution. Thus, for $d = 8$ and 10, we propose the following alternative. We choose the transformation matrix $X(\alpha)$ in such a

way that the image of $X_1, \ldots, X_{n_0}$ has the same variance-covariance matrix as the density $\tilde{g}$ (identity in our example). In other words, we replace $X(\alpha)$ with $\hat{\Sigma}^{1/2}$ where $\hat{\Sigma}$ is an affine equivariant estimate of the variance-covariance matrix of the distribution (computed from $X_1, \ldots, X_{n_0}$). The rest of the construction does not change. Note that the corresponding estimate no longer depends on α but on $\hat{\Sigma}$. In this regard it will be denoted by $f_n^{\hat{\Sigma}}$ and for the sake of clarity the associated reference density g_α and vector b_X^α will be denoted by $g_{\hat{\Sigma}}$ and $b_X^{\hat{\Sigma}}$. Consistency Corollary 2.2.1 is still true for $f_n^{\hat{\Sigma}}$. We take $n_0 = 1000$ for $d = 8$ and 10. $D(f, f_n)$ and $D(f, f_n^\alpha)$ are computed from Monte-Carlo method. The results are displayed in Table 2.3.

2.5 Concluding Remarks

1. Our examples demonstrate rather strikingly that f_n^α is on the whole better than f_n. The difference increases with the correlation and the dimension. Moreover, keeping n and d fixed, $D(f, f_n^\alpha)$ is stable whatever the correlation. It is worth pointing out that the partition is well adapted to the data cloud even with high correlation (see Figure 2.5).

2. In contrast with the first data set, performances of f_n^α and f_n are similar when $\rho = 0$ for the second set. It is probably due to the fact that the symmetric Laplace distribution and the standard gaussian distribution have the same median. Therefore the translation vector is close to zero and the reference density and the density to estimate are close enough without the translation. On the other hand, for the first data set, the two distributions do not have the same coordinatewise median. The translation can be seen as a "bias corrector" between the two densities.

3. For the second data set, the partition is not equivariant by affine transformation of the data sample when $d = 8$ or $d = 10$. All the same, the estimate is performant and the computation is quick even in large dimension. We emphasize that the transformation-retransformation procedure just allows to select a reference density which is not "too far" from the density to estimate. In other words our choice of the transformation matrix is motivated by the fact that the reference density should be as close as possible to the density f to estimate. Since f is unknown in practice, we select the affine transformation such that the image of $X_1, \ldots, X_{n_0}$ and the random variable with density $\tilde{g}$ have

    - the same variance-covariance matrix (with the help of the linear transformation $\hat{\Sigma}^{-1/2}$);

$d;n$	ρ	$D(f,f_n)$	$D(f,f_n^{\alpha})$ or $D(f,f_n^{\hat{\Sigma}})$	Ga
2;250	0	0.12	0.12	0
	0.5	0.18	0.13	0.28
	0.95	0.73	0.12	0.84
4;1000	0	0.34	0.36	-0.06
	0.5	0.55	0.37	0.33
	0.95	2.32	0.37	0.84
8;10000	0	0.90	0.90	0
	0.5	1.58	0.93	0.41
	0.95	6.10	0.93	0.85
10;500000	0	1.08	1.12	-0.04
	0.5	1.90	1.14	0.40
	0.95	7.52	1.12	0.85

Table 2.3: Regular modified histograms versus data-driven modified histograms.

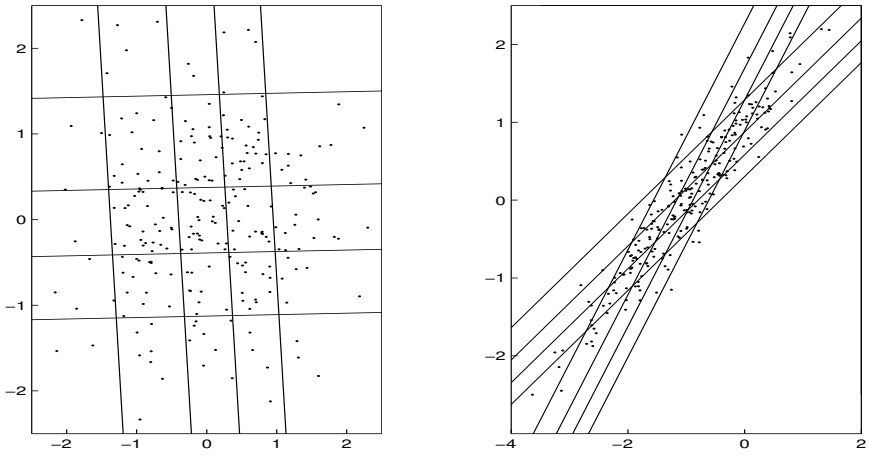

Figure 2.5: Partition of data-driven modified histogram. Simulated samples are the same as in Figure 2.3. $\tilde{g}$ is a product of Gumbel densities and $\ell_1 = \ell_2 = 5$.

- the same median (with the translation $b_X^{\hat{\Sigma}}$).

Note that when f is elliptically symmetric, similar conditions on the choice of the transformation matrix are discussed by Chakraborty [10] in the asymptotic study of the affine equivariant multivariate quantiles.

Bibliography

[1] D.N. Anderson. A multivariate linnik distribution. *Statistic and Probability Letters*, 14:333–336, 1992.

[2] A.R. Barron. The convergence in information of probability density estimators. In *Proceedings of the International Symposium of IEEE on Information Theory*, Kobe : Japan, June 19-24 1988.

[3] A.R. Barron, L. Györfi, and E.C. van der Meulen. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Transaction on Information Theory*, 38:1437–1454, 1992.

[4] A. Berlinet and G. Biau. Iterated modified hitograms as dynamical systems. *Journal of Nonparametric Statistics*, 16:385–401, 2004.

[5] A. Berlinet and E. Brunel. Cross-validated density estimates based on Kullback-Leibler information. *Journal of Nonparametric Statistics*, 16:493–513, 2004.

[6] A. Berlinet, L. Györfi, and E.C. van der Meulen. The asymptotic normality of relative entropy in multivariate density estimation. *Publication de l'Institut de Statistique de l'Université de Paris*, 41:3–27, 1997.

[7] A. Berlinet, I. Vajda, and E.C. van der Meulen. About the asymptotic accuracy of barron density estimates. *IEEE Transactions on Information Theory*, 38:1437–1454, 1998.

[8] B.M. Brown. Statistical use of the spatial median. *Journal of the Royal Statistical Society, Series B*, 45:25–30, 1983.

[9] B.M. Brown and T.P. Hettmansperger. Affine invariant rank methods in the bivariate location model. *Journal of the Royal Statistical Society, Series B*, 49:301–310, 1987.

[10] B. Chakraborty. On affine equivariant multivariate quantiles. *The Institute of Statistical Mathematics*, 53:380–403, 2001.

[11] P. Chaudhuri. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91:862–872, 1996.

[12] P. Chaudhuri and D. Sengupta. Sign tests in multidimension : Inference based on the geometry of the data cloud. *Journal of the American Statistical Association*, 88:1363–1370, 1993.

[13] I. Csizár. Information-type measures of divergence of probability distributions and indirect observations. *Studia Sci. Math. Hungar*, 2:299–318, 1967.

[14] I. Csizár. Generalized entropy and quantization problems. In *Trans. Sixth Prague Conf. Information Theory, Statistical Decision Functions, Random Process*, pages 159–174, Prague : Academia, 1973.

[15] L. Devroye. On arbitrary slow rates of global convergence in density estimation. *Z. f. W.*, 62:475–483, 1983.

[16] W.F. Eddy. Set valued ordering of bivariate data. In *Stochastic Geometry, Geometric Statistics, and Stereology*, pages 79–90. R.V. Ambartsumian and W. Weil, Leipzig, 1983.

[17] W.F. Eddy. Ordering of multivariate data. In *Computer Science and Statistics : The Interface*, pages 25–30. L. Billard, Amsterdam : North-Holland, 1985.

[18] L. Györfi, F. Liese, I. Vajda, and E.C. van der Meulen. Distribution estimates consistent in χ^2-divergence. *Statistics*, 32:31–57, 1998.

[19] P. Hall. On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, 15:1491–1519, 1987.

[20] J.H.B. Kemperman. On the optimum rate of transmitting information. *The Annals of Mathematical Statistics*, 40:2156–2177, 1969.

[21] J.H.B. Kemperman. The median of a finite measure on a Banach space. In *Statistical Data Analysis Based on L_1 norm and Related Methods*, pages 217–230. Y. Dodge, Amsterdam North-Holland, 1987.

[22] S. Kullback. A lower bound for discrimination in terms of variation. *IEEE Transactions on Information Theory*, 13:126–127, 1967.

[23] R.Y. Liu, J.M. Parelius, and K. Singh. Multivariate analysis by data depth : descriptive statistics, graphics and inference (with discussion). *The Annals of Statistics*, 18:783–858, 1999.

[24] C.J. Stone. An asymptotically optimal histogram selection rule. In L. Le Cam and R. A. Olshen, editors, *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer*, volume 2, pages 513–520, Wadsworth, Belmont, CA, 1985.

[25] J.W. Tuckey. Mathematics and picturing data. In *Proc. Intern. Congr. Math*, volume 2, pages 523–531, Vancouver 1974, 1975.

Deuxième partie

Méthodes Combinatoires en Estimation de la Densité

Chapter 1

Parameter Selection in Modified Histogram Estimates*

Abstract

A multivariate modified histogram density estimate depending on a reference density g and a partition P has recently been proved to have good consistency properties according to several information theoretic criteria. Given an i.i.d. sample, we show how to select automatically both g and P so that the expected L_1 error of the corresponding selected estimate is within a given constant multiple of the best possible error plus an additive term which tends to zero under mild assumptions. Our method is inspired by the combinatorial tools developed in Devroye and Lugosi [9] and it includes a wide range of reference density and partition models. Results of simulations are presented.

1.1 Introduction

General ϕ-divergences (Liese and Vajda [12]) are widely used in many fields of statistics (data compression, telecommunication networks, classification, pattern recognition, neural networks...), particularly in decision processes based on density estimates and functionals of them. Many authors have put forward their attractive properties as criteria of accuracy. However, considering convergence of estimates of a density in the sense of ϕ-divergences causes some trouble. With standard histograms the situation is even hopeless as empty cells, occurring with high probability, make most of divergences infinite. The *modified histograms* in-

*Article écrit en collaboration avec Alain Berlinet et Gérard Biau, et publié dans la revue *Statistics*.

troduced by Barron [1] and Barron, Györfi and van der Meulen [2] circumvent this problem. They are defined as follows.

Suppose that we observe independent $\mathbb{R}^d$-valued random variables $X_1, \ldots, X_n$ with common unknown density f.

- Denote by g a known density on $\mathbb{R}^d$ and by ν_g the associated probability measure;
- Define a sequence of integers $\{\ell_n\}_{n\geq 1}$ such that $2 \leq \ell_n$ and let $h_n = 1/\ell_n$;
- Introduce a sequence of partitions $P = \{A_{n1}, \ldots, A_{n\ell_n}\}$ such that $\nu_g(A_{ni}) = h_n$ for $i = 1, \ldots, \ell_n$;
- Finally consider, for $a_n = 1/(nh_n + 1)$, the following density estimate f_n:

$$f_n(x) = \left[(1-a_n)\frac{\mu_n\big(A_n(x)\big)}{h_n} + a_n\right] g(x) = \frac{n\mu_n\big(A_n(x)\big)+1}{nh_n+1}\, g(x)\,, \quad (1.1)$$

  where μ_n stands for the empirical measure associated with the sample $X_1, \ldots, X_n$, *i.e.*, $\mu_n(A) = (1/n)\sum_{i=1}^n \mathbf{1}_{[X_i \in A]}$, and $A_n(x)$ equals A_{ni} if $x \in A_{ni}$.

The estimate (1.1) is a mixture of a histogram-type density estimate and the known density g. It can also be regarded as a piecewise transformation of g itself: roughly speaking, this modified histogram results from the comparison of the quantiles of g – the *reference density* – with the empirical quantiles (see Figure 1.1 for an example).

For further results on modified histograms, we refer the reader to Barron, Györfi and van der Meulen [2] who prove consistency in the sense of information divergence, Berlinet, Györfi and van der Meulen [6] who prove a central limit theorem for Kullback-Leibler divergence, Györfi, Liese, Vajda and van der Meulen [10], and Berlinet, Vajda and van der Meulen [7] who extend the information divergence consistency properties respectively to the χ^2-divergence and to more general ϕ-divergences.

Once the observations are given two parameters have to be chosen to build the modified histogram, namely a reference density g and a partition P. Recent univariate results obtained by Berlinet and Brunel (see [4], [5]) show that the Kullback-Leibler cross-validation technique works well to select the partition from the data and that it is asymptotically optimal. As far as we know, no work has been devoted so far to select g and P simultaneously. This article proposes to fill this gap, using a general multivariate data-based combinatorial methodology

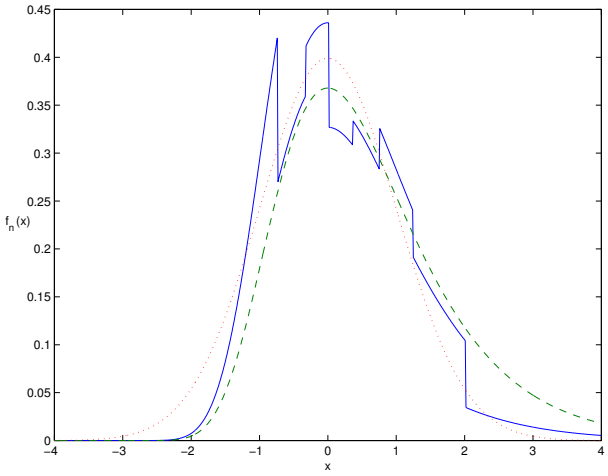

Figure 1.1: Modified histogram estimate (continuous line) of $\mathbf{n = 100}$ Gaussian (dotted line) data ($\ell_{\mathbf{n}} = \mathbf{8}$). The reference density is Gumbel (dashed line).

presented in Devroye and Lugosi [9]. More precisely, we will show how to select both g and P – within given classes – so that the expected L_1 error of the corresponding selected estimate is up to a given constant multiple of the best possible error plus an additive term which tends to zero under mild assumptions. The paper is organized as follows. In Section 2, we present the multivariate selection procedure and give the main results. Examples are worked out in Section 3 and simulations are presented in Section 4. Proofs are gathered in Section 5.

1.2 Automatic parameter selection

1.2.1 The combinatorial method

Using ideas from Yatracos [16], Devroye and Lugosi [9] explore a new paradigm for the data-based or automatic selection of the free parameters of density estimates in general so that the expected L_1 error is within a given constant multiple of the best possible error. To summarize in the present context, assume we are given a class of density estimates parameterized by $\theta \in \Theta$ such that $f_{n,\theta}$ denotes the density estimate with parameter θ. Let $m < n$ be an integer which splits the data $X_1, \ldots, X_n$ into

- a set $X_1, \ldots, X_{n-m}$ used for the construction of the density estimate;
- a validation set $X_{n-m+1}, \ldots, X_n$.

Introduce the class of random sets

$$\mathcal{A}_\Theta = \Big\{ \big\{ x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x) \big\} : (\theta, \theta') \in \Theta^2 \Big\}$$

($\mathcal{A}_\Theta$ is the so-called *Yatracos class* associated with Θ) and define

$$\Delta_\theta = \sup_{A \in \mathcal{A}_\Theta} \left| \int_A f_{n-m,\theta} - \mu_m(A) \right|,$$

where $\mu_m(A) = (1/m) \sum_{i=n-m+1}^{n} \mathbf{1}_{[X_i \in A]}$ is the empirical measure associated with the sample $X_{n-m+1}, \ldots, X_n$. Then the *minimum distance estimate* f_n is defined as any density estimate selected among the candidates $f_{n-m,\theta}$ with

$$\Delta_\theta < \inf_{\theta^* \in \Theta} \Delta_{\theta^*} + \frac{1}{n}.$$

Note that the $1/n$ term is added to ensure the existence of such a density estimate. According to Devroye and Lugosi [9], Chapter 10, whenever $f_{n-m,\theta}$ integrates to one, the selected f_n satisfies the following inequality:

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \leq 3 \inf_{\theta \in \Theta} \mathbf{E}\left\{ \int |f_{n-m,\theta} - f| \right\} + 8\, \mathbf{E}\left\{ \sqrt{\frac{\log 2\, \mathbf{S}_{\mathcal{A}_\Theta}(m)}{m}} \right\} + \frac{3}{n}. \quad (1.2)$$

Here, $\mathbf{S}_{\mathcal{A}_\Theta}(m)$ is the *Vapnik-Chervonenkis shatter coefficient* of the class of sets $\mathcal{A}_\Theta$ (Vapnik and Chervonenkis [15]), defined by

$$\mathbf{S}_{\mathcal{A}_\Theta}(m) = \max_{x_1, \ldots, x_m \in \mathbb{R}^d} \mathrm{Card}\big\{ \{x_1, \ldots, x_m\} \cap A : A \in \mathcal{A}_\Theta \big\}.$$

This general methodology provides an automatic procedure to construct a density estimate f_n whose L_1 error is (almost) as small as that of the best estimate among the $f_{n,\theta}$, $\theta \in \Theta$. We emphasize that inequality (1.2) is nonasymptotic, that is, the bound is valid for all n. The rest of the analysis is then purely combinatorial and merely consists in obtaining upper bounds for the value of $\mathbf{S}_{\mathcal{A}_\Theta}(m)$.

A challenging question is whether the combinatorial L_1 selection procedure of Devroye and Lugosi [9] can be extended to L_p norms ($1 < p \leq \infty$) or to more general ϕ-divergences, such as Kullback-Leibler information or Hellinger distance. According to the authors' experience, the extension to L_p criteria seems feasible, at the price of some technical requirements extending Scheffé's identity [14]. On the other hand, the divergence case presents a more delicate problem. Here, one needs to carefully assess the divergence between two measures as a supremum of functionals over a suitable class of functions. Dual representations of divergences should provide a good starting point, see for example Keziou [11].

1.2.2 Selecting a modified histogram

In this paragraph, we will be concerned with the selection of a density g and a partition P in the modified histogram estimate, using the general combinatorial tools presented above. Let us first describe the mathematical model. We let $\mathcal{G}$ be a given class of candidate reference densities on $\mathbb{R}^d$, and we denote by ν_g the probability measure associated with each $g \in \mathcal{G}$. Consider $\mathcal{P}$ a family of candidate partitions of $\mathbb{R}^d$ such that each $P \in \mathcal{P}$ has at most r cells ($r \geq 2$, possibly function of n, and to be made precise later on). To each density $g \in \mathcal{G}$ and each partition $P = \{A_1, \ldots, A_\ell\} \in \mathcal{P}$ such that $\nu_g(A_i) = 1/\ell$, $i = 1, \ldots, \ell$, assign the corresponding modified histogram $f_{n,\theta}$ defined as in (1.1), with $\theta = (g, P)$. We use the minimum distance estimate to select θ from

$$\Theta = \left\{(g, P) : g \in \mathcal{G}, P = \{A_1, \ldots, A_\ell\} \in \mathcal{P},\, \ell \leq r,\, \nu_g(A_i) = 1/\ell\right\}, \tag{1.3}$$

the set of all possible pairs of reference densities and partitions. Denote by f_n the resulting minimum distance estimate. Now, to apply (1.2), we need to obtain upper bounds for the mth shatter coefficient $\mathbf{S}_{\mathcal{A}_\Theta}(m)$ of the Yatracos class associated with Θ. The following theorem is a key combinatorial result towards this direction. Denote by $\mathbf{S}_{\mathcal{D}}(j)$ the jth shatter coefficient of the class of sets

$$\mathcal{D} = \left\{\left\{(x, z) \in \mathbb{R}^d \times \mathbb{R}_+^* : \alpha z g(x) - g'(x) > 0\right\} : \alpha \in \mathbb{R}_+^*, (g, g') \in \mathcal{G}^2\right\},$$

and, with a slight abuse of notation, denote by $\mathbf{S}_{\mathcal{P}}(j)$ the jth shatter coefficient of the class of sets which are cells of any partition in $\mathcal{P}$.

Theorem 1.2.1 *If $\mathcal{A}_\Theta$ is the Yatracos class defined by* (1.3)*, then*

$$\mathbf{S}_{\mathcal{A}_\Theta}(m) \leq \mathbf{S}_{\mathcal{D}}(m)\left[\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\right]^{4r}.$$

Consequently

$$\mathbf{E}\left\{\int |f_n - f|\right\} \leq 3 \inf_{\theta \in \Theta} \mathbf{E}\left\{\int |f_{n-m,\theta} - f|\right\} + 8\sqrt{\frac{\log 2 + \log \mathbf{S}_{\mathcal{D}}(m) + 4r \log \mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)}{m}} + \frac{3}{n}. \tag{1.4}$$

Since in most cases of interest, bounds for $\mathbf{S}_{\mathcal{D}}(m)$ and $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$ are polynomial in m and n (detailed examples are presented in Section 3), one can choose m and r as functions of n such that the terms on the right hand side of (1.4) are balanced. More precisely:

Corollary 1.2.1 *Assume that the shatter coefficients* $\mathbf{S}_{\mathcal{D}}(m)$ *and* $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$ *are polynomial in their arguments. Then the choices*

$$m = \frac{n}{\log n} \qquad \textit{and} \qquad r = n^a, \quad a > 0,$$

lead to

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \le 3 \inf_{\theta \in \Theta} \mathbf{E}\left\{ \int |f_{n-m,\theta} - f| \right\} + \mathrm{O}\left(\frac{\log n}{n^{(1-a)/2}} \right).$$

The optimal L_1 error of the univariate modified histogram is known to go to zero, under standard smoothness assumptions, at the rate $n^{-1/3}$, provided $r \sim n^{1/3}$. Therefore, the bound above essentially says that for polynomial shatter coefficients $\mathbf{S}_{\mathcal{D}}(m)$ and $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$ and $a = 1/3$, we have asymptotically a performance that is guaranteed to be, up to a logarithm term, within a factor of three of the optimal performance. Roughly, the logarithm term appears as the price to be paid for using unrestricted classes of reference densities.

In order to use Theorem 1.2.1, we have to make sure that $\inf_{\theta\in\Theta} \mathbf{E} \int |f_{n-m,\theta} - f|$ is not much larger than $\inf_{\theta\in\Theta} \mathbf{E} \int |f_{n,\theta} - f|$, that is, holding out m observations does not cause much trouble. Whereas this result holds for parameter selection by the combinatorial method for most classical nonparametric density estimates (such as histograms, kernel estimates or wavelet estimates, see Devroye and Lugosi [9], Chapter 10), things turn out to be more complicated for the modified histogram estimate under study. Our result is as follows.

Theorem 1.2.2 *Denote by* μ *the common distribution of the* X_i*'s, and suppose that there exists a positive real number* α *such that* $\forall \theta \in \Theta$ $\big(\theta = (g, P), P = \{A_1, \ldots, A_\ell\}\big)$

$$\alpha \le \mu(A_i), \qquad i = 1, \ldots, \ell. \tag{1.5}$$

Then, for all $m \le n/2$*, we have*

$$\begin{aligned}
&\mathbf{E}\left\{ \int |f_n - f| \right\} \le \\
&\quad 3\left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}} + \frac{\sqrt{8}\, mr}{(n-m)\sqrt{n}\,\alpha(1-\alpha)} \right) \inf_{\theta\in\Theta} \mathbf{E}\left\{ \int |f_{n,\theta} - f| \right\} \\
&\quad + 8\sqrt{\frac{\log 2 + \log \mathbf{S}_{\mathcal{D}}(m) + 4r \log \mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)}{m}} + \frac{3}{n}.
\end{aligned}$$

Corollary 1.2.2 *Assume that the conditions of Theorem 1.2.2 are satisfied, and that the shatter coefficients* $\mathbf{S}_{\mathcal{D}}(m)$ *and* $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$ *are polynomial in their arguments. Then the choices*

$$m = \frac{n}{\log n} \qquad \textit{and} \qquad r = n^a, \quad 0 < a \leq 1/2,$$

lead to

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \leq 3\left(1 + \mathrm{O}\Big(\frac{1}{\sqrt{\log n}}\Big)\right) \inf_{\theta\in\Theta} \mathbf{E}\left\{ \int |f_{n,\theta} - f| \right\} + \mathrm{O}\left(\frac{\log n}{n^{(1-a)/2}}\right).$$

Roughly speaking, condition (1.5) means that the set of candidate reference densities $\mathcal{G}$ is not too far from the target f. It is in particular satisfied when $\mathcal{G}$ is finite or when $\mathcal{G}$ is the class of Gaussian densities with bounded mean and variance parameters, and $\nu_g \ll \mu$ for all $g \in \mathcal{G}$. Let us now discuss some examples.

1.3 Examples

In this section, we provide various useful bounds for the shatter coefficients $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$ and $\mathbf{S}_{\mathcal{D}}(m)$. We first recall that the *Vapnik-Chervonenkis dimension* V (Vapnik and Chervonenkis [15]) of a class $\mathcal{H}$ of sets is defined as the largest integer p such that

$$\mathbf{S}_{\mathcal{H}}(p) = 2^p.$$

If $\mathbf{S}_{\mathcal{H}}(p) = 2^p$ for all p, then we say that $V = \infty$. A classical consequence of Sauer's lemma [13] shows that if $\mathcal{H}$ has Vapnik-Chervonenkis dimension $V < \infty$, then

$$\mathbf{S}_{\mathcal{H}}(j) \leq (j+1)^V. \tag{1.6}$$

Let us first derive $\mathbf{S}_{\mathcal{P}}(j)$ for several classes of partitions $\mathcal{P}$ – recall that $\mathbf{S}_{\mathcal{P}}(j)$ means the jth shatter coefficient of the class of sets which are cells of any partition in $\mathcal{P}$. We first consider the univariate case $d = 1$.

1.3.1 Univariate modified histograms

As a simple but important example, consider $d = 1$, and let $\mathcal{P}$ be the class containing all partitions of the real line into at most r intervals. Denoting by G the distribution function associated with any reference density g, the intervals A_i for $P = \{A_1, \ldots, A_\ell\} \in \mathcal{P}$ are defined as follows:

$$\begin{aligned} A_i &= \Big(G^{-1}\Big(\frac{i-1}{\ell}\Big), G^{-1}\Big(\frac{i}{\ell}\Big)\Big], \ i = 1, \ldots, \ell - 1, \\ A_\ell &= \Big(G^{-1}\Big(1 - \frac{1}{\ell}\Big), G^{-1}(1)\Big), \end{aligned}$$

where G^{-1} denotes the quantile function defined on $[0,1]$ by $G^{-1}(u) = \inf\{x \in \mathbb{R} : G(x) \geq u\}$. Within this framework, $\mathbf{S}_{\mathcal{P}}(j)$ is at most the jth shatter coefficient of the class of all intervals, which equals $j(j+1)/2+1$. Note that Berlinet and Brunel [4], [5] study a univariate cross-validation-based method to select ℓ (but not g and ℓ simultaneously).

Let us now focus attention on the shatter coefficient $\mathbf{S}_{\mathcal{D}}(m)$ for two useful classes of univariate reference densities $\mathcal{G}$. Recall that

$$\mathcal{D} = \left\{ \left\{ (x,z) \in \mathbb{R}^d \times \mathbb{R}_+^* : \alpha z g(x) - g'(x) > 0 \right\} : \alpha \in \mathbb{R}_+^*, (g,g') \in \mathcal{G}^2 \right\}.$$

Exponential family. A family $\mathcal{G}$ of densities on $\mathbb{R}$ is called an *exponential family* if each density in $\mathcal{G}$ may be written in the form

$$g_\xi(x) = c\gamma(\xi)\beta(x)e^{\sum_{i=1}^k \pi_i(\xi)\psi_i(x)}, \tag{1.7}$$

where ξ belongs to some parameter set Ξ, $\psi_1, \ldots, \psi_k : \mathbb{R} \to \mathbb{R}$, $\beta : \mathbb{R} \to [0,\infty)$, $\gamma > 0, \pi_1, \ldots, \pi_k : \Xi \to \mathbb{R}$ are fixed functions, and c is a positive normalization constant. Examples of exponential families include classes of Gaussian, gamma, beta, Rayleigh, and Maxwell densities. Note that for $\alpha > 0$, $\alpha z g_\xi(x) > g_{\xi'}(x)$ if and only if

$$\log z + \sum_{i=1}^k \big(\pi_i(\xi) - \pi_i(\xi')\big)\psi_i(x) + \log \frac{\alpha\gamma(\xi)}{\gamma(\xi')} > 0\,. \tag{1.8}$$

By a mapping that makes each of the functions of x and z a new variable, it is easy to see that inequality (1.8) is just a homogeneous linear inequality $a_1\lambda_1 + \ldots + a_{k+2}\lambda_{k+2} > 0$, with the coefficients a_i depending upon the pair (ξ,ξ') only. The Vapnik-Chervonenkis dimension for a collection of linear halfspaces in $\mathbb{R}^{k+2}$ is not more than $k+2$ (Devroye and Lugosi [9], Corollary 4.2). As a consequence, by (1.6),

$$\mathbf{S}_{\mathcal{D}}(m) \leq (m+1)^{k+2}\,.$$

Series estimates. Let $\psi_1, \ldots, \psi_k$ be fixed nonnegative basis functions from $\mathbb{R}^d$ to $\mathbb{R}$ such that $\int \psi_i = t_i$ for $1 \leq i \leq k$. We define the class $\mathcal{G}$ as the collection of all linear combinations

$$g_\xi(x) = \sum_{i=1}^k a_i\psi_i(x)$$

with coefficient $\xi = (a_1, \ldots, a_k)$ satisfying $\sum_{i=1}^k a_i t_i = 1$. Clearly, for $\alpha > 0$, $\alpha z g_\xi(x) > g_{\xi'}(x)$ if and only if

$$\sum_{i=1}^k \alpha a_i z\psi_i(x) - \sum_{i=1}^k a_i'\psi_i(x) > 0\,.$$

Making again each of the functions $\psi_i(x)$ and $z\psi_i(x)$ a new variable, we are led to a homogeneous linear inequality $b_1\lambda_1 + \ldots + b_{2k}\lambda_{2k} > 0$, with coefficients b_i depending upon the pair (ξ, ξ') only. Therefore

$$\mathbf{S}_{\mathcal{D}}(m) \leq (m+1)^{2k}\,.$$

1.3.2 Multivariate modified histograms

The aim of this paragraph is to study multivariate modified histograms defined via a multinormal reference density. This leads us to consider the class

$$\mathcal{G} = \left\{ g_{m,\Sigma}(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}}\, e^{-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)} \right\},$$

where m is an arbitrary element of $\mathbb{R}^d$ and Σ is a symmetric positive definite $d \times d$ matrix. For a given reference density $g_{m,\Sigma} \in \mathcal{G}$ and a given integer $\ell \geq 2$, we let the partition P be as follows.

- Set $\ell = \ell_1 \ldots \ell_d$, with $\ell_1, \ldots, \ell_d$ positive integers, and let $h_j = 1/\ell_j$ for $j = 1, \ldots, d$;
- For $j = 1, \ldots, d$ and $i_j = 1, \ldots, \ell_j - 1$, compute the quantiles of order $i_j h_j$ of a univariate standard normal $\mathcal{N}(0,1)$; denote by q_{j,i_j} these quantiles, with the convention $q_{j,0} = -\infty$ and $q_{j,\ell_j} = +\infty$;
- Consider the grid defined by the above family $\{q_{j,i_j}\}$; this grid leads to a partition of $\mathbb{R}^d$ into ℓ hyperrectangles, say $\tilde{A}_{i_1,\ldots,i_d}$, $1 \leq j \leq d, 1 \leq i_j \leq \ell_j$;
- Fix $T_{m,\Sigma}$ the affine transformation

  $$T_{m,\Sigma}(x) = \Sigma^{1/2}x + m\,,$$

  and let $\{A_{i_1,\ldots,i_d}\}$ be the image-partition of $\{\tilde{A}_{i_1,\ldots,i_d}\}$ by $T_{m,\Sigma}$ (see Figure 1.2 that depicts a bivariate example).

Finally take

$$P = \left\{ A_{i_1,\ldots,i_d} \right\}_{\substack{1 \leq j \leq d \\ 1 \leq i_j \leq \ell_j}}.$$

Denote by $\nu_{m,\Sigma}$ the probability measure associated with the reference $g_{m,\Sigma}$. It is easily seen that, for any cell $A_{i_1,\ldots,i_d}$ of the partition P,

$$\nu_{m,\Sigma}(A_{i_1,\ldots,i_d}) = 1/\ell\,.$$

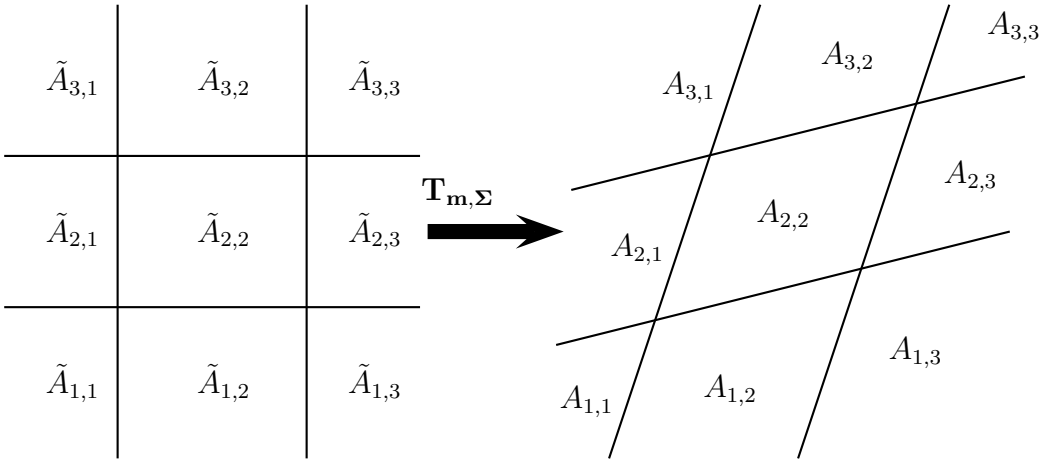

Figure 1.2: Transformation of a partition in $\mathbb{R}^2$.

Note however that the decomposition $\ell = \ell_1 \dots \ell_d$ is not necessarily unique. Thus, given $g_{m,\Sigma} \in \mathcal{G}$ and $\ell \geq 2$, we have just constructed a partition of $\mathbb{R}^d$ into ℓ sets of $\nu_{m,\Sigma}$-measure $1/\ell$. Clearly, each set in any such partition is an intersection of at most $2d$ hyperplanes (it is a polytope with at most $2d$ faces). Therefore

$$\mathbf{S}_{\mathcal{P}}(j) \leq (j+1)^{2d(d+1)}$$

(see for example Devroye, Györfi and Lugosi [8]).

Let us now consider the shatter coefficient $\mathbf{S}_{\mathcal{D}}(m)$. Here, $\mathcal{G}$ is the class of multinormal densities, hence it is a multivariate exponential family. More precisely, setting $\xi = (m, \Sigma)$, each g_ξ in $\mathcal{G}$ may be written in the form

$$g_\xi(x) = c\gamma(\xi)\beta(x)e^{\sum_{i=1}^{k} \pi_i(\xi)\psi_i(x)},$$

with the notation of (1.7) – just replace $\mathbb{R}$ with $\mathbb{R}^d$ – and with $k = d(d+3)/2$. We conclude that

$$\mathbf{S}_{\mathcal{D}}(m) \leq (m+1)^{d(d+3)/2+2}.$$

Note that the bounds on the shatter coefficients in the examples presented above are polynomial in their arguments, so that Corollary 1.2.1 and Corollary 1.2.2 apply. One can argue that the bound $r = n^a$ is somewhat restrictive. However, extensive simulations (see Berlinet and Biau [3]) reveal that the number of cells ℓ should be very small with respect to n. Therefore, in practice, the bound $r = n^a$ does not harm too much. Moreover, it is consistent with the results of Barron, Györfi and van der Meulen [2], who proved that a univariate Kullback-Leibler-based choice of ℓ is of order $n^{1/3}$.

1.4 Simulations

In this section, we illustrate the theory with univariate simulation results enlightening the efficiency of the combinatorial method. The density to be estimated, a Beta $(2, 2)$, is shown in Figure 1.3.

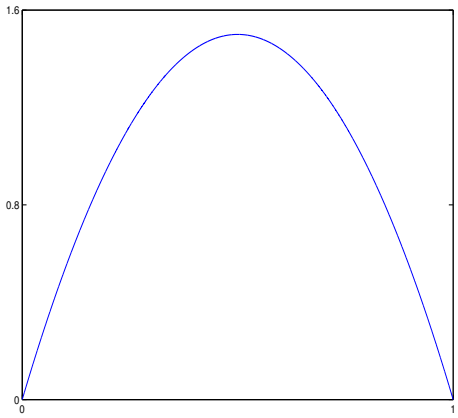

Figure 1.3: Density Beta $(2, 2)$ to be estimated.

We consider a class $\mathcal{G}$ of references including 9 densities, presented in Figure 1.4.

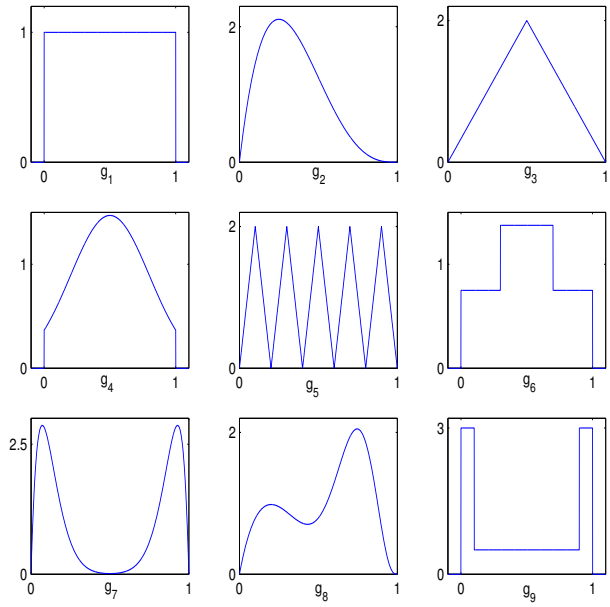

Figure 1.4: Collection of reference densities.

Given a reference g in the collection $\mathcal{G}$ and an integer ℓ, the associated partition is constructed via the quantiles of the density g, as explained in Paragraph 3.1. Thus, in this context, the method will automatically select a parameter θ from the set

$$\Theta = \left\{(g, \ell) : g \in \mathcal{G}, 2 \leq \ell \leq r\right\}.$$

The resulting minimum distance estimate is denoted f_n.

We also shed light on the advantages of selecting both the partition and the reference density in contrast to the case where only the partition is selected. To this aim, for each *fixed* reference density $g \in \mathcal{G}$, we run the combinatorial method to select the sole number of cells ℓ from the set $\Theta_g = \{\ell : 2 \leq \ell \leq r\}$, and we denote by $f_{n,g}$ the elected estimate.

To assess the quality of the selected estimates, we compare the L_1 performances of the elected f_n and $f_{n,g}$ with the best estimates $f_{n,\theta^\star}$ and $f_{n,\theta_g^\star}$ in the corresponding classes, that is

$$\theta^\star \in \operatorname*{argmin}_{\theta \in \Theta} \left\{ \int |f_{n,\theta} - f| \right\},$$

and, for a fixed g,

$$\theta_g^\star \in \operatorname*{argmin}_{\theta \in \Theta_g} \left\{ \int |f_{n,\theta} - f| \right\}.$$

Table 1.1 and Table 1.2 summarize the results. For each of the references g, we display in Table 1.1 the L_1 error of $f_{n,g}$ and $f_{n,\theta_g^\star}$, and we present in Table 1.2 the error of the estimates f_n and $f_{n,\theta^\star}$. We also show the number $\hat{\ell}_n$ of selected classes. All results are averaged over 50 repetitions.

The L_1 error ratios selected / optimal never exceed 2.26, and all of these results enlighten the good performances of the combinatorial method in general. They also clearly show the advantages of selecting both the partition and the reference density in contrast to the case where only the partition is selected. As a matter of fact, the L_1 performances of f_n over the $f_{n,g}$'s are significantly better for 5 reference models out of 9, and roughly similar for 2. Unsurprisingly, the best performances of $f_{n,g}$ are obtained for the densities g_3 (triangle) and g_4 (truncated Gaussian $\mathcal{N}(0.5, 1)$), which resemble the most the density Beta $(2, 2)$. In practice, when one has no or few a priori information on the target density, the selection approach presented in the present paper is preferable.

	$n = 200, m = 50, r = 16$			$n = 1000, m = 150, r = 30$		
g	$\int \lvert f_{n,g} - f \rvert$	$\int \lvert f_{n,\theta_g^\star} - f \rvert$	$\hat{\ell}_n$	$\int \lvert f_{n,g} - f \rvert$	$\int \lvert f_{n,\theta_g^\star} - f \rvert$	$\hat{\ell}_n$
g_1	0.2060	0.1536	9.68	0.1205	0.0958	17.20
g_2	0.3254	0.2961	12.92	0.2379	0.2228	24.24
g_3	0.1677	0.1103	7.28	0.1043	0.0695	15.12
g_4	0.1767	0.1036	8.28	0.1119	0.0849	14.08
g_5	0.4327	0.4000	14.28	0.3358	0.3176	24.72
g_6	0.2340	0.1891	10.84	0.1419	0.1141	18.08
g_7	0.8241	0.8135	15.64	0.6714	0.6633	29.44
g_8	0.2241	0.1743	9.04	0.1424	0.1144	17.04
g_9	0.2399	0.1728	10.92	0.1370	0.1089	19.12

Table 1.1: Combinatorial method results for the selection of P.

$n = 200, m = 50, r = 16$			$n = 1000, m = 150, r = 30$		
$\int \lvert f_n - f \rvert$	$\int \lvert f_{n,\theta^\star} - f \rvert$	$\hat{\ell}_n$	$\int \lvert f_n - f \rvert$	$\int \lvert f_{n,\theta^\star} - f \rvert$	$\hat{\ell}_n$
0.2249	0.0995	8.28	0.1469	0.0694	16.32

Table 1.2: Combinatorial method results for the selection of the pair (g, P).

1.5 Proofs

1.5.1 Proof of Theorem 1.2.1

We just have to prove that

$$\mathbf{S}_{\mathcal{A}_\Theta}(m) \leq \mathbf{S}_{\mathcal{D}}(m) \big[\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\big]^{4r} ,$$

and the second part of the theorem will directly follow from inequality (1.2).

Let $y_1, \ldots, y_m$ be m distinct vectors in $\mathbb{R}^d$. For each $\theta = (g, P) \in \Theta$, $P = \{A_1, \ldots, A_\ell\}$, consider the $m \times r$ matrix z_θ such that the element in its tth row and jth column is

$$z_\theta^{(t,j)} = \begin{cases} \mathbf{1}_{[y_t \in A_j]} \sum_{i=1}^{n-m} \mathbf{1}_{[X_i \in A_j]} & \text{for } t \leq m,\ j \leq \ell\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Clearly,

$$\mathbf{1}_{[y_t \in A_j]} \mathbf{1}_{[X_i \in A_j]} = 1 \quad \text{if and only if} \quad (y_t, X_i) \in A_j \times A_j\,.$$

Since there are $m(n-m)$ different pairs (y_t, X_i), the number of different values the jth column $(z_\theta^{(1,j)}, \dots, z_\theta^{(m,j)})$ of the matrix z_θ can take as we vary $\theta \in \Theta$ is at most the shatter coefficient $\mathbf{S}_{\mathcal{C}}\big(m(n-m)\big)$ of the class of sets $\mathcal{C}$ of the form $A \times A$, where A is any set in any possible partition in $\mathcal{P}$. This shatter coefficient is clearly bounded by the square of the shatter coefficient $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$. Hence the jth column of the matrix z_θ can take at most $\big[\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\big]^2$ values. But since the matrix z_θ has r columns, it can take at most

$$\big[\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\big]^{2r}$$

values. Thus if we set

$$\mathcal{W} = \big\{(z_\theta, z_{\theta'}) : (\theta, \theta') \in \Theta^2\big\},$$

we have

$$\operatorname{Card} \mathcal{W} \le \big[\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\big]^{4r}.$$

For fixed $(w, w') \in \mathcal{W}$, let $U_{(w,w')}$ denote the collection of all (θ, θ') such that $(z_\theta, z_{\theta'}) = (w, w')$. For $(\theta, \theta') \in U_{(w,w')}$ $\big(\theta = (g, P)$, $\theta' = (g', P')$, $P = \{A_1, \dots, A_\ell\}$, $P' = \{A'_1, \dots, A'_{\ell'}\}\big)$ and $t \le m$, we have

$$y_t \in A_{\theta,\theta'} = \big\{x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x)\big\}$$

if and only if

$$\frac{\sum_{j=1}^{\ell} z_\theta^{(t,j)} + 1}{(n-m)h + 1}\, g(y_t) > \frac{\sum_{j=1}^{\ell'} z_{\theta'}^{(t,j)} + 1}{(n-m)h' + 1}\, g'(y_t)\,,$$

where $h = 1/\ell$ and $h' = 1/\ell'$. Within the set $U_{(w,w')}$, $z_\theta^{(t,j)}$ and $z_{\theta'}^{(t,j)}$ are fixed for all t and j. Therefore, with the notation

$$z_t = \frac{\sum_{j=1}^{\ell} z_\theta^{(t,j)} + 1}{\sum_{j=1}^{\ell'} z_{\theta'}^{(t,j)} + 1} \qquad \text{for } 1 \le t \le m\,,$$

we obtain that $y_t \in A_{\theta,\theta'}$ if and only if

$$\frac{(n-m)h' + 1}{(n-m)h + 1}\; z_t\, g(y_t) - g'(y_t) > 0\,.$$

It follows that

$$\begin{aligned}
&\operatorname{Card}\big\{\{\mathbf{1}_{[y_1 \in A_{\theta,\theta'}]}, \dots, \mathbf{1}_{[y_m \in A_{\theta,\theta'}]}\} : (\theta, \theta') \in U_{(w,w')}\big\} \\
&\quad \le \operatorname{Card}\big\{\{\mathbf{1}_{[\alpha z_1 g(y_1) - g'(y_1) > 0]}, \dots, \mathbf{1}_{[\alpha z_m g(y_m) - g'(y_m) > 0]}\} : \alpha \in \mathbb{R}_+^*, (g, g') \in \mathcal{G}^2\big\} \\
&\quad \le \mathbf{S}_{\mathcal{D}}(m)\,.
\end{aligned}$$

Putting all pieces together, we obtain

$$\begin{aligned}\operatorname{Card}\bigl\{\{y_1,\ldots,y_m\}\cap A_{\theta,\theta'}:(\theta,\theta')\in\Theta^2\}\bigr\} &\le \mathbf{S}_{\mathcal{D}}(m)\operatorname{Card}\mathcal{W}\\ &\le \mathbf{S}_{\mathcal{D}}(m)\bigl[\mathbf{S}_{\mathcal{P}}\bigl(m(n-m)\bigr)\bigr]^{4r}.\end{aligned}$$

The proof of Theorem 1.2.1 is finished. ■

1.5.2 Proof of Theorem 1.2.2

The proof of Theorem 1.2.2 is a consequence of Theorem 1.2.1 and the following lemma.

Lemma 1.5.1 *Denote by μ the common distribution of the X_i's, and suppose that there exists a positive real number α such that $\forall\theta\in\Theta$ $\bigl(\theta=(P,g), P=\{A_1,\ldots,A_\ell\}\bigr)$*

$$\alpha\le\mu(A_i)\,,\qquad i=1,\ldots,\ell.$$

Introduce

$$J_{n,\theta}=\int|f_{n,\theta}-f|\,.$$

If m is a positive integer such that $2m\le n$, then

$$\frac{\inf_{\theta\in\Theta}\mathbf{E}\bigl\{J_{n-m,\theta}\bigr\}}{\inf_{\theta\in\Theta}\mathbf{E}\bigl\{J_{n,\theta}\bigr\}}\le 1+\frac{2m}{n-m}+8\sqrt{\frac{m}{n}}+\frac{\sqrt{8}\,mr}{(n-m)\sqrt{n}\,\alpha(1-\alpha)}\,.$$

Proof of Lemma 1.5.1 Note first that the modified histogram is *not* an additive estimate in the sense of Devroye and Lugosi [9] so that their Theorem 10.2 does not apply. Nevertheless we can start with the inequality that they prove:

$$\inf_{\theta\in\Theta}\mathbf{E}\{J_{n-m,\theta}\}\le\inf_{\theta\in\Theta}\mathbf{E}\{J_{n,\theta}\}\left(1+2\sup_{\theta\in\Theta}\frac{\mathbf{E}\bigl\{\int|f_{n-m,\theta}-f_{n,\theta}|\,\mathrm{d}x\bigr\}}{\mathbf{E}\bigl\{\int|f_{n,\theta}-\mathbf{E}f_{n,\theta}|\,\mathrm{d}x\bigr\}}\right).$$

Fix x and $\theta=(g,P)$ for now and define $K_\theta(x,X_i)=\mathbf{1}_{[X_i\in A(x)]}$. Recall that $A(x)$ denotes the cell of the partition P (which has ℓ cells) in which x falls. Observe that

$$f_{n,\theta}(x)=\frac{1}{nh+1}\Bigl(1+\sum_{i=1}^{n}K_\theta(x,X_i)\Bigr)\,g(x)\,,$$

where $h=1/\ell$. Introduce

$$Y_i=K_\theta(x,X_i)-\mathbf{E}\bigl\{K_\theta(x,X_i)\bigr\}\,,$$

and denote the partial sums of Y_i's by $S_j=Y_1+\ldots+Y_j$. Observe the following:

$$
\begin{aligned}
&(nh+1)|f_{n-m,\theta}(x) - f_{n,\theta}(x)| \\
&\quad = \left| \frac{nh+1}{(n-m)h+1} \Big(1 + \sum_{i=1}^{n-m} K_\theta(x, X_i)\Big) - \Big(1 + \sum_{i=1}^{n} K_\theta(x, X_i)\Big) \right| g(x) \\
&\quad = \left| \frac{mh}{(n-m)h+1} \Big(1 + \sum_{i=1}^{n-m} K_\theta(x, X_i)\Big) - \sum_{i=n-m+1}^{n} K_\theta(x, X_i) \right| g(x) \\
&\quad = \left| \frac{mh}{(n-m)h+1} (Y_1 + \ldots + Y_{n-m}) - (Y_{n-m+1} + \ldots + Y_n) \right. \\
&\qquad \left. + \frac{m}{(n-m)h+1} \Big(h - \mathbf{E}\big\{K_\theta(x, X_1)\big\}\Big) \right| g(x)\,,
\end{aligned}
$$

so that

$$
\begin{aligned}
\mathbf{E}\big\{(nh+1)|f_{n-m,\theta}(x) - f_{n,\theta}(x)|\big\} \le \Bigg[&\frac{m}{n-m} \mathbf{E}\big\{|S_{n-m}|\big\} + \mathbf{E}\big\{|S_m|\big\} \\
&+ \frac{m}{(n-m)h+1} \big|h - \mathbf{E}\big\{K_\theta(x, X_1)\big\}\big| \Bigg] g(x)\,.
\end{aligned}
$$

Also,

$$
(nh+1)|f_{n,\theta}(x) - \mathbf{E} f_{n,\theta}(x)| = |S_n|\, g(x)\,,
$$

which implies

$$
\mathbf{E}\big\{(nh+1)|f_{n,\theta}(x) - \mathbf{E} f_{n,\theta}(x)|\big\} = \mathbf{E}\big\{|S_n|\big\}\, g(x)\,.
$$

If $2m \le n$, a straightforward consequence of Lemma 10.1 and Lemma 10.3 in Devroye and Lugosi (2001) leads to

$$
\frac{\mathbf{E}\big\{|f_{n-m,\theta} - f_{n,\theta}|\big\}}{\mathbf{E}\big\{|f_{n,\theta} - \mathbf{E} f_{n,\theta}|\big\}} \le \frac{m}{n-m} + 4\sqrt{\frac{m}{n}} + \frac{m}{(n-m)h+1}\, \frac{\sqrt{8}\,\big|h - \mathbf{E}\big\{K_\theta(x, X_1)\big\}\big|}{\sqrt{n}\,\mathbf{E}\big\{|Y_1|\big\}}\,. \tag{1.9}
$$

Let $p(x)$ stand for $\mu\big(A(x)\big)$. Clearly,

$$
\begin{cases}
\mathbf{E}\big\{K_\theta(x, X_1)\big\} = p(x) \\
\mathbf{E}\big\{|Y_1|\big\} = 2p(x)\big(1 - p(x)\big)\,.
\end{cases}
$$

By assumption, and using the fact that $\ell \ge 2$, we obtain, still holding x fixed,

$$
\alpha \le p(x) \le 1 - \alpha\,.
$$

Note that $0 < \alpha \le 1/2$. By (1.9)

$$
\frac{\mathbf{E}\big\{|f_{n-m,\theta} - f_{n,\theta}|\big\}}{\mathbf{E}\big\{|f_{n,\theta} - \mathbf{E} f_{n,\theta}|\big\}} \le \frac{m}{n-m} + 4\sqrt{\frac{m}{n}} + \frac{m}{(n-m)h+1}\, \frac{\sqrt{8}\,|h - p(x)|}{2\sqrt{n}\,p(x)\big(1-p(x)\big)}\,.
$$

Moreover

$$\frac{1}{p(x)\big(1-p(x)\big)} \le \frac{1}{\alpha(1-\alpha)}\,.$$

On the other hand,

$$|h-p(x)| \le \max\left(1, \frac{p(x)}{h}\right) h \le rh\,.$$

Putting all pieces together, we obtain

$$\begin{aligned}\frac{m}{(n-m)h+1}\,\frac{\sqrt{8}\,|h-p(x)|}{2\sqrt{n}\,p(x)\big(1-p(x)\big)} \le& \frac{mh}{(n-m)h+1}\,\frac{\sqrt{2}\,r}{\sqrt{n}\,\alpha(1-\alpha)}\\ \le& \frac{\sqrt{2}\,mr}{(n-m)\sqrt{n}\,\alpha(1-\alpha)}\,.\end{aligned}$$

This implies that for any fixed θ

$$\begin{aligned}&\mathbf{E}\Big\{\int |f_{n-m,\theta}-f_{n,\theta}|\,\mathrm{d}x\Big\}\\ &\le \left(\frac{m}{n-m}+4\sqrt{\frac{m}{n}}+\frac{\sqrt{2}\,mr}{(n-m)\sqrt{n}\,\alpha(1-\alpha)}\right)\mathbf{E}\Big\{\int |f_{n,\theta}-\mathbf{E}f_{n,\theta}|\,\mathrm{d}x\Big\}\,.\end{aligned}$$

This completes the proof of the lemma. ■

Bibliography

[1] A. R. Barron. The convergence in information of probability density estimators. In *Proceedings of the International Symposium of IEEE on Information Theory*, Kobe: Japan, June 19-24 1988.

[2] A. R. Barron, L. Györfi, and E. C. van der Meulen. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Transactions on Information Theory*, 38:1437–1454, 1992.

[3] A. Berlinet and G. Biau. Iterated modified histograms as dynamical systems. *Journal of Nonparametric Statistics*, 16:385–401, 2004.

[4] A. Berlinet and E. Brunel. Choix optimal du nombre de classes pour l'estimateur de Barron de la densité. *Comptes Rendus de l'Académie des Sciences de Paris*, 331:713–716, 2000.

[5] A. Berlinet and E. Brunel. Cross-validated density estimates based on Kullback-Leibler information. *Journal of Nonparametric Statistics*, 16:493–513, 2004.

[6] A. Berlinet, L. Györfi, and E. C. van der Meulen. Asymptotic normality of relative entropy in multivariate density estimation. *Publications de l'Institut de Statistique de l'Université de Paris*, 41:3–27, 1997.

[7] A. Berlinet, I. Vajda, and E. C. van der Meulen. About the asymptotic accuracy of Barron density estimates. *IEEE Transactions on Information Theory*, 44:999–1009, 1998.

[8] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer–Verlag, New York, 1996.

[9] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation.* Springer–Verlag, New York, 2001.

[10] L. Györfi, F. Liese, I. Vajda, and E. C. van der Meulen. Distribution estimates consistent in χ^2-divergence. *Statistics*, 32:31–57, 1998.

[11] A Keziou. Dual representation of ϕ-divergences and applications. *Comptes Rendus de l'Académie des Sciences de Paris*, 336:857–862, 2003.

[12] F. Liese and I. Vajda. *Convex Statistical Distances*. Teubner–Verlag, Leipzig, 1987.

[13] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13:145–147, 1972.

[14] H. Scheffé. A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, 18:434–438, 1947.

[15] V. N. Vapnik and Chervonenkis, A. Ya. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[16] Y. G. Yatracos. Rates of convergence of minimun distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13:768–774, 1985.

Chapter 2

Optimal L_1 Bandwidth Selection for Variable Kernel Density Estimates*

Abstract

It is well established that one can improve performance of kernel density estimates by varying the bandwidth with the location and/or the sample data at hand. Our interest in this paper is in the data-based selection of a variable bandwidth within an appropriate parameterized class of functions. We present an automatic selection procedure inspired by the combinatorial tools developed in Devroye and Lugosi [10]. It is shown that the expected L_1 error of the corresponding selected estimate is up to a given constant multiple of the best possible error plus an additive term which tends to zero under mild assumptions.

2.1 Introduction

Assume we are given an *i.i.d.* sample $X_1, \ldots, X_n$ drawn from an unknown probability density f on $\mathbb{R}^d$. One of the most popular estimates of f is the *fixed bandwidth kernel estimate* defined by

$$f_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big), \quad x \in \mathbb{R}^d, \tag{2.1}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a kernel with $\int K = 1$ and $h > 0$ is the bandwidth (or smoothing parameter), see Rosenblatt [16] or Parzen [15]. The terminology

*Article écrit en collaboration avec Alain Berlinet et Gérard Biau, et publié dans la revue *Statistics and Probability Letters.*

fixed bandwidth means that the parameter h is held constant across x and the X_i's (but it can depend on n). While the estimate (2.1) performs well for most regular densities, its capabilities are known to decrease when estimating more complex functions such as multimodal densities (Sain and Scott [18]). Moreover, as the dimensionality increases, the so-called curse of dimensionality affects the quality of the estimation. Due to the sparseness of data in higher dimensions, multivariate neighborhoods are often empty, particularly in the tails of the density. Therefore, larger and larger bandwidths are necessary in the tails. However, this also has adverse effect of oversmoothing the main features (such as bumps and modes, see Sain [17]). These drawbacks can be overcome, to some extent, by varying the bandwidth in order to better capture the local behavior of the underlying density. For that purpose, two big families of *variable* (*bandwidth*) *kernel estimates* have been considered in the literature.

The variable estimates of the first family have a bandwidth which is allowed to vary with the location x. Its members are often referred to as *balloon estimates* and take the form

$$f_n(x) = \frac{1}{nh(x)} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h(x)}\Big) .$$

Such estimates lead to substantial gains over the fixed bandwidth in higher dimensional spaces and, to some extent, circumvent the curse of dimensionality (Terrell and Scott [20]).
The second family of variable kernel estimates was originally considered by Breiman, Meisel and Purcell [7], who suggested varying the bandwidth at each sample point, leading to the so-called *sample point estimates*

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(X_i)} K\Big(\frac{x - X_i}{h(X_i)}\Big) . \tag{2.2}$$

An appealing property of the above estimate is that a good choice of $h(X_i)$ allows to reduce the bias. As a matter of fact, Abramson [1] shows that the bias-rate usually reserved for fixed kernel estimates using negative fourth order kernels is actually achievable by estimates of the form (2.2). For a complete and comprehensive description of variable kernel estimates and their properties, we refer the reader to Jones [12] who also discusses a variable bandwidth depending on both the location and the sample points.

To exploit the advantages offered by variable kernel estimates, one has to design a good data-dependent way of determining the bandwidth function. As an important (but negative) result towards this direction, Devroye and Lugosi [9] show that it is impossible to find an optimal way of selecting the smoothing function

if this latter is allowed to depend on the location x only. More precisely, consider the class of univariate variable estimates

$$f_{n,h(x)}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(x)} K\Big(\frac{x - X_i}{h(x)}\Big) ,$$

where the bandwidth $h(x)$ is allowed to be any measurable function $h : \mathbb{R} \to (0, \infty)$ of x. Then a data-based variable kernel estimate has the form

$$f_{n,H(x)}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{H(x)} K\Big(\frac{x - X_i}{H(x)}\Big) ,$$

where it is understood that $H(x) = H(x; X_1, \ldots, X_n)$. Ideally, one would like to find $H(x)$ so that the expected error $\mathbf{E}\{\int |f_{n,H(x)}(x) - f(x)| \, \mathrm{d}x\}$ is close to the ideal value $\inf_{h:\mathbb{R}\to(0,\infty)} \mathbf{E}\{\int |f_{n,h(x)}(x) - f(x)| \, \mathrm{d}x\}$ for all densities. Unfortunately, Devroye and Lugosi [9] prove that if K is a symmetric nonnegative square-integrable kernel with compact support, then

$$\inf_{H:\mathbb{R}^{n+1}\to(0,\infty)} \sup_{f\in\mathcal{F}_B} \frac{\mathbf{E}\big\{ \int |f_{n,H(x)}(x) - f(x)| \, \mathrm{d}x\}}{\inf_{h:\mathbb{R}\to(0,\infty)} \mathbf{E}\{\int |f_{n,h(x)}(x) - f(x)| \, \mathrm{d}x\}} \geq C n^{\frac{1}{10}} ,$$

where $\mathcal{F}_B$ denotes the class of nondecreasing, convex-shaped densities f on $[0, 1]$ with $\sup_{(0,1)} f(x) \leq B$ and C is a positive universal constant. This inequality shows that even with the knowledge that $f \in \mathcal{F}_B$, one cannot efficiently design a variable bandwidth. In other words this class of variable bandwidth kernel estimates is too large to be optimized.

Thus, one should constrain the class of possible bandwidth functions from which selection is made. This is precisely the problem that we address in the present paper, using a general multivariate data-based combinatorial methodology presented in Devroye and Lugosi [10]. More precisely, we will show how to select the smoothing function within an appropriate class so that the expected L_1 error of the corresponding selected estimate is up to a given constant multiple of the best possible error plus an additive term which tends to zero under mild assumptions. The paper is organized as follows. In Section 2, we present the multivariate selection procedure. We then specify the algorithm to the bandwidth function selection problem in Section 3. Examples are worked out in Section 4 for different models of variable kernel estimates, and Section 5 is devoted to the proofs.

2.2 Automatic parameter selection

Using ideas from Yatracos [22], Devroye and Lugosi [10] explore a new paradigm for the data-based or automatic selection of the free parameters of density estimates in general so that the expected L_1 error is within a given constant multiple

of the best possible error. To summarize in the present context, assume we are given a class of density estimates parameterized by $\theta \in \Theta$ such that $f_{n,\theta}$ denotes the density estimate with parameter θ. Moreover, assume that each $f_{n,\theta}$ may be written in the form

$$f_{n,\theta}(x) = \frac{1}{n} \sum_{i=1}^{n} K_\theta(x, X_i)\,,$$

where $K_\theta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a measurable function such that $\mathbf{E}\{|K_\theta(x, X)|\} < \infty$ for each x. Such estimates are called additive and regular (Devroye and Lugosi [10], Chapter 10). Examples of additive and regular estimates include the kernel, histogram, series and wavelet estimates. Now, let $m < n$ be an integer which splits the data $X_1, \ldots, X_n$ into

- a set $X_1, \ldots, X_{n-m}$ used for the construction of the density estimates;
- a validation set $X_{n-m+1}, \ldots, X_n$.

Introduce the class of random sets

$$\mathcal{A}_\Theta = \Big\{ \big\{ x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x) \big\} : (\theta, \theta') \in \Theta^2 \Big\}$$

($\mathcal{A}_\Theta$ is the so-called *Yatracos class* associated with Θ) and define

$$\Delta_\theta = \sup_{A \in \mathcal{A}_\Theta} \left| \int_A f_{n-m,\theta} - \mu_m(A) \right|,$$

where $\mu_m(A) = (1/m) \sum_{i=n-m+1}^{n} \mathbf{1}_{[X_i \in A]}$ is the empirical measure associated with the subsample $X_{n-m+1}, \ldots, X_n$. Then the *minimum distance estimate* f_n is defined as any density estimate selected among those $f_{n-m,\theta}$ with

$$\Delta_\theta < \inf_{\theta^* \in \Theta} \Delta_{\theta^*} + \frac{1}{n}\,.$$

Note that the $1/n$ term is added to ensure the existence of such a density estimate. According to Devroye and Lugosi [10], Chapter 10, whenever $f_{n-m,\theta}$ is integrable (and not necessarily nonnegative), the selected f_n satisfies the following inequality, valid for all n and $m \le n/2$:

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \le 5\left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right) \inf_{\theta \in \Theta} \mathbf{E}\left\{ \int |f_{n,\theta} - f| \right\} + 8\,\mathbf{E}\left\{ \sqrt{\frac{\log 2\,\mathbf{S}_{\mathcal{A}_\Theta}(m)}{m}} \right\} + \frac{5}{n} \tag{2.3}$$

(for the sake of clarity, we drop the "dx" notation when no confusion is possible). Here, $\mathbf{S}_{\mathcal{A}_\Theta}(m)$ is the *Vapnik-Chervonenkis shatter coefficient* of the class of sets $\mathcal{A}_\Theta$ (Vapnik and Chervonenkis [21]), defined by

$$\mathbf{S}_{\mathcal{A}_\Theta}(m) = \max_{x_1,\ldots,x_m \in \mathbb{R}^d} \text{Card}\big\{\{x_1,\ldots,x_m\} \cap A : A \in \mathcal{A}_\Theta\big\}.$$

This general methodology provides us with an automatic procedure to construct a density estimate f_n whose L_1 error is (almost) as small as that of the best estimate among the $f_{n,\theta}$, $\theta \in \Theta$. We emphasize that inequality (2.3) is nonasymptotic, that is, the bound is valid for all n. The rest of the analysis consists in obtaining upper bounds for the value of $\mathbf{S}_{\mathcal{A}_\Theta}(m)$.

2.3 Selecting a variable kernel estimate

In this section, we will be concerned with the selection of a bandwidth function in the variable kernel estimate, using the general combinatorial tools presented above. Moreover, to improve performance over ordinary kernel estimates for densities with varying behavior in different regions of the space, we shall use different parameterized smoothing functions in different regions of $\mathbb{R}^d$. For that purpose, let $\mathcal{P}_1$ (*resp.* $\mathcal{P}_2$) be a class of partitions of $\mathbb{R}^d$ such that each $P_1 = \{B_1^1,\ldots,B_{r_1}^1\} \in \mathcal{P}_1$ (*resp.* $P_2 = \{B_1^2,\ldots,B_{r_2}^2\} \in \mathcal{P}_2$) has at most r_1 (*resp.* r_2) cells. Denote by J the set $\{1,\ldots,r_1\}\times\{1,\ldots,r_2\}$ and by $\underline{\lambda} = (\lambda_{j_1 j_2} : (j_1,j_2) \in J)$ a generic vector of $\mathbb{R}^{r_1 r_2 p}$.

To go straight to the point, we will assume that the variable bandwidth is a parameterized measurable function $h : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{r_1 r_2 p} \to (0,\infty)$ of the form

$$h(x,X_i,\theta) = \sum_{(j_1,j_2)\in J} \phi(x,X_i,\lambda_{j_1 j_2}) \mathbf{1}_{B_{j_1}^1 \times B_{j_2}^2}(x,X_i)\,,$$

where the parameter $\theta = (P_1,P_2,\underline{\lambda})$ and, for fixed x and X_i, each map $\lambda_{j_1 j_2} \mapsto \phi(x,X_i,\lambda_{j_1 j_2})$ is a polynomial function over $\mathbb{R}^p$ (the monomials are combinations of the components $\lambda_{j_1 j_2}$) of degree no more than ℓ. To each partition $(P_1,P_2) \in \mathcal{P}_1 \times \mathcal{P}_2$ and parameter vector $\underline{\lambda} = (\lambda_{j_1 j_2} : (j_1,j_2) \in J)$ we may now associate the corresponding variable bandwidth estimate $f_{n,\theta}(x)$ defined with $\theta = (P_1,P_2,\underline{\lambda})$. In other words, for $x \in \mathbb{R}^d$,

$$f_{n,\theta}(x) = \frac{1}{n}\sum_{i=1}^n \sum_{(j_1,j_2)\in J} \frac{\mathbf{1}_{[\phi(x,X_i,\lambda_{j_1 j_2})>0]}}{\phi(x,X_i,\lambda_{j_1 j_2})} K\Big(\frac{x-X_i}{\phi(x,X_i,\lambda_{j_1 j_2})}\Big) \mathbf{1}_{B_{j_1}^1 \times B_{j_2}^2}(x,X_i)\,, \quad (2.4)$$

with the usual convention $0\times\infty = 0$. Observe that $f_{n,\theta}$ is not a bona fide density since it usually fails to integrate to one. Note also that we require the functions

ϕ to be polynomial in their parameters $\lambda_{j_1 j_2}$ *only*. This allows us to deal with a large choice of bandwidth models, see the examples in Section 2.4.

Now, we can use the combinatorial method described in Section 2.2 to select θ from the set

$$\Theta = \left\{(P_1, P_2, \underline{\lambda}) : P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2, \underline{\lambda} \in \mathbb{R}^{r_1 r_2 p}\right\},$$

and we let f_n be the resulting minimum distance estimate. To apply (2.3), we merely need to obtain upper bounds for the mth shatter coefficient $\mathbf{S}_{\mathcal{A}_\Theta}(m)$ of the Yatracos class associated with Θ. With a slight abuse of notation we denote by $\mathbf{S}_{\mathcal{P}}(j)$ the jth shatter coefficient of the class of sets $B^1 \times B^2$, where B^1 (*resp.* B^2) is any cell of any partition in $\mathcal{P}_1$ (*resp.* $\mathcal{P}_2$), and, for simplicity, we assume that $K(x) = c\mathbf{1}_{[||x|| \leq 1]}$, where c is an appropriate normalizing factor.

Proposition 2.3.1 *If $\mathcal{A}_\Theta$ is the Yatracos class defined by*

$$\Theta = \left\{(P_1, P_2, \underline{\lambda}) : P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2, \underline{\lambda} \in \mathbb{R}^{r_1 r_2 p}\right\},$$

then

$$\begin{aligned}\mathbf{S}_{\mathcal{A}_\Theta}(m) \leq & 2^{1+(2+4p)r_1 r_2}\, \ell^{4r_1 r_2 p}\big[(m(n-m)+1)(2(n-m)-1)(m+1)\big]^{2r_1 r_2 p} \\ & \times \mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)^{2r_1 r_2}.\end{aligned}$$

Note that the above upper bound is non random. The proof of Proposition 2.3.1 relies on a lemma of Bartlett, Maiorov, and Meir [3]. This lemma bounds the number of distinct sign vectors that can be generated using polynomial functions. However, we emphasize that other types of functional dependencies are feasible. For example, Theorem 8.14, page 124 in Anthony and Bartlett [2] provides a portmanteau result for more general function classes in terms of the number of arithmetic operations required for computing the functions. Combining the result of Proposition 2.3.1 with (2.3) leads to the following performance bound for the minimum distance estimates f_n.

Theorem 2.3.1 *Let $\mathcal{A}_\Theta$ be the Yatracos class defined by*

$$\Theta = \left\{(P_1, P_2, \underline{\lambda}) : P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2, \underline{\lambda} \in \mathbb{R}^{r_1 r_2 p}\right\}.$$

Assume that for every $(a, \lambda) \in \mathbb{R}^d \times \mathbb{R}^p$

$$\int \frac{1}{\phi(x, a, \lambda)} K\Big(\frac{x-a}{\phi(x, a, \lambda)}\Big)\, \mathrm{d}x < \infty.$$

Then, for $m \le n/2$, we have

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \le 5\left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right) \inf_{\theta \in \Theta} \mathbf{E}\left\{ \int |f_{n,\theta} - f| \right\}$$
$$+ 8\left[\frac{\big(1 + (2+4p) r_1 r_2\big) \log 2 + 4 r_1 r_2 p \log \ell + 2 r_1 r_2 \log \mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)}{m} \right.$$
$$\left. + \frac{2 r_1 r_2 p \log\big[(m(n-m)+1)(2(n-m)-1)(m+1)\big]}{m} \right]^{1/2} + \frac{5}{n} \,.$$

Theorem 2.3.1 generalizes to more complex bandwidths a result of Devroye, Lugosi, and Udina [11], who partition the sample and use a different fixed bandwidth in each cell of the partition. The complexity of the class of possible partitions among which we select appears in the bounds. Larger families of partitions offer better flexibility, but it is more difficult to select the best among them. To make the above theorem useful, the classes of partitions $\mathcal{P}_k$ $(k = 1, 2)$ have to be restricted in such a way that

$$\frac{\log \mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)}{m} \to 0 \quad \text{as } n \to \infty \,.$$

Recall that $\mathbf{S}_{\mathcal{P}}(m) \le \mathbf{S}_{\mathcal{P}_1}(m) \mathbf{S}_{\mathcal{P}_2}(m)$, where $\mathbf{S}_{\mathcal{P}_k}(m)$ stands for the mth shatter coefficient of the class of sets which are cells of any partition in $\mathcal{P}_k$ $(k = 1, 2)$. As a simple but important example, consider $d = 1$, and let $\mathcal{P}_k$ be the class containing all partitions of the real line into at most r_k intervals. Then $\mathbf{S}_{\mathcal{P}_k}(m)$ is just the mth shatter coefficient of the class of all intervals, which equals $m(m+1)/2 + 1$. More generally, if $\mathcal{P}_k$ stands for the class of partitions of $\mathbb{R}^d$ into at most r_k rectangles, then the shatter coefficient $\mathbf{S}_{\mathcal{P}_k}(m)$ is known to be bounded by $(m+1)^{2d}$. Of course, other multivariate examples, such as tree or Voronoi partitions, are feasible (see Devroye, Györfi, and Lugosi [8], Chapter 13). In all those standard examples,

$$\log \mathbf{S}_{\mathcal{P}_k}\big(m(n-m)\big) = \mathrm{O}(\log n) \,.$$

Therefore, keeping r_k, p and ℓ fixed, and considering for example the choice $m = n/\log n$, we obtain

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \le 5\left(1 + \mathrm{O}\Big(\frac{1}{\sqrt{\log n}}\Big)\right) \inf_{\theta \in \Theta} \mathbf{E}\left\{ \int |f_{n,\theta} - f| \right\} + \mathrm{O}\left(\frac{\log n}{\sqrt{n}}\right) .$$

Since in most cases of interest, the optimal L_1 error tends to zero much slower than $1/\sqrt{n}$, this inequality means that, asymptotically, the error of the minimum distance estimate stays within a constant factor multiple of the best possible

error. Note also that r_k, p and ℓ are allowed to tend to infinity with n, but should not increase so fast that the second term in the upper bound in Theorem 2.3.1 starts dominating. We will not be concerned with the actual details of the minimization algorithm. We realize however that more work is needed to make the present method computationally feasible.

Of course, most kernel functions used in practice are not naive kernels. However, most kernels can be well approximated by Riemann kernels of the form

$$K(x) = \sum_{i=1}^{k} \alpha_i \mathbf{1}_{A_i}(x)\,,$$

$k < \infty$ and $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$. Thus, at the price of slightly worse constants in the bounds, the results presented in the present paper can be easily adapted to all important kernels. For a complete presentation, we refer the reader to Devroye and Lugosi [10], Chapter 11.

We close this section by exhibiting a collection of densities for which the error commited by the selected variable kernel density estimate is less than the usual kernel rate $\mathrm{O}(n^{-2/5})$. We start with a result of Devroye and Lugosi [9] (Lemma 1). These authors use ideas of Sain and Scott [19] to prove that

$$\sup_{f \in \mathcal{F}_B} \mathbf{E}\left\{ \int |f_{n,h(x)} - f| \right\} \leq \sqrt{\frac{4B}{n}}\,,$$

where $\mathcal{F}_B$ is the class of non decreasing, convex-shaped densities on $[0,1]$ with $\sup_{[0,1]} f(x) \leq B$, and $f_{n,h(x)}$ is the variable kernel density estimate corresponding to the variable bandwidth

$$h(x) = \sup\left\{ z > 0 : f * K_z(x) = f(x) \right\}. \tag{2.5}$$

Here

$$K = \mathbf{1}_{[-\frac{1}{2},\frac{1}{2}]} \quad \text{and} \quad K_z(x) = \frac{1}{z} K\Big(\frac{x}{z}\Big)\,.$$

Let us particularize formula (2.5) to the toy-class of linear densities on $[0,1]$, defined by

$$\mathcal{F} = \left\{ f(x) = ax + 1 - \frac{a}{2} : 0 \leq a \leq 2 \right\}.$$

Working out expression (2.5), one obtains that $h(x)$ falls into the general class of bandwidth functions of the form

$$h(x,\theta) = \sum_{j=1}^{3} \phi(x,\lambda_j) \mathbf{1}_{B_j}(x)\,,$$

where, for $j = 1, 2, 3$ and $\lambda_j = (\lambda_j^1, \lambda_j^2, \lambda_j^3) \in \mathbb{R}^3$,

$$\phi(x, \lambda_j) = \frac{1}{\lambda_j^1 x + 1 - \lambda_j^1/2} + \lambda_j^2 x + \lambda_j^3 \,,$$

and $P = \{B_1, B_2, B_3\}$ belongs to $\mathcal{P}$, the class of partitions of $[0, 1]$ into at most three intervalls.

Using the notation $\underline{\lambda} = \left(\lambda_j^k : 1 \leq j, k \leq 3\right)$ for a generic vector of $\mathbb{R}^6$, we can run the combinatorial method to select θ from the set

$$\Theta = \left\{(P, \underline{\lambda}) : P \in \mathcal{P}, \underline{\lambda} \in \mathbb{R}^6\right\}.$$

Note that the class of considered bandwidth functions is not polynomial in its parameters. Nevertheless our results are still valid in this extended framework (see just before Theorem 2.3.1). With the choice $m = n/\log n$, the selected minimum distance estimate f_n then satisfies the inequality

$$\sup_{f \in \mathcal{F}} \mathbf{E}\left\{ \int |f_n - f| \right\} = \mathrm{O}\left(\frac{\log n}{\sqrt{n}}\right),$$

a much faster rate than the usual rate $\mathrm{O}(n^{-2/5})$.

2.4 Examples

Example 1 Biau and Devroye [4] study the L_1 minimax risk over the multivariate class of bounded block decreasing densities. Extending former results of Birgé [5], [6], these authors show first that a suitable variable kernel estimate with linear varying bandwidth achieves the optimal minimax rate. Second, they exhibit by the present combinatorial method a data-dependent bandwidth of the form

$$h(x) = \sum_{i=0}^{q} a_i x^i \,,$$

and prove that the corresponding variable kernel estimate uniformly adapts over the class of bounded block decreasing densities. Clearly, this model falls into the general definition (2.4): just choose $P = \mathbb{R}^d$ and $\underline{\lambda} = (a_0, \ldots, a_q) \in \mathbb{R}^{q+1}$.

Example 2 Denote by V_d the volume of the unit sphere in $\mathbb{R}^d$ and let k be a positive real number. Terrell and Scott [20] show that the following variable bandwidth

$$h(x, k) = \left(\frac{k}{n V_d f(x)}\right)^{1/d} \tag{2.6}$$

is asymptotically equivalent to the k-nearest neighbor bandwidth presented by Loftsgaarden and Quesenberry [13] (known to perform well as dimensionality increases). The problem is the selection of k. Since f is unknown, it has to be replaced by a pilot estimate computed from an independent sample. One can choose, for example, a fixed kernel estimate $\hat{f}$ designed with a Gaussian kernel and a (fixed) bandwidth selected by a data-driven method (Park and Marron [14]). The resulting estimate is of the general form (2.4): just take $P = \mathbb{R}^d$, $\lambda = k^{1/d}$ and

$$\phi(x, \lambda) = \frac{\lambda}{\left(nV_d \hat{f}(x)\right)^{1/d}} .$$

Example 3 Abramson [1] suggests using a variable bandwidth inversely proportional to the square root of the density at X_i, *i.e.*, $h(X_i) = \alpha f(X_i)^{-1/2}$. This adaptive choice performs well for small sample size and reduces the pointwise bias. We may still replace f by a pilot estimate computed from an independent sample of size q. For simplicity, assume that f is a density on $[0, 1]$ and, in place of f, plug the histogram estimate $\hat{f}$ anchored at 0 using at most r cells defined by the bin width vector $\tilde{h} = (\tilde{h}_1, \ldots, \tilde{h}_r)$, *i.e.*,

$$\hat{f}(x) = \sum_{j=1}^{r} \frac{\tilde{\mu}_q(B_j)}{\tilde{h}_j} \mathbf{1}_{B_j}(x) ,$$

where $B_j = (\sum_{k=1}^{j-1} \tilde{h}_k, \sum_{k=1}^{j} \tilde{h}_k]$ for $j = 1, \ldots, r$, and $\tilde{\mu}_q$ denotes the empirical measure computed from the independent sample. Here, we wish to select α and $\tilde{h}$ with the combinatorial method. In this context, the estimate suggested by Abramson [1] reads

$$f_{n,\theta}(x) = \sum_{i=1}^{n} \sum_{j=1}^{r} \frac{\sqrt{\tilde{\mu}_q(B_j)}}{\alpha \sqrt{\tilde{h}_j}} K\left(\frac{\sqrt{\tilde{\mu}_q(B_j)}(x - X_i)}{\alpha \sqrt{\tilde{h}_j}} \right) \mathbf{1}_{B_j}(X_i) , \tag{2.7}$$

where $\theta = (\alpha, \lambda_1, \ldots, \lambda_r)$, and $\lambda_j = \sqrt{\tilde{h}_j}$. Note that the vector $\tilde{h}$ entirely determines the partition $(B_1, \ldots, B_r)$. Each estimate (2.7) is a member of the family (2.4).

Example 4 Jones [12] proposes to modify Abramson's estimate by considering α as a function of the estimation point. Keeping the same notation as in Example 3, with $\alpha = \alpha(x, \mu)$ polynomial in μ, the corresponding estimates may be written as

$$f_{n,\theta}(x) = \sum_{i=1}^{n} \sum_{j=1}^{r} \frac{\sqrt{\tilde{\mu}_q(B_j)}}{\alpha(x, \mu) \sqrt{\tilde{h}_j}} K\left(\frac{\sqrt{\tilde{\mu}_q(B_j)}(x - X_i)}{\alpha(x, \mu) \sqrt{\tilde{h}_j}} \right) \mathbf{1}_{B_j}(X_i) ,$$

with $\theta = (\mu, \lambda_1, \ldots, \lambda_r)$. The estimate of Jones still falls in the general class of variable bandwidth models (2.4).

2.5 Proofs

The proof of Proposition 2.3.1 will strongly rely on the following lemma. We make use of the function sgn(.) defined by

$$\mathrm{sgn}(x) = \begin{cases} 1 & \text{if} \quad x \geq 0 \\ 0 & \text{if} \quad x < 0\,. \end{cases}$$

Lemma 2.5.1 (Bartlett, Maiorov, and Meire, 1998) *Suppose that $f_1(.)$, $\ldots$, $f_m(.)$ are fixed polynomials of degree at most ℓ in p variables. Then the number of distinct sign vectors*

$$\Big(\mathrm{sgn}\big(f_1(a)\big), \ldots, \mathrm{sgn}\big(f_m(a)\big)\Big)$$

that can be generated by varying $a \in \mathbb{R}^p$ is at most

$$2(2\ell)^p(m+1)^p\,.$$

Proof of Proposition 2.3.1 We introduce the notation $x_1, \ldots, x_{n-m}$ for the sample from $\mathbb{R}^{d(n-m)}$ used in the definition of $f_{n-m,\theta}$. It is deterministic and the bounds below will hold uniformly over all such samples. To compute the shatter coefficient, we will use $y_1, \ldots, y_m$ as the sample from $\mathbb{R}^{dm}$ to be employed. We start as in Lemma 12.3, page 123 of Devroye and Lugosi [10]. For each $\theta \in \Theta$, consider the $m \times (n-m) \times r_1 \times r_2$ array z_θ with current element

$$z_\theta^{(t,i,j_1,j_2)} = \mathbf{1}_{[\phi(y_t,x_i,\lambda_{j_1 j_2})>0]} \mathbf{1}_S\left(\frac{y_t - x_i}{\phi(y_t, x_i, \lambda_{j_1 j_2})}\right) \mathbf{1}_{B^1_{j_1} \times B^2_{j_2}}(y_t, x_i)\,,$$

$t \leq m, i \leq n-m, j_1 \leq r_1$, $j_2 \leq r_2$, and $S = \{x : ||x|| \leq 1\}$. We shall first bound the number of different values the array z_θ can take as θ ranges through Θ. Fix temporarily $j_1 = J_1$ and $j_2 = J_2$, and consider the submatrix $u_\theta^{(J_1,J_2)}$ with current element $z_\theta^{(t,i,J_1,J_2)}$. Since there are $m(n-m)$ different pairs (y_t, x_i), the bit vector

$$(\mathbf{1}_{B^1_{J_1} \times B^2_{J_2}}(y_t, x_i) : t \leq m, i \leq n-m)$$

can take at most $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$ values as θ ranges through Θ. Consequently, the number of different values the matrix $u_\theta^{(J_1,J_2)}$ can take is bounded by the product of $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$ and the number of values the bit vector

$$(\mathbf{1}_{[\phi(y_t,x_i,\lambda_{J_1 J_2})>0]} \mathbf{1}_{[\phi(y_t,x_i,\lambda_{J_1 J_2}) \geq ||y_t - x_i||]} : t \leq m, i \leq n-m)$$

can take as $\lambda_{J_1J_2}$ runs through $\mathbb{R}^p$. Since

$$\mathbf{1}_{[\phi(y_t,x_i,\lambda_{J_1J_2})>0]}\mathbf{1}_{[\phi(y_t,x_i,\lambda_{J_1J_2})\geq||y_t-x_i||]}$$

equals

$$\begin{cases} 1-\mathbf{1}_{[-\phi(y_t,x_i,\lambda_{J_1J_2})\geq 0]} & \text{if } y_t = x_i \\ \mathbf{1}_{[\phi(y_t,x_i,\lambda_{J_1J_2})-||y_t-x_i||\geq 0]} & \text{otherwise,} \end{cases}$$

we conclude that

$$\begin{aligned} &\mathrm{Card}\{u_\theta^{(J_1,J_2)} : \theta\in\Theta\} \\ &\leq \mathrm{Card}\Big\{\big(\mathbf{1}_{[R_1(\lambda_{J_1J_2})\geq 0]},\ldots,\mathbf{1}_{[R_{m(n-m)}(\lambda_{J_1J_2})\geq 0]}\big) : \lambda_{J_1J_2}\in\mathbb{R}^p\Big\}\,\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\,, \end{aligned}$$

where the $m(n-m)$ functions R_k's are defined by

$$R_k(\lambda_{J_1J_2}) = \begin{cases} -\phi(y_t,x_i,\lambda_{J_1J_2}) & \text{if } y_t = x_i \\ \\ \phi(y_t,x_i,\lambda_{J_1J_2}) - ||y_t-x_i|| & \text{otherwise}\,. \end{cases}$$

Since the R_k's are polynomials over $\mathbb{R}^p$ of degree no more than ℓ, Lemma 2.5.1 shows that $u_\theta^{(J_1,J_2)}$ can take at most

$$2(2\ell)^p\,(m(n-m)+1)^p\,\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$$

values. If follows that the array z_θ can take at most

$$\Big[2(2\ell)^p\,(m(n-m)+1)^p\,\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\Big]^{r_1r_2},$$

and, similarly, that

$$\mathrm{Card}\big\{(z_\theta,z_{\theta'}) : (\theta,\theta')\in\Theta^2\big\} \leq \Big[2(2\ell)^p\,(m(n-m)+1)^p\,\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\Big]^{2r_1r_2}.$$

Write now $\mathcal{W} = \big\{(w,w') : (w,w') = (z_\theta,z_{\theta'}) \text{ for some } (\theta,\theta')\in\Theta^2\big\}$. For fixed $(w,w')\in\mathcal{W}$, let $U_{(w,w')}$ denote the collection of all (θ,θ') such that $(z_\theta,z_{\theta'}) = (w,w')$. For $(\theta,\theta')\in U_{(w,w')}$, we will use the following notation:

$$\theta = (P_1,P_2,\underline{\lambda}) \quad\text{and}\quad \theta' = (P_1',P_2',\underline{\lambda}')\,,$$

with

$$P_1 = \{B_1^1,\ldots,B_{r_1}^1\},\ P_2 = \{B_1^2,\ldots,B_{r_2}^2\},\underline{\lambda} = (\lambda_{j_1j_2} : (j_1,j_2)\in J)$$

and

$$P_1' = \{B_1^{'1},\ldots,B_{r_1}^{'1}\},\ P_2' = \{B_1^{'2},\ldots,B_{r_2}^{'2}\},\ \underline{\lambda}' = (\lambda'_{j_1j_2} : (j_1,j_2)\in J).$$

For every $t \leq m$, consider the sets

$$I_t = \left\{(i, j_1, j_2) : z_\theta^{(t,i,j_1,j_2)} \neq 0, i \leq n-m, j_1 \leq r_1, j_2 \leq r_2\right\}$$

and

$$I'_t = \left\{(i, j_1, j_2) : z_{\theta'}^{(t,i,j_1,j_2)} \neq 0, i \leq n-m, j_1 \leq r_1, j_2 \leq r_2\right\}.$$

Observe that y_t belongs to

$$A_{\theta,\theta'} = \left\{x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x)\right\}$$

if and only if

$$\sum_{(i,j_1,j_2)\in I_t} \frac{1}{\phi(y_t, x_i, \lambda_{j_1 j_2})}\, z_\theta^{(t,i,j_1,j_2)} > \sum_{(i,j_1,j_2)\in I'_t} \frac{1}{\phi(y_t, x_i, \lambda'_{j_1 j_2})}\, z_{\theta'}^{(t,i,j_1,j_2)}\,.$$

Within the set $U_{(w,w')}$, the values $z_\theta^{(t,i,j_1,j_2)}$ and $z_{\theta'}^{(t,i,j_1,j_2)}$ are fixed for all t, i, j_1 and j_2. Therefore, since the x_i's and y_t's are fixed,

$$\sum_{(i,j_1,j_2)\in I_t} \frac{1}{\phi(y_t, x_i, \lambda_{j_1 j_2})}\, z_\theta^{(t,i,j_1,j_2)}$$

may be written in the form

$$\frac{P_t(\underline{\lambda})}{Q_t(\underline{\lambda})}\,,$$

where the functions $\underline{\lambda} \mapsto P_t(\underline{\lambda})$ (*resp.* $\underline{\lambda} \mapsto Q_t(\underline{\lambda})$) are polynomials over $\mathbb{R}^{r_1 r_2 p}$ of degree no more than $(n-m-1)\ell$ (*resp.* $(n-m)\ell$). Similarly, the quantity

$$\sum_{(i,j_1,j_2)\in I'_t} \frac{1}{\phi(y_t, x_i, \lambda'_{j_1 j_2})}\, z_{\theta'}^{(t,i,j_1,j_2)}$$

may be written in the form

$$\frac{P'_t(\underline{\lambda}')}{Q'_t(\underline{\lambda}')}\,,$$

where P'_t (*resp.* Q'_t) are polynomials over $\mathbb{R}^{r_1 r_2 p}$ of degree no more than $(n-m-1)\ell$ (*resp.* $(n-m)\ell$). Set now $\tilde{\underline{\lambda}} = (\underline{\lambda}, \underline{\lambda}') \in \mathbb{R}^{2r_1 r_2 p}$ and, for $t = 1, \ldots, m$, define

$$\mathcal{R}_t(\tilde{\underline{\lambda}}) = P_t(\underline{\lambda}) Q'_t(\underline{\lambda}') - P'_t(\underline{\lambda}') Q_t(\underline{\lambda})\,.$$

Observe that each $\mathcal{R}_t$ is a polynomial function over $\mathbb{R}^{2r_1 r_2 p}$ of degree no more than $(2(n-m)-1)\ell$. Therefore, applying again Lemma 2.5.1 we obtain

$$\begin{aligned}
&\mathrm{Card}\left\{\left\{\mathbf{1}_{[y_1 \in A_{\theta,\theta'}]}, \ldots, \mathbf{1}_{[y_m \in A_{\theta,\theta'}]}\right\} : (\theta, \theta') \in U^2_{(w,w')}\right\} \\
&\qquad \leq \mathrm{Card}\left\{\left\{\mathbf{1}_{[\mathcal{R}_1(\tilde{\underline{\lambda}})>0]}, \ldots, \mathbf{1}_{[\mathcal{R}_m(\tilde{\underline{\lambda}})>0]}\right\} : \tilde{\underline{\lambda}} \in \mathbb{R}^{2r_1 r_2 p}\right\} \\
&\qquad \leq 2(2\ell)^{2r_1 r_2 p} \left((2(n-m)-1)(m+1)\right)^{2r_1 r_2 p}.
\end{aligned}$$

Putting all pieces together, we obtain

$$\begin{aligned}\mathbf{S}_{\mathcal{A}_\Theta}(m) &\leq 2(2\ell)^{2r_1r_2p}\big((2(n-m)-1)(m+1)\big)^{2r_1r_2p}\operatorname{Card}\mathcal{W}\\ &\leq 2^{1+(2+4p)r_1r_2}\,\ell^{4r_1r_2p}\big[(m(n-m)+1)(2(n-m)-1)(m+1)\big]^{2r_1r_2p}\\ &\qquad\times \mathbf{S}_{\mathcal{P}}(m(n-m))^{2r_1r_2}\,.\end{aligned}$$

■

Bibliography

[1] I. Abramson. On bandwidth variation in kernel estimates. *The Annals of Statistics*, 10:1217–1223, 1982.

[2] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.

[3] P. L. Bartlett, V. Maiorov, and R. Meir. Almost linear VC-dimension bounds for piecewise polynomial networks. *Neural Computation*, 10:2159–2173, 1998.

[4] G. Biau and L. Devroye. On the risk of estimates for block decreasing densities. *Journal of Multivariate Analysis*, 86:143–165, 2003.

[5] L. Birgé. Estimating a density under order restrictions: nonasymptotic minimax risk. *The Annals of Mathematical Statistics*, 15:995–1012, 1987a.

[6] L. Birgé. On the risk of histograms for estimating decreasing densities. *The Annals of Mathematical Statistics*, 15:1013–1022, 1987b.

[7] L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19:135–144, 1977.

[8] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer–Verlag, New-York, 1996.

[9] L. Devroye and G. Lugosi. Variable kernel estimates: On the impossibility of tunning the parameters. In E. Giné and D. Mason, editors, *High-Dimensional Probability II*, pages 405–424. Springer–Verlag, New York, 2000.

[10] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer–Verlag, New York, 2001.

[11] L. Devroye, G. Lugosi, and F. Udina. Inequalities for a new data-based method for selecting nonparametric density estimates. In Madan L. Puri, editor, *Asymptotics in Statistics and Probability: Papers in Honor of George*

Gregory Roussas, pages 133–154. VSP International Science Publishers, Zeist, The Netherlands, 2000.

[12] M. C. Jones. Variable kernel density estimates and variable kernel density estimates. *Australian Journal of Statistics*, 32:361–371, 1990.

[13] D. O. Loftsgaarden and C. P. Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36:1049–1051, 1965.

[14] B. U. Park and J. S. Marron. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, 85(409):66–72, 1990.

[15] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[16] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27:832–837, 1956.

[17] S. R. Sain. Multivariate locally adaptive density estimators. *Computational Statistics and Data Analysis*, 39:165–186, 2002.

[18] S. R. Sain and D. W. Scott. On locally adaptive density estimation. *Journal of the American Statistical Association*, 436:1525–1534, 1996.

[19] S. R. Sain and D. W. Scott. Zero-bias locally adaptive density estimators. *Scandinavian Journal of Statistics*, 29:431–450, 2002.

[20] G. R. Terrell and D. W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20:1236–1265, 1992.

[21] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[22] Y. G. Yatracos. Rates of convergence of minimun distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13:768–774, 1985.

Troisième partie

Classification de Courbes

Chapter 1

Functional Classification with Wavelets*

Abstract

Let X be a random variable taking values in a Hilbert space and let Y be a random label with values in $\{0,1\}$. Given a collection of classification rules and a learning sample of independent copies of the pair (X,Y), we show how to select optimally a classifier and derive its consistency properties. As a general strategy, we first expand the learning sample data on a wavelet basis and reduce the overall infinite dimension to a finite one via a suitable data-dependent thresholding. Then, a finite-dimensional classification rule is performed on the non-zero coefficients. Both the dimension and the classifier are automatically selected by data-splitting and empirical risk minimization. Applications of this technique to a signal discrimination problem involving speech recordings and simulated data are presented. Finally, it is shown that the consistency of the method is preserved by using a suitable sampling scheme.

1.1 Introduction

1.1.1 Functional classification

The problem of classification (or pattern recognition or discrimination) is about guessing or predicting the unknown class of an observation. An observation is usually a collection of numerical measurements represented by a d-dimensional

*Article écrit en collaboration avec Alain Berlinet et Gérard Biau, et soumis à la revue *IEEE Transactions on Information Theory.*

vector. However, in many real-life problems, input data are in fact (sampled) functions rather than standard high dimensional vectors, and this casts the classification problem into the class of Functional Data Analysis.

The last few years have witnessed important new developments in both the theory and practice of functional classification and related learning problems. Nonparametric techniques have been proved useful for analyzing such functional data, and the literature is growing at a fast pace: Hastie, Buja and Tibshirani [17] set out the general idea of Functional Discriminant Analysis (these authors make use of a roughness penalty approach to regularization, and they illustrate their method with examples in speech recognition and handwritten character recognition); Kulkarni and Posner [19] study rates of convergence of k-nearest neighbor regression estimates in general spaces; Hall, Poskitt and Presnell [16] employ a functional data-analytic method for dimension reduction based on Principal Component Analysis and perform Quadratic Discriminant Analysis on the reduced space, so do Ramsay and Silverman [24], [25]; Ferraty and Vieu [15] estimate nonparametrically the posterior probability of an incoming curve in a given class, whereas Rossi and Villa [26] investigate the use of Support Vector Machines in the context of Functional Data Analysis.

Although standard pattern recognition techniques appear to be feasible, the intrinsic infinite-dimensional structure of the observations makes learning suffer from the curse of dimensionality (see Abraham, Biau and Cadre [1] for a detailed discussion, examples and counterexamples). In practice, before applying any learning technique to model real data, a preliminary dimension reduction or model selection step reveals crucial for appropriate smoothing and circumvention of the dimensionality effect. As a matter of fact, filtering is a popular dimension reduction method in signal processing and this is the central approach we take in this paper.

Roughly, filtering reduces the infinite dimension of the observations by considering only the first d coefficients of the data expanded on an appropriate basis. This approach was followed by Kirby and Sirovich [18], Comon [8], Belhumeur, Hepana and Kriegman [3], Hall, Poskitt and Presnell [16], or Amato, Antoniadis and De Feis [2], among others. Given a collection of functions we wish to classify, Biau, Bunea and Wegkamp [4] propose to use first Fourier filtering on each signal, and then perform k-nearest neighbor classification in $\mathbb{R}^d$. These authors study finite sample and asymptotic properties of a data-driven procedure that selects simultaneously both the dimension d and the optimal number of neighbors k.

The aim of the present paper is to extend the data-based filtering approach of Biau, Bunea and Wegkamp [4] to wavelet bases and to more general discrimination rules. Our motivation is twofold.

- First, as pointed out for example in Amato, Antoniadis and De Feis [2], wavelet bases offer some significant advantages over other bases. Indeed, wavelets can be used successfully for compression of a stochastic process, in the sense that the sample paths can be accurately reconstructed from a fraction of the full set of wavelet coefficients. Further, the wavelet decomposition of the sample paths is a local one, so that if the information relevant to the classification problem is contained in a particular part (or parts) of the sample functions, as typically it is, this information will be carried by a very small number of wavelet coefficients. Moreover, the ability of wavelets to model the signal at different levels of resolution means that we have the option of selecting from the paths in a range of bandwidths.

- Second, we seek for general performance bounds and consistency results when using (finite-dimensional approximations of) the sample data in the selection of a discrimination rule and/or its parameters. This article offers both a practical methodology and general performance results for all those who are willing to use wavelet filtering as a dimension reduction step before effective classification.

Throughout the manuscript, we will adopt the point of view of automatic pattern recognition described, to a large extent, in Devroye [12]. In this setup, one uses a test sequence to select the best rule from a rich class of discrimination rules defined in terms of a training sequence. For the clarity of the paper, all important concepts and inequalities regarding this classification paradigm are summarized in the next paragraph. In Section 2, we outline the method and state its finite sample performance and consistency. Section 3 offers some experimental results both on real-life and simulated data. In practice, due to the limited precision of the measuring instruments, it is clearly impossible to observe the functional data continuously. In other words, the problem of classification for *sampled data* is of extreme interest in practical issues. This problem will be discussed in Section 5.

1.1.2 Automatic pattern recognition

This section gives a brief exposition and sets up terminology of automatic pattern recognition. For a detailed introduction, the reader is referred to Devroye [12].

To model the automatic learning problem, we introduce a probabilistic setting. Denote by $\mathcal{F}$ some abstract Hilbert space, and keep in mind that the choice $\mathcal{F} = L_2([0,1])$ (that is, the space of all square integrable functions on $[0,1]$) will be a leading example throughout the paper. The data consist of a sequence of $n+m$ i.i.d. $\mathcal{F} \times \{0,1\}$-valued random variables $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$.

The X_i's are the *observations*, and the $Y_i's$ are the *labels*[†]. Note that we artificially split the data into two independent sequences, one of length n, and one of length m: we call the n sequence the *training sequence*, and the m sequence the *testing sequence*. A discrimination rule is a (measurable) function $g : \mathcal{F} \times (\mathcal{F} \times \{0,1\})^{n+m} \to \{0,1\}$. It classifies a new observation $x \in \mathcal{F}$ as coming from class $g\left(x, (X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})\right)$. We will write $g(x)$ for the sake of convenience.

The probability of error of a given rule g is

$$L_{n+m}(g) = \mathbf{P}\big\{g(X) \neq Y | (X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})\big\},$$

where (X, Y) is independent of the data sequence and is distributed as (X_1, Y_1). Although we would like $L_{n+m}(g)$ to be small, we know (see e.g. Devroye, Györfi and Lugosi [13], Theorem 2.1, page 10) that $L_{n+m}(g)$ cannot be smaller than the Bayes probability of error

$$L^* = \inf_{s:\mathcal{F}\to\{0,1\}} \mathbf{P}\{s(X) \neq Y\}.$$

In the learning process, we aim at constructing rules with probability of error as close as possible to L^*. To do this, we employ the learning sequence to design a class of data-dependent discrimination rules and we use the testing sequence as an impartial judge in the selection process. More precisely, we denote by $\mathbf{D}_n$ a (possibly infinite) collection of functions $g : \mathcal{F} \times (\mathcal{F} \times \{0,1\})^n \to \{0,1\}$, from which a particular function $\hat{g}$ is selected by minimizing the *empirical risk* based upon the testing sequence:

$$\hat{L}_{n,m}(\hat{g}) = \frac{1}{m} \sum_{i=n+1}^{n+m} \mathbf{1}_{[\hat{g}(X_i) \neq Y_i]} = \min_{g \in \mathbf{D}_n} \frac{1}{m} \sum_{i=n+1}^{n+m} \mathbf{1}_{[g(X_i) \neq Y_i]}.$$

At this point, observe that in the formulation above, for $x \in \mathcal{F}$,

$$g(x) = g\big(x, (X_1, Y_1), \ldots, (X_n, Y_n)\big)$$

and

$$\hat{g}(x) = \hat{g}\big(x, (X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})\big),$$

i.e., the discriminators g themselves are based upon the training sequence only, whereas the chosen classifier $\hat{g}$ depends on the *entire data set*, as the rest of the data is used for selecting the classifiers.

[†]In this study we restrict our attention to binary classification. The reason is simplicity and that the binary problem already captures many of the main features of more general problems.

Since, conditionally to the training sequence, $\hat{L}_{n,m}(g)$ is an unbiased estimate of $L_n(g)$, we expect that $L_{n+m}(\hat{g})$ is close to $\inf_{g\in\mathbf{D}_n} L_n(g)$. This is captured in the following inequality (see Devroye, Györfi and Lugosi [13], Lemma 8.2, page 126):

$$L_{n+m}(\hat{g}) - \inf_{g\in\mathbf{D}_n} L_n(g) \le 2 \sup_{g\in\mathbf{D}_n} \left|\hat{L}_{n,m}(g) - L_n(g)\right|. \tag{1.1}$$

Thus, upper bounds for $\sup_{g\in\mathbf{D}_n} |\hat{L}_{n,m}(g) - L_n(g)|$ provide us with upper bounds for the suboptimality of $\hat{g}$ within $\mathbf{D}_n$. When the class of rules $\mathbf{D}_n$ is finite with (random) cardinality bounded by N_n, upper bounds can be obtained via a direct application of Hoeffding's inequality:

$$\mathbf{E}_n\left\{\sup_{g\in\mathbf{D}_n}\left|\hat{L}_{n,m}(g) - L_n(g)\right|\right\} \le \sqrt{\frac{\log(2N_n)}{2m}} + \frac{1}{\sqrt{8m\log(2N_n)}}, \tag{1.2}$$

where the notation $\mathbf{E}_n$ means the expectation conditional on the training sequence of length n. The inequality above is useless when $N_n = \infty$. It is here that we can apply the inequality of Vapnik and Chervonenkis [29] or one of its modifications. We first need some more notation. For fixed training sequence $(x_1, y_1), \ldots, (x_n, y_n)$, denote by $\mathbf{C}_n$ the collection of all sets

$$\mathbf{C}_n = \Big\{\{x \in \mathcal{F} : g(x) = 1\} : g \in \mathbf{D}_n\Big\},$$

and define the shatter coefficient as

$$\mathbb{S}_{\mathbf{C}_n}(m) = \max_{(x_1,\ldots,x_m)\in\mathcal{F}^m} \text{Card}\left\{\{x_1,\ldots,x_m\} \cap C : C \in \mathbf{C}_n\right\}.$$

Then

$$\mathbf{E}_n\left\{\sup_{g\in\mathbf{D}_n}\left|\hat{L}_{n,m}(g) - L_n(g)\right|\right\} \le \sqrt{\frac{8\log\left(4\mathbb{S}_{\mathbf{C}_n}(2m)\right)}{m}} + \frac{1}{\sqrt{(m/2)\log\left(4\mathbb{S}_{\mathbf{C}_n}(2m)\right)}}. \tag{1.3}$$

For more information and improvements on these inequalities, we refer the reader to the monograph of Devroye, Györfi and Lugosi [13], and to the comprehensive surveys of Boucheron, Bousquet and Lugosi [5], [6].

1.2 Dimension reduction for classification

The theory of wavelets has recently undergone a rapid development with exciting implications for nonparametric estimation. Wavelets are functions that can cut

up a signal into different frequency components with a resolution matching its scale. Unlike the traditional Fourier bases, wavelet bases offer a degree of localization in space as well as frequency. This enables development of simple function estimates that respond effectively to discontinuities and spatially varying degree of oscillations in a signal, even when the observations are contaminated by noise. The books of Daubechies [11], Meyer [21], Mallat [20] and Cohen [7] give detailed expositions of the mathematical aspects of wavelets.

As for now, to avoid useless technical notation, we will suppose that the feature space $\mathcal{F}$ is equal to the Hilbert space $L_2([0,1])$, and we will sometimes refer to the observations X_i as "the curves". Extension to more general Hilbert spaces is routine and is left to the reader. We recall that $L_2([0,1])$ is approximated by a *multiresolution analysis*, i.e., a ladder of closed subspaces

$$V_0 \subset V_1 \subset \ldots \subset L_2\big([0,1]\big)$$

whose union is dense in $L_2([0,1])$, and where each V_j is spanned by 2^j orthonormal scaling functions $\phi_{j,k}, k = 0, \ldots, 2^j - 1$, such that $\mathrm{supp}(\phi_{j,k}) \subset [k2^{-j}, (k+1)2^{-j}]$. At each resolution level $j \geq 0$, the orthonormal complement W_j between V_j and V_{j+1} is generated by 2^j orthonormal wavelets $\psi_{j,k}, k = 0, \ldots, 2^j - 1$. Thus, the family

$$\bigcup_{j\geq 0} \{\psi_{j,k}\}_{k=0,\ldots,2^j-1}$$

completed by $\{\phi_{0,0}\}$ forms an orthonormal basis of $L_2([0,1])$. As a consequence, any observation X in $L_2([0,1])$ reads

$$X(t) = \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \zeta_{j,k}\psi_{j,k}(t) + \eta\phi_{0,0}(t), \quad t \in [0,1], \tag{1.4}$$

where

$$\zeta_{j,k} = \int_0^1 X(t)\psi_{j,k}(t)\mathrm{d}t \quad \text{and} \quad \eta = \int_0^1 X(t)\phi_{0,0}(t)\mathrm{d}t,$$

and the consistency (1.4) is understood in $L_2([0,1])$. We are now ready to introduce our classification algorithm and discuss its consistency properties. Using the notation of Paragraph 1.1.2, we suppose that the data consist of a sequence of $n+m$ i.i.d. $L_2([0,1])\times\{0,1\}$-valued random observations $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$. Given a multiresolution analysis of $L_2([0,1])$ as explicited above, each observation X_i is expressed as a series expansion

$$X_i(t) = \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \zeta_{j,k}^i\psi_{j,k}(t) + \eta^i\phi_{0,0}(t), \quad t \in [0,1]. \tag{1.5}$$

For the sake of coherence, it will be convenient to reindex the sequence $\{\phi_{0,0}, \psi_{0,0}, \psi_{1,0}, \psi_{1,1}, \psi_{2,0}, \psi_{2,1}, \psi_{2,2}, \psi_{3,0}, ...\}$ into $\{\psi_1, \psi_2, ...\}$. With this scheme, expression (1.5) may be rewritten as

$$X_i(t) = \sum_{j=1}^{\infty} X_{ij}\psi_j(t), \quad t \in [0,1], \tag{1.6}$$

with the random coefficients

$$X_{ij} = \int_0^1 X_i(t)\psi_j(t)\,\mathrm{d}t.$$

Denote by $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots)$ the sequence of coefficients associated with X_i. Recall that the Hilbert space $L_2([0,1])$ is isomorphic to $\ell_2 = \left\{\mathbf{x} = (x_1, x_2, \ldots) : \sum_{j=1}^{\infty} x_j^2 < \infty\right\}$. Consequently, knowing X_i is the same as knowing $\mathbf{X}_i$. In our quest of dimension reduction, we first fix in (1.5) a maximum (large) resolution level J ($J \geq 0$, possibly function of n) so that

$$X_i(t) \approx \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \zeta_{j,k}^i \psi_{j,k}(t) + \eta^i \phi_{0,0}(t), \quad t \in [0,1]$$

or equivalently, using (1.6),

$$X_i(t) \approx \sum_{j=1}^{2^J} X_{ij}\psi_j(t), \quad t \in [0,1].$$

At this point, we could try to use these finite-dimensional approximations of the observations, and let the data select optimally one of the $2^{2^J} - 1$ non-empty subbases of $\{\psi_1, \ldots, \psi_{2^J}\}$. By doing so, we would be faced with an unreasonable overall algorithmic complexity, and therefore catastrophic subsequent performance bounds. Thus, in order to reduce the overall complexity of the problem, we suggest the following procedure.

First, for each $d = 1, \ldots, 2^J$, we assume to be given beforehand a (possibly infinite) collection $\mathbf{D}_n^{(d)}$ of rules $g^{(d)} : \mathbb{R}^d \times (\mathbb{R}^d \times \{0,1\})^n \to \{0,1\}$ working in $\mathbb{R}^d$ and using the n d-dimensional training data as input. We will denote by $\mathbb{S}_{\mathbf{C}_n^{(d)}}(m)$ the corresponding shatter coefficients (see Paragraph 1.1.2) and, with a slight abuse of notation, by $\mathbb{S}_{\mathbf{C}_n}^{(J)}(m)$ the shatter coefficient corresponding to the collection $\cup_{d=1}^{2^J}\mathbf{D}_n^{(d)}$ of all rules embedded in $\mathbb{R}^{2^J}$. Observe that

$$\mathbb{S}_{\mathbf{C}_n}^{(J)}(m) \leq \sum_{d=1}^{2^J} \mathbb{S}_{\mathbf{C}_n^{(d)}}(m). \tag{1.7}$$

Second, we let the n training data reorder the first 2^J basis functions $\{\psi_1, \ldots, \psi_{2^J}\}$ into $\{\psi_{j_1}, \ldots, \psi_{j_{2^J}}\}$ via the scheme

$$\sum_{i=1}^{n} X_{ij_1}^2 \geq \sum_{i=1}^{n} X_{ij_2}^2 \geq \ldots \geq \sum_{i=1}^{n} X_{ij_{2^J}}^2. \tag{1.8}$$

In other words, we just let the training sample decide by itself which basis functions carry the most significant information.

We finish the procedure by a **third** selection step: pick the *effective* dimension $d \leq 2^J$ and a classification rule $g^{(d)}$ in $\mathbf{D}_n^{(d)}$ by approximating each X_i by $\mathbf{X}_i^{(d)} = (X_{ij_1}, \ldots, X_{ij_d})$.

The dimension d and the classifier $g^{(d)}$ are simultaneously selected using the data-splitting device described in Paragraph 1.1.2. Precisely, we choose both d and $g^{(d)}$ optimally by minimizing the empirical probability of error based on the independent validation set, that is

$$\left(\hat{d}, \hat{g}^{(\hat{d})}\right) \in \underset{d=1,\ldots,2^J, g \in \mathbf{D}_n^{(d)}}{\operatorname{argmin}} \left[\frac{1}{m} \sum_{i=n+1}^{n+m} \mathbf{1}_{[g^{(d)}(\mathbf{X}_i^{(d)}) \neq Y_i]}\right]. \tag{1.9}$$

Note that the second step of our algorithm is somewhat related to wavelet shrinkage, that is, certain wavelet coefficients are reduced to zero. Wavelet shrinkage and thresholding methods constitute a powerful way to carry out signal analysis, especially when the underlying process has sparse wavelet representation. They are computationally fast and automatically adapt to spatial and frequency inhomogeneities of the signal. A review of the advantages of wavelet shrinkage appears in Donoho, Johnstone, Kerkyacharian and Picard [14]. In our functional classification context, the preprocessing step (1.8) allows to shrink *globally* all learning data. This point is crucial, as individual shrinkages would lead to different significant bases for each function in the learning set.

Apart from being conceptually simple, this method leads to the classifier $\hat{g}(\mathbf{x}) = \hat{g}^{(\hat{d})}(\mathbf{x}^{(\hat{d})})$ with a probability of misclassification

$$L_{n+m}(\hat{g}) = \mathbf{P}\big\{\hat{g}(\mathbf{X}) \neq Y \,|\, (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_{n+m}, Y_{n+m})\big\},$$

where, for a generic X, $\mathbf{X}^{(d)} = (\mathbf{X}_{j_1}, \ldots, \mathbf{X}_{j_d})$ denotes the first d coefficients reordered via the scheme (1.8). The selected rule $\hat{g}$ satisfies the following optimal inequality, whose proof is clear from (1.1) and (1.3):

Theorem 1.2.1

$$\mathbf{E}\{L_{n+m}(\hat{g})\} - L^* \le L^*_{2^J} - L^* + \mathbf{E}\Big\{ \inf_{d=1,\ldots,2^J, g^{(d)} \in \mathbf{D}_n^{(d)}} L_n(g^{(d)}) \Big\} - L^*_{2^J}$$
$$+ 2\mathbf{E}\left\{ \sqrt{\frac{8 \log\left(4\mathbb{S}^{(J)}_{\mathbf{C}_n}(2m)\right)}{m}} + \frac{1}{\sqrt{(m/2)\log\left(4\mathbb{S}^{(J)}_{\mathbf{C}_n}(2m)\right)}} \right\}.$$

Here

$$L^*_{2^J} = \inf_{s:\mathbb{R}^{2^J} \to \{0,1\}} \mathbf{P}\{s(\mathbf{X}^{(2^J)}) \neq Y\}$$

stands for the Bayes probability of error when the feature space is $\mathbb{R}^{2^J}$.

We may view the first term, $L^*_{2^J} - L^*$, on the right of the inequality as an approximation term – the price to pay for using a finite-dimensional approximation. This term converges to zero by Lemma 1.2.1 below.

Lemma 1.2.1 *We have*

$$L^*_{2^J} - L^* \to 0 \quad \textit{as } J \to \infty.$$

Proof From Devroye, Györfi and Lugosi [13] (Theorem 2.2, page 16), we have, for every $J \ge 1$,

$$L^*_{2^J} - L^* \le 2\mathbf{E}\big|\mathbf{E}\{Y|\mathbf{X}^{(2^J)}\} - \mathbf{E}\{Y|\mathbf{X}\}\big|.$$

Note that $M_J = \mathbf{E}\{Y|\mathbf{X}^{(2^J)}\} = \mathbf{E}\{Y|X_1,\ldots,X_{2^J}\}$ is a uniformly bounded martingale with respect to the natural filtration $\sigma(X_1,\ldots,X_{2^J})$. By the martingale convergence theorem (cf. Pollard [22], Corollary 27, page 151), it follows that M_J converges in L_1 to a limit M_∞, which equals $\mathbf{E}\{Y|\mathbf{X}\}$ (cf. Pollard [22], Theorem 36, page 154). The claim of the lemma follows. ■

The second term, $\mathbf{E}\{\inf_{d=1,\ldots,2^J, g^{(d)} \in \mathbf{D}_n^{(d)}} L_n(g^{(d)})\} - L^*_{2^J}$, can be handled by standard results on classification. Let us first recall the definition of a *consistent* rule: a rule g is consistent for a class of distributions $\mathcal{D}$ if $\mathbf{E}\{L_n(g)\} \to L^*$ as $n \to \infty$ for all distributions $(X,Y) \in \mathcal{D}$.

Corollary 1.2.1 *Let* $\mathcal{D}$ *be a class of distributions. For fixed* $J \ge 0$, *assume that we can pick from each* $\mathbf{D}_n^{(2^J)}$, $n \ge 1$, *one* $g_n^{(2^J)}$ *such that the sequence* $(g_n^{(2^J)})_{n\ge 1}$ *is consistent for* $\mathcal{D}$. *If*

$$\lim_{n\to\infty} m = \infty, \quad \textit{and, for each } J, \quad \lim_{n\to\infty} \mathbf{E}\left\{ \frac{\log \mathbb{S}^{(J)}_{\mathbf{C}_n}(2m)}{m} \right\} = 0,$$

then the automatic rule $\hat{g}$ defined in (1.9) is consistent for $\mathcal{D}$ in the sense

$$\lim_{J\to\infty}\lim_{n\to\infty}\mathbf{E}\big\{L_{n+m}(\hat{g})\big\}=L^*.$$

Proof The proof uses Theorem 1.2.1, Lemma 1.2.1, and the upper bound

$$\mathbf{E}\left\{\inf_{d=1,\ldots,2^J,g^{(d)}\in\mathbf{D}_n^{(d)}}L_n(g^{(d)})\right\}-L_{2^J}^*\le\mathbf{E}\big\{L_n(g_n^{(2^J)})\big\}-L_{2^J}^*.$$

■

This consistency result is new and is especially valuable since few approximation results have been established for functional classification. Corollary 1.2.1 shows that a consistent rule is selected if, for each fixed $J\ge 0$, the sequence of $\mathbf{D}_n^{(2^J)}$'s contains a consistent rule, even if we do not know which functions from $\mathbf{D}_n^{(2^J)}$ lead to consistency. If we are only concerned with consistency, Corollary 1.2.1 reassures us that nothing is lost as long as we take m much larger than $\log\mathbf{E}\{\mathbb{S}_{\mathbf{C}_n^{(J)}}(2m)\}$. Often, this reduces to a very weak condition on the size m of the testing set. Note also that it is usually possible to find upper bounds on the random variable $\mathbb{S}_{\mathbf{C}_n^{(J)}}(2m)$ that depend on n, m and J, but not on the actual values of the random variables $(X_1,Y_1),\ldots,(X_n,Y_n)$. In this case, the bound is distribution-free, and the problem is purely combinatorial: count $\mathbb{S}_{\mathbf{C}_n^{(J)}}(2m)$. Examples are worked out below.

1.2.1 Examples

Linear discrimination For fixed $d\ge 1$, consider all rules that split $\mathbb{R}^d$ in two by virtue of a half plane, and assign class 1 to one half space, and class 0 to the other. In other words, x is classified in one of the two classes according to whether

$$a_0+\sum_{i=1}^d a_i x^{(i)}$$

is positive or negative, where $x^{(1)},\ldots,x^{(d)}$ denote the components of $x\in\mathbb{R}^d$. Choose one half space and decide that points on the border belong to this half space. Because the training sequence is not even used in the definition of the collection, $\mathbb{S}_{\mathbf{C}_n^{(d)}}(2m)$ does only depend on d and m. According to Cover [9], one has

$$\mathbb{S}_{\mathbf{C}_n^{(d)}}(2m)\le 2\left((2m)^d+1\right),$$

and thus, using (1.7),

$$\mathbb{S}_{\mathbf{C}_n}^{(J)}(2m)\le 2\sum_{d=1}^{2^J}\left((2m)^d+1\right)=2^{J+1}+4m\frac{(2m)^{2^J}-1}{2m-1}$$

Here the algorithm uses the training sequence to select the optimal dimension $\hat{d}$ only – the coefficients are adjusted just by minimizing the error committed on the testing sample. The resulting discriminator picks the best $\hat{d}$-dimensional separating hyperplane based upon the testing sample.

k-NN rules In the k-nearest neighbor rule (k-NN), a majority vote decision is made over the labels based upon the k nearest neighbors of x in the training set. This procedure is among the most popular nonparametric methods used in statistical pattern recognition with over 900 research articles published on the method since 1981 alone! Dasarathy [10] has provided a comprehensive collection of around 140 key papers.

If $\mathbf{D}_n^{(d)}$ contains all NN-rules (all values of k) in dimension d, then, unlike the collection of the previous example, $\mathbf{D}_n^{(d)}$ increases with n, and it depends very much on the training set. A trivial bound in this case is

$$\mathbb{S}_{\mathbf{C}_n^{(d)}}(2m) \leq n$$

because there are only n members in $\mathbf{D}_n^{(d)}$. Consequently,

$$\mathbb{S}_{\mathbf{C}_n}^{(J)}(2m) \leq 2^J n.$$

Stone [27] proved the striking result that k-NN classifiers are universally consistent if $X \in \mathbb{R}^d$, provided $k \to \infty$ and $k/n \to 0$. Therefore, we see that our strategy leads to a consistent rule whenever $J/m \to 0$ and $\log n/m \to 0$ as $n \to \infty$. Thus, we can take m equal to a small fraction of n without loosing consistency. Consistent classifiers can also be obtained by other local averaging methods as long as $\mathcal{F} = \mathbb{R}^d$, see e.g. Devroye, Györfi and Lugosi [13]. On the other hand, the story is radically different in general spaces $\mathcal{F}$. Abraham, Biau and Cadre [1] present counterexamples indicating that the moving window rule (Devroye, Györfi and Lugosi [13], Chapter 10) is not consistent for general $\mathcal{F}$, and they argue that restrictions on the space $\mathcal{F}$ (in terms of metric covering numbers) and on the regression function $\eta(x) = \mathbb{E}\{Y|X = x\}$ cannot be given up. By adapting the arguments in Abraham, Biau and Cadre [1], it can be shown that the k-NN classifier is consistent, provided η is continuous on the separable Hilbert space $L_2([0,1])$, $k \to \infty$ and $k/n \to 0$.

Binary tree classifiers Classification trees partition $\mathbb{R}^d$ into regions, often hyperrectangles parallel to the axes. Among these, the most important are the binary classification trees, since they have just two children per node and are thus easiest to manipulate and update. Many strategies have been proposed for

constructing the binary decision tree (in which each internal node corresponds to a cut, and each terminal node corresponds to a set in the partition). For examples and list of references, we refer the reader to Devroye, Györfi and Lugosi [13], Chapter 20.

If we consider for example all binary trees in which each internal node corresponds to a split perpendicular to one of the axes, then

$$\mathbb{S}_{\mathbf{C}_n^{(d)}}(2m) \leq \big(1 + d(n+2m)\big)^k,$$

where k is the maximum number of consecutive orthogonal cuts (or internal nodes). Therefore,

$$\mathbb{S}^{(J)}_{\mathbf{C}_n}(2m) \leq \sum_{d=1}^{2^J} \big(1 + d(n+2m)\big)^k \leq 2^J \big(1 + 2^J(n+2m)\big)^k.$$

1.3 Applications

In this section, we propose to illustrate the performance of the wavelet classification algorithm presented in Section 1.2. The method has been tested with three collections of rules $\mathbf{D}_n^{(d)}$ performing in the finite-dimensional space $\mathbb{R}^d$. We will use the following acronyms:

- W-QDA when $\mathbf{D}_n^{(d)}$ consists of the Quadratic Discriminant Analysis rule performed in dimension d.
- W-NN when $\mathbf{D}_n^{(d)}$ consists of all d-dimensional nearest-neighbor classifiers.
- W-T when $\mathbf{D}_n^{(d)}$ contains binary trees in $\mathbb{R}^d$ in which each internal node corresponds to a split perpendicular to one of the axes.

In addition, our functional classification methodology is compared with two alternative approaches:

- F-NN refers to the Fourier filtering approach combined with the k-NN rule studied in Biau, Bunea and Wegkamp [4]. In this method, the k-NN discrimination rule is performed on the first d coefficients of a Fourier series expansion of each curve. The effective dimension d and the number of neighbors k are selected by minimizing the empirical probability of error based on the testing sequence plus an additive penalty term $\lambda_d/\sqrt{m}$ which avoids overfitting. We choose the penalty term as suggested by the authors, namely $\lambda_d = 0$ for $d \leq n$ and $\lambda_d = \infty$ for $d > n$.

- MPLSR refers to the Multivariate Partial Least Square Regression for functional classification. This approach is studied in detail in Preda and Saporta [23]. The number of PLS components is selected by minimizing the empirical probability of error based on the testing sequence.

1.3.1 Speech recognition

We first tested the method in a speech recognition problem. We study a part of TIMIT database which was investigated in Hastie, Buja and Tibshirani [17]. The data are log-periodograms corresponding to recording phonemes of 32 ms duration. We are concerned with the discrimination of five speech frames corresponding to five phonemes transcribed as follows: "sh" as in "she" (872 items), "dcl" as in "dark" (757 items), "iy" as the vowel in "she" (1163 items), "aa" as the vowel in "dark" (695 items) and "a0" as the first vowel in "water" (1022 items). The database is a multispeaker database. Each speaker is recorded at a 16 kHz sampling rate and we retain only the first 256 frequencies (see Figure 1.1). Thus the data consists of 4509 series of length 256 with known class word membership.

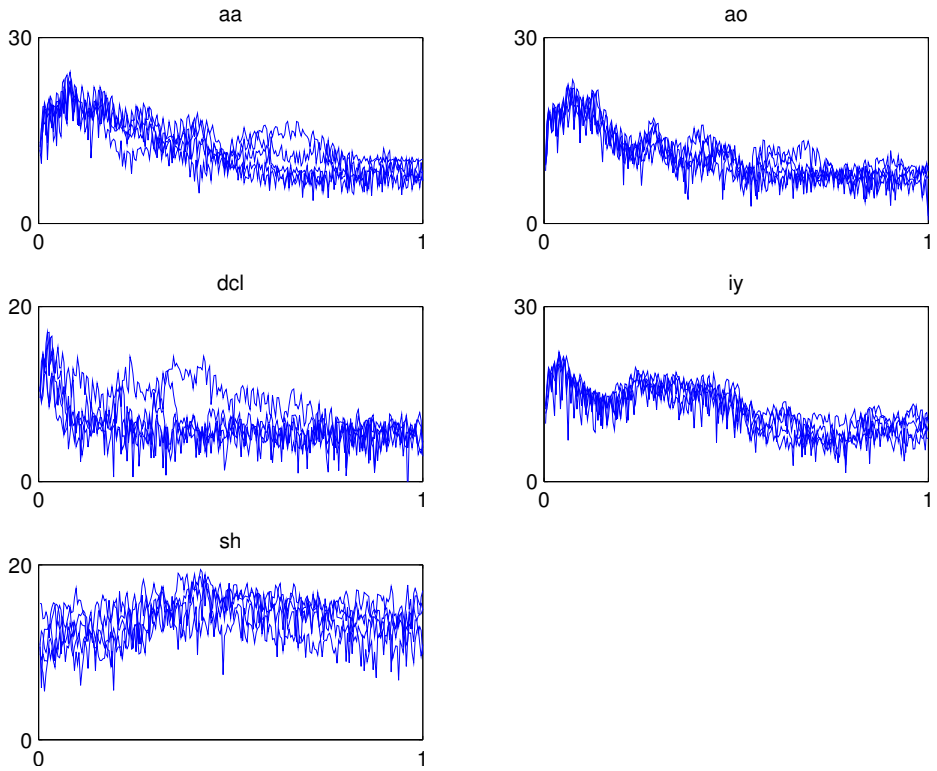

Figure 1.1: A sample of 5 log-periodograms, one in each class.

We decided to retain 250 observations for training and 250 observations for testing. The error rate (**ER**) of the elected rule $\hat{g}$ for classifying new observations is unknown, but it can be estimated consistently using the rest of the data $(X_{501}, Y_{501}), \ldots, (X_{4509}, Y_{4509})$, via the formula

$$\mathbf{ER} = \frac{1}{4009} \sum_{i=501}^{4509} \mathbf{1}_{[\hat{g}(X_i) \neq Y_i]}.$$

Table 1.1 displays the estimated error rates for the different methods together with the selected dimensions (number of PLS components for MPLSR). All results are averaged over 50 random partitions of the data.

Method	**ER**	$\hat{d}$
W-QDA	0.1042	7.30
W-NN	0.1096	19.52
W-T	0.1253	9.10
F-NN	0.1277	48.76
MPLSR	0.0904	5.96

Table 1.1: Estimated error rates and selected dimensions.

Table 1.1 shows that the three methods using wavelets perform well on the data at hand. The method MPLSR seems a really competitive procedure in regards with the others, and the results of the Fourier-based algorithm are still acceptable. However, the performance of the latter methods can considerably deteriorate for time/frequencies inhomogeneous signals, as illustrated in the next paragraph.

1.3.2 A small simulation study

We propose to investigate the performance of our method in the following simulated scenario. For each $i = 1, \ldots, n$, we generate pairs $(X_i(t), Y_i)$ via the scheme:

$$X_i(t) = \frac{1}{50}\Big(\sin(F_i^1 \pi t) f_{\mu_i,\sigma_i}(t) + \sin(F_i^2 \pi t) f_{1-\mu_i,\sigma_i}(t) \Big) + \varepsilon_i,$$

where

- $f_{\mu,\sigma}$ stands for the normal density with mean μ and variance σ^2;
- F_i^1 and F_i^2 are uniform random variables on $[50, 150]$;

- μ_i is randomly uniform on $[0.1, 0.4]$;
- σ_i is randomly uniform on $[0, 0.005]$;
- The ε_i's are mutually independent normal random variables with mean 0 and standard deviation 0.5.

The label Y_i associated to X_i is then defined, for $i = 1, \ldots, n$, by

$$Y_i = \begin{cases} 0 & \text{if } \mu_i \leq 0.25 \\ 1 & \text{otherwise.} \end{cases}$$

Figure 1.2 displays six typical realizations of the X_i's. We see that each curve $X_i(t)$, $t \in [0, 1]$, is composed of two different but symmetric signals, and the problem is thus to detect if the two signals are close (label 0) or enough distant (label 1).

All the algorithms were tested over samples of size 50 for learning and 50 for testing. The error rates (**ER**) were estimated on independent samples of size 500. Figure 1.3 and Figure 1.4 display the boxplots of the estimated error rates and dimensions for 50 replications of the different methods under study. Table 1.2 summarizes the results.

Method	**ER**	$\hat{d}$
W-QDA	0.0849	8.16
W-NN	0.1255	23.92
W-T	0.1275	20.48
F-NN	0.3493	75.88
MPLSR	0.4413	3.68

Table 1.2: Estimated error rates and selected dimensions.

Table 1.2 enlightens the good results achieved by the wavelet classification algorithms. We note in particular the excellent performance of W-QDA which achieves, in average, the best rates. The rather poor results obtained by the method F-NN are not surprising. Due to the penalty term ($\lambda_d = 0$ for $d \leq n$ and $\lambda_d = \infty$ for $d > n$), this procedure retains only the first n coefficients of the Fourier expansion. This maximal number of coefficients is definitely too low here since frequencies of the two signals can typically approach 150 Hz. The problem

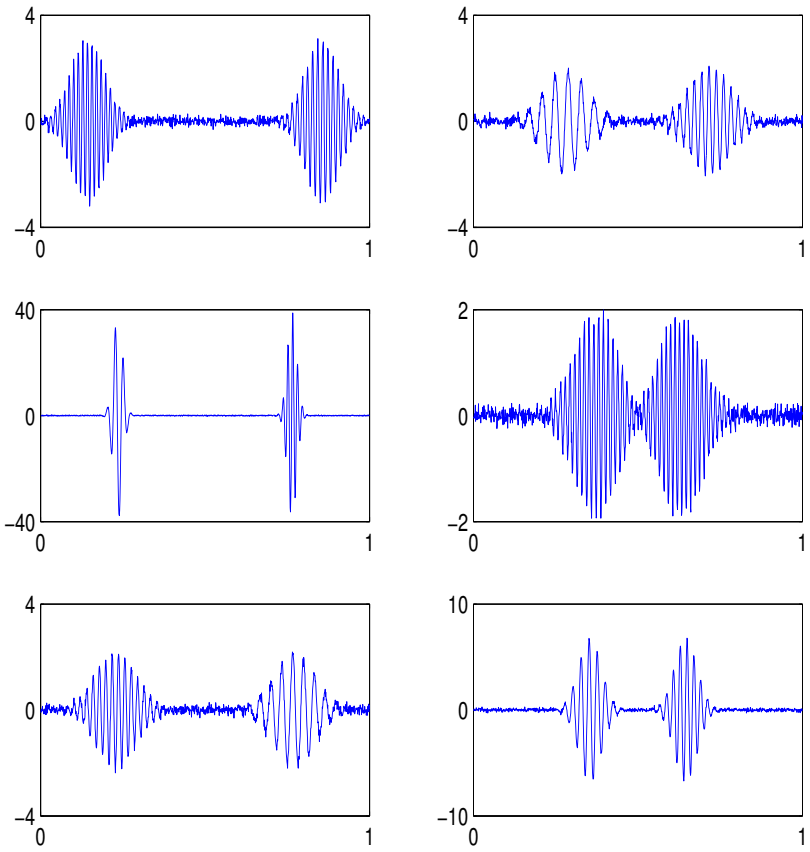

Figure 1.2: Six typical realizations of simulated curves with label 1 (left) and label 0 (right).

of the calibration of the penalty term is discussed in detail in Biau, Bunea and Wegkamp [4] and Tuleau [28].

Finally, Figure 1.4 strongly supports the idea that our wavelet approach allows to considerably reduce the dimension. For example, the typical dimensions selected by the algorithm W-NN are around 20, whereas there are over 70 for the Fourier procedure.

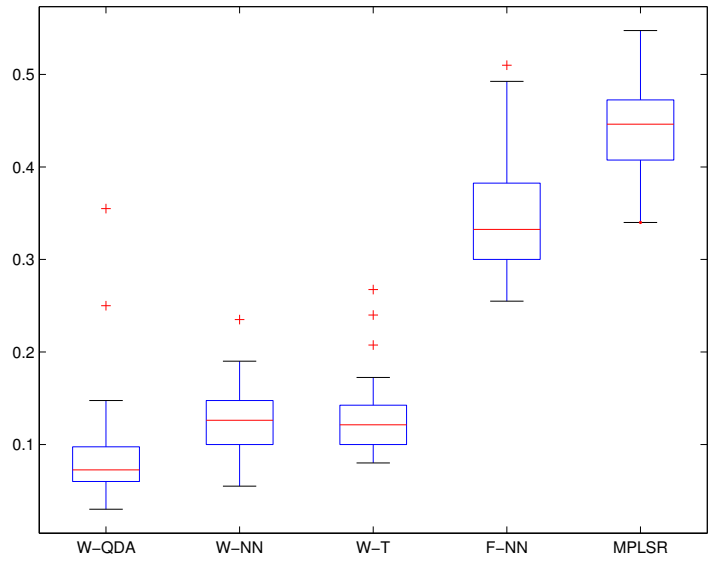

Figure 1.3: Boxplots of the estimated error rates for 50 replications of the tested methods.

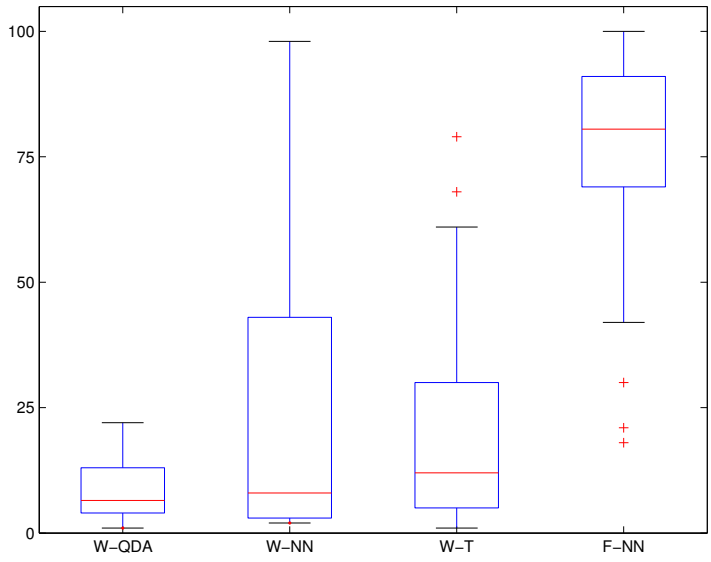

Figure 1.4: Boxplots of the selected dimensions for 50 replications of the tested methods.

1.4 Sampled data classification

In all the previous sections, we implicitly assumed that each process X_i was completely observed over the interval $[0,1]$. In other words, all learning curves were, in some sense, "perfectly" known. In practice, however, it is clearly impossible to observe the curves continuously, and each process is observed only at a *finite* number of *sampling points*. Based on these discrete observations, an estimate or a statistic is formed for use in the classification problem at hand. Usually, two kinds of sampling schemes are considered: *deterministic* and *random*. In deterministic sampling, the sampling points are chosen according to a deterministic rule, such as periodic sampling. In random sampling, the sampling points are chosen according to a randomized rule, such as Poisson sampling. It is noteworthy that imperfections in deterministic sampling schemes (e.g. periodic sampling with jitter) may lead to random sampling.

The problem of functional statistics for *sampled data* is thus of extreme interest in practical issues. As a modest step towards this direction, we will be concerned in this section with the consistency of the wavelet-based classification approach combined with the k-nearest neighbor rule for deterministic sampled curves. We realize however that more work is needed to understand the phenomenon in all generality. In particular, we will not address the important issue of a signal observation perturbed by an additive random noise.

In our sampling model, the data consist of $n+m$ random observations $(\mathbb{Z}_1, Y_1), \ldots, (\mathbb{Z}_{n+m}, Y_{n+m})$, where each observational unit (curve) is now a vector of ℓ measurements $\mathbb{Z}_i = (Z_i^1, \ldots, Z_i^\ell)$ along equidistant sampling points $t_1, \ldots, t_\ell$. That is, for every $i = 1, \ldots, n$,

$$Z_i^p = X_i(t_p), \quad p = 1, \ldots, \ell.$$

Recall that each X_i was expanded on a wavelet basis via (1.6) as

$$X_i(t) = \sum_{j=1}^{\infty} X_{ij}\psi_j(t), \quad t \in [0,1].$$

In the sampling framework, the coefficients X_{ij} are unknown, and they are consistently estimated by

$$\bar{X}_{ij} = \frac{1}{\ell}\sum_{p=1}^{\ell} Z_i^p \psi_j(t_p),$$

which yields the following estimates of the X_i's:

$$\bar{X}_i(t) = \sum_{j=1}^{\infty} \bar{X}_{ij}\psi_j(t), \quad t \in [0,1].$$

Just as in Section 1.2, we fix a maximum resolution level J, so that

$$\bar{X}_i(t) \approx \sum_{j=1}^{2^J} \bar{X}_{ij}\psi_j(t), \quad t \in [0,1]. \tag{1.10}$$

We denote by $\bar{\mathbf{X}}_i = (\bar{X}_{i1}, \bar{X}_{i2}, \ldots)$ the sequence of coefficients associated with $\bar{X}_i$. With this notation, the wavelet-based dimension reduction strategy presented in Section 1.2 is applied to the "sampled" data $\bar{\mathbf{X}}_i$ instead of the "perfect" ones $\mathbf{X}_i$. Using the k-NN rule for classification, the strategy reads as follows:

- Let the n training data reorder the 2^J basis functions $\{\psi_1, \ldots, \psi_{2^J}\}$ into $\left\{\psi_{\bar{j}_1}, \ldots, \psi_{\bar{j}_{2^J}}\right\}$ via the scheme

$$\sum_{i=1}^{n} \bar{X}_{i\bar{j}_1}^2 \geq \sum_{i=1}^{n} \bar{X}_{i\bar{j}_2}^2 \geq \ldots \geq \sum_{i=1}^{n} \bar{X}_{i\bar{j}_{2^J}}^2. \tag{1.11}$$

- For $d = 1, \ldots, 2^J$, approximate each $\bar{X}_i$ by the sum $\sum_{p=1}^{d} \bar{X}_{i\bar{j}_p}\psi_{\bar{j}_p}$.

- Perform nearest neighbor classification using the finite-dimensional data $\bar{X}_i^{(d)} = (\bar{X}_{i\bar{j}_1}, \ldots, \bar{X}_{i\bar{j}_d})$, $i = 1, \ldots, n$.

The dimension d and the number of neighbors k are simultaneously selected by data-splitting and empirical risk minimization. Precisely, set

$$\bar{\mathcal{D}}_n = \left\{(\bar{\mathbf{X}}_1, Y_1), \ldots, (\bar{\mathbf{X}}_n, Y_n)\right\}$$

and denote by $g_{n,k}^{(d)}(x, \bar{\mathcal{D}}_n)$ the k-NN rule performed on the first d wavelet coefficients ordered via (1.11). The rule $g_{n,k}^{(d)}(x, \bar{\mathcal{D}}_n)$ is defined as follows. The training data are first reordered

$$\left(\bar{\mathbf{X}}_{(1)}^{(d)}(x), Y_{(1)}(x)\right), \ldots, \left(\bar{\mathbf{X}}_{(n)}^{(d)}(x), Y_{(n)}(x)\right)$$

according to increasing Euclidean distances $\|\bar{\mathbf{X}}_i^{(d)} - x\|$ of the $\bar{\mathbf{X}}_i^{(d)}$'s to $x \in \mathbb{R}^d$. In other words, $\bar{\mathbf{X}}_{(i)}^{(d)}(x)$ is the i-th nearest neighbor of x amongst $\bar{\mathbf{X}}_j^{(d)}, j = 1, \ldots, n$. If distance ties occur, a tie-breaking strategy must be defined. For example, in case of $\|\bar{\mathbf{X}}_i^{(d)} - x\| = \|\bar{\mathbf{X}}_j^{(d)} - x\|$, $\bar{\mathbf{X}}_i^{(d)}$ may be declared closer to x if $i < j$, i.e., tie-breaking is done by indices. The k-NN classification rule is finally defined as

$$g_{n,k}^{(d)}\left(x, \bar{\mathcal{D}}_n\right) = \begin{cases} 0 & \text{if } \sum_{i=1}^{k} \mathbf{1}_{\{Y_{(i)}(x)=0\}} \geq \sum_{i=1}^{k} \mathbf{1}_{\{Y_{(i)}(x)=1\}} \\ 1 & \text{otherwise.} \end{cases} \tag{1.12}$$

The free parameters d and k are optimally selected by minimizing the empirical probability of error based on the independent validation set, that is

$$(\tilde{d}, \tilde{k}) \in \underset{d=1,\ldots,2^J, k=1,\ldots,n}{\operatorname{argmin}} \left[\frac{1}{m} \sum_{i=n+1}^{n+m} \mathbf{1}_{[g_{n,k}^{(d)}(\bar{\mathbf{X}}_i^{(d)}, \bar{\mathcal{D}}_n) \neq Y_i]} \right]. \tag{1.13}$$

Denoting by $\bar{X}$ the truncated decomposition (1.10) of a generic X and by $\bar{\mathbf{X}}^{(d)} = (\bar{X}_{\bar{j}_1}, \ldots, \bar{X}_{\bar{j}_d})$ the first d coefficients reordered via (1.11), this method leads to the classifier $\tilde{g}(\bar{\mathbf{x}}) = g_{n,\tilde{k}}^{(\tilde{d})}(\bar{\mathbf{x}}^{(\tilde{d})}, \bar{\mathcal{D}}_n)$ with a probability of misclassification

$$\bar{L}_{n+m}(\tilde{g}) = \mathbf{P}\big\{ \tilde{g}(\bar{\mathbf{X}}) \neq Y | \bar{\mathcal{D}}_{n+m} \big\}.$$

As for now, to lighten the notation, we denote by $\bar{g}_{n,k}^{(d)}(.)$ the k-NN classifier (1.12) performing on the sample data $\bar{\mathcal{D}}_n$, and by $g_{n,k}^{(d)}(.)$ the k-NN classifier performing on the "true" data set

$$\mathcal{D}_n = \big\{ (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n) \big\}.$$

For fixed $d \leq 2^J$ and $k \leq n$, the probability of error of $\bar{g}_{n,k}^{(d)}$ and $g_{n,k}^{(d)}$ are respectively given by

$$\bar{L}_n\big(\bar{g}_{n,k}^{(d)}\big) = \mathbf{P}\big\{ \bar{g}_{n,k}^{(d)}(\bar{\mathbf{X}}) \neq Y | \bar{\mathcal{D}}_n \big\}$$

and

$$L_n\big(g_{n,k}^{(d)}\big) = \mathbf{P}\big\{ g_{n,k}^{(d)}(\mathbf{X}) \neq Y | \mathcal{D}_n \big\}.$$

The following lemma establishes that these two error terms are close when ℓ, the number of sampling points, tends to infinity. Recall that, given a generic X, we denote by $\mathbf{X}^{(d)} = (\mathbf{X}_{j_1}, \ldots, \mathbf{X}_{j_d})$ the first "true" d coefficients reordered via the scheme (1.8).

Lemma 1.4.1 *Suppose that for each $d = 1, \ldots, 2^J$, the random variable $\mathbf{X}^{(d)}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. Then,* a.s.*, for all $d \leq 2^J$ and all $k \leq n$, we have*

$$\lim_{\ell \to \infty} \bar{L}_n(\bar{g}_{n,k}^{(d)}) = L_n(g_{n,k}^{(d)}). \tag{1.14}$$

Consequently,

$$\lim_{\ell \to \infty} \mathbf{E}\Big\{ \inf_{d=1,\ldots,2^J, k=1,\ldots,n} \bar{L}_n(\bar{g}_{n,k}^{(d)}) \Big\} = \mathbf{E}\Big\{ \inf_{d=1,\ldots,2^J, k=1,\ldots,n} L_n(g_{n,k}^{(d)}) \Big\}. \tag{1.15}$$

Proof Observe first that

$$\begin{aligned}\bar{L}_n\big(\bar{g}_{n,k}^{(d)}\big) &= \mathbf{P}\big\{\bar{g}_{n,k}^{(d)}(\bar{\mathbf{X}}) \neq Y | \bar{\mathcal{D}}_n\big\} \\ &= \mathbf{P}\big\{\bar{g}_{n,k}^{(d)}(\bar{\mathbf{X}}) \neq Y | \mathcal{D}_n\big\}. \end{aligned} \tag{1.16}$$

Now, denote by $\bar{S}_J$ (*resp.* S_J) the vector $\{\bar{j}_1, \ldots, \bar{j}_{2^J}\}$ (*resp.* the vector $\{j_1, \ldots, j_{2^J}\}$) corresponding to the data-based coefficient reordering (1.11) (*resp.* (1.8)). Define

$$\delta = \min_{1 \leq j < j' \leq 2^J} \left| \sum_{i=1}^n X_{ij}^2 - \sum_{i=1}^n X_{ij'}^2 \right|,$$

and observe, since $\mathbf{X}^{(2^J)}$ is absolutely continuous with respect to the Lebesgue measure, that $\delta > 0$ *a.s.* Moreover, we have

$$\big\{\bar{S}_J \neq S_J\big\} = \big\{\exists (k,k'), 1 \leq k < k' \leq 2^J : \sum_{i=1}^n \bar{X}_{ij_k}^2 < \sum_{i=1}^n \bar{X}_{ij_{k'}}^2 \big\}. \tag{1.17}$$

We first proceed to show that, *a.s*,

$$\bar{S}_J = S_J \tag{1.18}$$

for all ℓ large enough. To this aim, note that the condition $k < k'$ implies

$$\sum_{i=1}^n X_{ij_k}^2 \geq \sum_{i=1}^n X_{ij_{k'}}^2 + \delta.$$

Consequently, using inequality (1.17), we obtain

$$\big\{\bar{S}_J \neq S_J\big\} \subseteq \bigcup_{1 \leq k < k' \leq 2^J} \Big\{ \sum_{i=1}^n (\bar{X}_{ij_{k'}}^2 - X_{ij_{k'}}^2) - \sum_{i=1}^n (\bar{X}_{ij_k}^2 - X_{ij_k}^2) > \delta \Big\}.$$

Since, for every $i = 1, \ldots, n$ and $j = 1, \ldots, 2^J$, $\lim_{\ell \to \infty} \bar{X}_{ij} = X_{ij}$, we conclude that (1.18) holds, as desired.

To finish the proof of equality (1.14), fix $d \leq 2^J$, $k \leq n$, and reorder the "true" training data set $\mathcal{D}_n$

$$\big(\mathbf{X}_{i_1}^{(d)}(\mathbf{X}^{(d)}), Y_{i_1}(\mathbf{X}^{(d)})\big), \ldots, \big(\mathbf{X}_{i_n}^{(d)}(\mathbf{X}^{(d)}), Y_{i_n}(\mathbf{X}^{(d)})\big)$$

according to the increasing Euclidean distances $\|\bar{\mathbf{X}}_i^{(d)} - \mathbf{X}^{(d)}\|$ of the $\bar{\mathbf{X}}_i^{(d)}$'s to $\mathbf{X}^{(d)} \in \mathbb{R}^d$, and set

$$\gamma = \min_{1 \leq j < j' \leq n} \Big\{ \Big| \|\mathbf{X}^{(d)} - \mathbf{X}_j^{(d)}\| - \|\mathbf{X}^{(d)} - \mathbf{X}_{j'}^{(d)}\| \Big| \Big\}.$$

Invoking the absolute continuity of $\mathbf{X}^{(d)}$, we see that $\gamma > 0$ *a.s.* Clearly, for $j' > j$, the inequality

$$\|\mathbf{X}^{(d)} - \mathbf{X}_{i_j}^{(d)}\| + \gamma \leq \|\mathbf{X}^{(d)} - \mathbf{X}_{i_{j'}}^{(d)}\|$$

holds. Now the following chain of inequalities is valid:

$$\begin{aligned}
&\left|\mathbf{P}\left\{\bar{g}_{n,k}^{(d)}(\bar{\mathbf{X}}) \neq Y|\bar{\mathcal{D}}_n\right\} - \mathbf{P}\left\{g_{n,k}^{(d)}(\mathbf{X}) \neq Y|\mathcal{D}_n\right\}\right| \\
&\quad \leq \mathbf{P}\left\{\bar{g}_{n,k}^{(d)}(\bar{\mathbf{X}}) \neq g_{n,k}^{(d)}(\mathbf{X})|\mathcal{D}_n\right\} \\
&\qquad \text{(by equality (1.16))} \\
&\quad \leq \mathbf{P}\left\{\bar{g}_{n,k}^{(d)}(\bar{\mathbf{X}}) \neq g_{n,k}^{(d)}(\mathbf{X}), \bar{S}_J = S_J|\mathcal{D}_n\right\} + \mathbf{P}\left\{\bar{S}_J \neq S_J|\mathcal{D}_n\right\} \\
&\quad \leq \sum_{1\leq j<j'\leq n} \mathbf{P}\left\{\|\bar{\mathbf{X}}^{(d)} - \bar{\mathbf{X}}_{i_{j'}}^{(d)}\| < \|\bar{\mathbf{X}}^{(d)} - \bar{\mathbf{X}}_{i_j}^{(d)}\| \ |\mathcal{D}_n\right\} + \mathbf{P}\left\{\bar{S}_J \neq S_J|\mathcal{D}_n\right\} \\
&\qquad \text{(by the union bound)} \\
&\quad \leq \sum_{1\leq j<j'\leq n} \mathbf{P}\Big\{\left(\|\bar{\mathbf{X}}^{(d)} - \bar{\mathbf{X}}_{i_j}^{(d)}\| - \|\mathbf{X}^{(d)} - \mathbf{X}_{i_j}^{(d)}\|\right) \\
&\qquad\qquad - \left(\|\bar{\mathbf{X}}^{(d)} - \bar{\mathbf{X}}_{i_{j'}}^{(d)}\| - \|\mathbf{X}^{(d)} - \mathbf{X}_{i_{j'}}^{(d)}\|\right) > \gamma|\mathcal{D}_n\Big\} \\
&\qquad \text{(by definition of } \gamma \text{ and by (1.18) for all } \ell \text{ large enough).}
\end{aligned}$$

The conclusion follows from the fact that $\lim_{\ell\to\infty} \bar{X}_{ij} = X_{ij}$. The second statement of the lemma is a consequence of Lebesgue's dominated convergence theorem. ■

Theorem 1.4.1 below states the consistency of the rule $\tilde{g}$ selected in (1.13) for a large class of distributions. If we are just worried about the consistency of the wavelet-based nearest neighbor classification rule, this theorem reassures us that nothing is lost as long as the curves are measured at more and more dense sampling points.

Theorem 1.4.1 *Suppose that for each $d = 1, \ldots, 2^J$, the random variable $\mathbf{X}^{(d)}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, and that*

$$\lim_{n\to\infty} m = \infty \quad \textit{and} \quad \lim_{n\to\infty} \frac{\log n}{m} = 0. \tag{1.19}$$

Then the automatic rule $\tilde{g}$ defined in (1.13) *is consistent in the sense*

$$\lim_{J\to\infty} \lim_{n\to\infty} \lim_{\ell\to\infty} \mathbf{E}\left\{\bar{L}_{n+m}(\tilde{g})\right\} = L^*.$$

Proof Consider the following decomposition:

$$\mathbf{E}\big\{\bar{L}_{n+m}(\tilde{g})\big\} - L^* = \mathbf{E}\Big[\bar{L}_{n+m}(\tilde{g}) - \inf_{d=1,\ldots,2^J,k=1,\ldots,n} \bar{L}_n(\bar{g}_{n,k}^{(d)})\Big] \tag{1.20}$$

$$+ \mathbf{E}\Big[\inf_{d=1,\ldots,2^J,k=1,\ldots,n} \bar{L}_n(\bar{g}_{n,k}^{(d)})\Big] - L^*. \tag{1.21}$$

From inequality (1.1) and (1.2), we already known that (1.20) is bounded above by

$$\sqrt{\frac{2\log(2*2^J n)}{m}} + \frac{1}{\sqrt{2m\log(2*2^J n)}}.$$

Therefore, using Lemma 1.4.1 and conditions (1.19), we conclude that

$$\lim_{J\to\infty}\lim_{n\to\infty}\lim_{\ell\to\infty} \mathbf{E}\Big[\bar{L}_{n+m}(\tilde{g}) - \inf_{d=1,\ldots,2^J,k=1,\ldots,n} \bar{L}_n(\bar{g}_{n,k}^{(d)})\Big] = 0.$$

Let us now turn to the analysis of the term (1.21). From Lemma 1.4.1, we know that

$$\lim_{\ell\to\infty} \mathbf{E}\Big\{\inf_{d=1,\ldots,2^J,k=1,\ldots,n} \bar{L}_n(\bar{g}_{n,k}^{(d)})\Big\} = \mathbf{E}\Big\{\inf_{d=1,\ldots,2^J,k=1,\ldots,n} L_n(g_{n,k}^{(d)})\Big\}.$$

Stone [27] proved that there exists a suitable sequence (k_n) with $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$ such that $\mathbf{E}\{L_n(g_{n,k_n}^{(2^J)})\} \to L_{2^J}^*$ as $n \to \infty$. Therefore, we have

$$\begin{aligned}
&\lim_{n\to\infty}\lim_{\ell\to\infty}\left(\mathbf{E}\Big[\inf_{d=1,\ldots,2^J,k=1,\ldots,n} \bar{L}_n(\bar{g}_{n,k}^{(d)})\Big] - L_{2^J}^*\right) \\
&\quad = \lim_{n\to\infty}\left(\mathbf{E}\Big[\inf_{d=1,\ldots,2^J,k=1,\ldots,n} L_n(g_{n,k}^{(d)})\Big] - L_{2^J}^*\right) \\
&\quad \le \lim_{n\to\infty}\left(\mathbf{E}\big\{L_n(g_{n,k_n}^{(2^J)})\big\} - L_{2^J}^*\right) \\
&\quad = 0 \\
&\qquad \text{(by Stone's theorem).}
\end{aligned}$$

Invoking Lemma 1.2.1, we finally obtain

$$\lim_{J\to\infty}\lim_{n\to\infty}\lim_{\ell\to\infty}\left(\mathbf{E}\Big[\inf_{d=1,\ldots,2^J,k=1,\ldots,n} \bar{L}_n(\bar{g}_{n,k}^{(d)})\Big] - L^*\right) = 0.$$

This concludes the proof of Theorem 1.4.1. ■

Bibliography

[1] C. Abraham, G. Biau, and B. Cadre. On the kernel rule for function classification. *Annals of the Institute of Statistical Mathematics*, 2005. In press.

[2] U. Amato, A. Antoniadis, and I. De Feis. Dimension reduction in functional regression with applications. *Computational Statistics and Data Analysis*, 2005. In press.

[3] P.N. Belhumeur, J.P. Hepana, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:711–720, 1997.

[4] G. Biau, F. Bunea, and M. H. Wegkamp. Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory*, 51:2163–2172, 2005.

[5] S. Boucheron, O. Bousquet, and G. Lugosi. *Advanced Lectures in Machine Learning*, chapter Introduction to statistical learning theory, pages 169–207. Springer, Heidelberg, 2004.

[6] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 2005. In press.

[7] A. Cohen. *Numerical Analysis of Wavelet Methods*. Elsevier, Amsterdam, 2003.

[8] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[9] T.M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC14:326–334, 1965.

[10] B.V. Dasarathy. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, 1991.

[11] I. Daubechies. *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[12] L. Devroye. Automatic pattern recognition: A study of the probability of error. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:530–543, 1988.

[13] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New-York, 1996.

[14] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? With discussion and a reply by the authors. *Journal of the Royal Statistical Society Series B*, 57:545–564, 2002.

[15] F. Ferraty and P. Vieu. The functional nonparametric model and application to spectrometric data. *Computational Statistics*, 17:545–564, 2002.

[16] P. Hall, D.S. Poskitt, and B. Presnell. A functional data-analytic approach to signal discrimination. *Technometrics*, 43:1–9, 2001.

[17] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23:73–102, 1995.

[18] M. Kirby and L. Sirovich. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:103–108, 1990.

[19] S.R. Kulkarni and S.E. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41:1028–1039, 1995.

[20] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, San Diego, 1999. 2nd edition.

[21] Y. Meyer. *Wavelet and Operators.* Cambridge University Press, Cambridge, 1992.

[22] D.B. Pollard. *A User's Guide to Measure Theoretic Probability.* Cambridge University Press, Cambridge, 2002.

[23] C. Preda and G. Saporta. Régression PLS sur un processus stochastique. *Revue de Statistique Appliquée*, 50(2), 2002.

[24] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis.* Springer-Verlag, New York, 1997.

[25] J.O. Ramsay and B.W. Silverman. *Applied Functional Data Analysis. Methods and Case Studies*. Springer-Verlag, New York, 2002.

[26] F. Rossi and N. Villa. Functional data analysis with support vector machine. In *Proceedings of ASMDA Conference 2005*, Brest, France, 2005.

[27] C.J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5:595–645, 1977.

[28] C. Tuleau. *Sélection de Variables pour la Discrimination en Grande Dimension et Classification de Données Fonctionnelles*. PhD thesis, University Paris XI-Orsay, 2005.

[29] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

Conclusion et perspectives

Dans ce travail de thèse, nous avons apporté une contribution supplémentaire aux thèmes de l'apprentissage statistique et de l'estimation non paramétrique. Nous nous sommes intéressés aux propriétés théoriques des différentes méthodes étudiées, tout en essayant, lorsque c'était possible, d'illustrer leurs performances à l'aide de jeux de données réelles ou simulées.

Dans la première partie du travail, nous nous sommes attachés à montrer l'intérêt d'appréhender les histogrammes modifiés comme des systèmes dynamiques à espace d'états de dimension infinie, tout en soulignant leurs bonnes performances pour estimer des densités complexes. Dans le cas multivarié, nous avons ensuite clairement mis en évidence le rôle déterminant joué par des partitions construites à partir du nuage des données relativement à la qualité de l'estimation. La combinaison de ces deux études devrait naturellement conduire, dans un avenir proche, à la généralisation de ces systèmes dynamiques fonctionnels au domaine multivarié. Il y a en effet fort à parier que la mise au point d'une procédure de sélection d'histogrammes modifiés itérés dans le cadre multivarié permettrait d'améliorer significativement les propriétés de certains estimateurs de la densité. Il s'agit néanmoins d'un problème délicat dont l'étude théorique reste à faire et qui devrait poser bon nombre de problèmes nouveaux dans sa mise en place effective.

Dans le premier chapitre de la deuxième partie, nous avons adapté les méthodes combinatoires développées par Devroye et Lugosi à la sélection des différents paramètres des histogrammes modifiés. Les résultats obtenus dans ce chapitre ont montré que, outre le choix de la partition et celui du nombre de classes nécessaires pour construire les histogrammes modifiés, un choix pertinent d'une densité de référence se révèle crucial pour obtenir de bonnes propriétés à distance finie. Dans le second chapitre nous avons utilisé les techniques combinatoires afin de développer une méthode originale permettant de sélectionner le paramètre de lissage (fonctionnel) d'estimateurs à noyau variables. Cette méthode permet, en particulier, de choisir de manière optimale le paramètre de lissage de la plupart des estimateurs à noyau variables proposés dans la littérature.

S'agissant de la recherche future, dans un premier temps, il nous semble important de s'attacher à développer des algorithmes permettant la mise en oeuvre effective sur machines des méthodes combinatoires. Ce travail a été amorcé dans le premier chapitre des deux premières parties, à l'aide d'algorithmes souvent très coûteux en temps de calcul, qu'il doit être possible d'améliorer. Dans un second temps, il serait également intéressant d'étendre les techniques combinatoires au cas de données dépendantes. Les inégalités exponentielles pour les processus empiriques obtenues par Rio en 2002 devraient constituer un bon point de départ pour cette nouvelle problématique.

La dernière partie de la thèse a finalement mis en évidence l'intérêt que pouvait apporter les ondelettes au problème de la classification fonctionnelle. Cette partie s'achève par la démonstration d'un résultat de convergence pour des règles calculées à partir de courbes discrétisées. Il s'agit d'un résultat qui nous semble pertinent dans la mesure où, en pratique, les individus sont toujours des versions discrétisées de phénomènes continus. Il n'en demeure pas moins que la tâche à accomplir dans ce domaine est encore immense. Dans l'avenir, nous proposerions d'étendre cette approche à des observations bruitées et mesurées en des points différents. Par ailleurs, toujours dans ce contexte, il serait intéressant d'obtenir des vitesses de convergence. En effet, ces dernières apporteraient de l'information sur le nombre de points de discrétisation "optimal" de la courbe en fonction de la taille de l'échantillon. Un tel résultat pourrait trouver de nombreuses applications, notamment dans le domaine des plans d'expériences. En effet, dans la pratique, les mesures ont souvent un coût. Il est donc important pour le praticien de connaître le nombre minimum de mesures nécessaires afin de garantir une bonne performance de la méthode utilisée.

Pour finir, à la lumière de premières expériences, nous pensons que les *algorithmes* et *méthodes à noyau* (SVM, dual clustering, kernel PCA...) adaptés au cas de la classification de courbes devraient fournir un ensemble d'outils prometteurs pour le domaine de l'estimation fonctionnelle.