Sur quelques problèmes d'apprentissage supervisé et non supervisé

Thomas Laloë

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

Sur Quelques Problèmes d'Apprentissage Supervisé et Non Supervisé

par

Thomas Laloë

Présentée et soutenue publiquement le **27/11/2009** devant le jury composé de :

MM.	P. BERTHET	Professeur, Université Toulouse III	Rapporteur
	N. BEZ	Chargé de recherche, IRD Sète	Examinateur
	G. BIAU	Professeur, Université Paris VI	Directeur de Thèse
	B. CADRE	Professeur, ENS Rennes	Directeur de Thèse
	J-M. MARIN	Professeur, Université Montpellier II	Examinateur
	J-M. POGGI	Professeur, Université Paris V	Rapporteur

Remerciements

Après trois années, voici venu le moment qu'ont attendu et qu'attendront avec impatience tous les doctorants au cours de leur thèse, la rédaction des remerciements. Ce n'est pas sans une certaine émotion que je me lance dans cette tâche, à la fois agréable, car elle me fait repenser à toutes ces personnes qui m'ont tant apporté, et ardue au vu de leur nombre important.

Je tiens tout d'abord à exprimer toute ma gratitude à Gérard Biau et Benoît Cadre qui m'ont accompagné, soutenu, poussé, guidé, tout au long de cette thèse, et dont je pense sincèrement qu'ils sont les meilleurs directeurs dont puisse rêver un doctorant. Tout en me laissant une grande liberté dans mon travail, ils ont toujours fait preuve d'une grande disponibilité, et ce en dépit de l'écart géographique qui nous séparait. Je garderai toujours avec moi le souvenir de nos discussions, de leurs conseils, et surtout de leur gentillesse. J'espère que nos relations et nos collaborations continuerons longtemps après cette thèse.

J'adresse également mes remerciements à Philippe Berthet et Jean-Michel Poggi qui m'ont fait l'honneur d'accepter de rapporter cette thèse, malgré un emploi du temps que je devine chargé. Je remercie Jean-Michel Marin de bien vouloir présider mon Jury de thèse, ainsi que Nicolas Bez qui a gentiment accepté de participer à mon Jury.

Je remercie l'ensemble des membres du département de mathématiques, et particulièrement les gens avec qui j'ai collaboré pour mon monitorat. Je tiens

à exprimer ma gratitude aux secrétaires, qui sont les véritables piliers de ce laboratoire. J'ai une pensée particulière pour les doctorants, ATER et maîtres de conférences avec qui j'ai tissé des liens que je sais solides et durables. Enfin je tiens à remercier celui qui a été mon co-bureau tout au long de cette thèse. Rémi, pour toutes nos discussions, nos rires, nos échanges, tes relectures, nos soirées, et sans oublier notre screenquizz auquel tu m'as initié, merci (et vive Michel Delpech).

Je remercie également Pierre Cartigny et tout l'équipe de l'UMR ASB qui m'ont accueilli à bras ouvert au sein de leur équipe. J'ai rencontré là-bas des gens formidables et j'ai pu tisser de solides liens d'amitié.

Je remercie tous mes amis, proches, famille qui m'ont soutenu durant cette aventure, et particulièrement ceux, ils se reconnaîtront, qui m'ont aidé pour des relectures tout au long de ces trois ans.

Je n'aurais pas pu arriver jusque là sans l'équilibre, la chaleur, le soutien et le bonheur dans lequel j'ai vécu, merci Maman, Papa, mes frères Christophe et Julien, mais aussi Myriam, Guy et Aurélie.

Et enfin, merci n'est qu'un petit mot pour exprimer à Amandine tout ce que je lui dois après 3 ans de partage, de bonheur et d'amour. Pour avoir su gérer mon stress, avoir su m'épauler dans les moments difficiles, et tout simplement être là avec moi au quotidien, je te le dis du fond du coeur, merci.

Je terminerai par un non remerciement. Je tiens à adresser un message personnel à ce qui a été ma bête noire durant cette thèse, l'orthographe :

Sleon une édtue de l' uvinertisé de Cmabrigde, l'odrre des ltteers dnas un mtos n'a pas d'ipmrotncae, la suele coshe ipmrotnate est que la pmeirère et la drenèire soniet à la bnnoe pclae. Le rsete peut êrte dnas un dsérorde ttoal et

vuos puoevz tujoruos lrie snas porlblème. C'est prace que le creaveu hmauin ne lit pas chuaqe ltetre elle-mmêe, mias le mot cmome un tuot.
La peruve...

Table des matières

Introduction générale

1.1 Présentation de la thèse

L'apprentissage statistique (Vapnik [27]) est un paradigme qui regroupe un ensemble de méthodes et d'algorithmes permettant d'extraire l'information pertinente des données, ou d'apprendre des comportements à partir d'exemples. Ses applications sont nombreuses et présentes dans des domaines aussi variés que la recherche d'informations dans de grands ensembles de données (segmentation thématique de texte, fouille d'images, etc.) ou la biologie (comportement de populations, puces ADN, etc.).

On distingue deux grandes problématiques en apprentissage statistique : l'apprentissage supervisé d'une part, et l'apprentissage non supervisé d'autre part.

L'apprentissage supervisé consiste à établir des règles de comportement à partir d'une base de données contenant des exemples de cas déjà étiquetés. Plus précisément, cette base de données est un ensemble de couples entrées-sorties $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ aléatoires. L'objectif est alors d'apprendre à prédire, pour toute nouvelle entrée X, la sortie Y. On parle de régression dans le cas où les sorties sont à valeurs continues (Györfi, Kohler, Krzyżak et Walk [15]), et de classification dans le cas où elles sont à valeurs discrètes (Devroye, Györfi et Lugosi [11]).

La théorie de l'apprentissage non supervisé traite le cas où l'on dispose seule-

ment des entrées $\{X_i\}_{1\leq i\leq n}$, sans les sorties. Le problème le plus important consiste alors à effectuer un partitionnement des données, également appelé clustering. Il s'agit de regrouper les observations en différents groupes homogènes (les clusters), en faisant en sorte que les données de chaque sous-ensemble partagent des caractéristiques communes.

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. En particulier, ces données peuvent se présenter sous la forme de fonctions, ou courbes, aléatoires. Néanmoins, le périmètre d'utilisation des méthodes statistiques traditionnelles se limite souvent au cas où l'espace des observations est de dimension finie. Dans ce nouveau contexte, l'enjeu est alors de proposer des méthodes permettant de traiter ces données fonctionnelles. L'objectif du présent travail de thèse consiste essentiellement à étudier et à approfondir des techniques d'apprentissage dans des espaces de dimension élevée. Plus précisément, notre travail se divise en trois parties.

La première partie, intitulée **régression fonctionnelle par la méthode des k-plus proches voisins**, est consacrée à l'étude d'un estimateur de la fonction de régression en dimension infinie, fondé sur la méthode des k-plus proches voisins, et proposé par Biau, Bunea et Wegkamp [3] dans le cadre de l'estimation de la fonction de densité. Ce travail s'inscrit dans la continuité des travaux de thèse de Laurent Rouvière [23] et de Christine Tuleau [26].

Dans la deuxième partie de la thèse, intitulée **classification fonctionnelle non supervisée**, on considère le problème du partitionnement de données fonctionnelles. Cette partie se divise en deux chapitres. Le premier est consacré à l'étude des performances de la méthode dite de quantification (Gersho et Gray [13]) dans des espaces de grande dimension, prolongeant ainsi les travaux de Támas Linder [17]. Dans le second nous appliquons nos

méthodes pour réaliser une étude du comportement de bancs d'anchois le long de la côte péruvienne. Ce dernier travail a été réalisé en collaboration avec François Gerlotto, de l'Institut de Recherche pour le Développement (IRD).

La troisième et dernière partie, intitulée **estimation non paramétrique des ensembles de niveau de la fonction de régression**, est dévolue au problème de l'estimation des ensembles de niveau de la fonction de régression. Ce travail fait suite aux travaux de Benoît Cadre [8], ainsi qu'à ceux de la première partie de cette thèse où nous avons étudié le problème de l'estimation de la fonction de régression.

1.2 Régression fonctionnelle par la méthode des k-plus proches voisins

Le principe de la régression est de prédire une variable (sortie) à partir d'une observation (entrée). Les observations prennent habituellement la forme de vecteurs de dimension d, rassemblant un certain nombre de mesures numériques. Cependant, dans beaucoup de problèmes pratiques, ces données se présentent sous la forme de fonctions aléatoires (enregistrement de voix, spectres, images, etc.). Cela recentre le problème de la régression dans le domaine général de l'analyse de données fonctionnelles (Ramsay et Silverman [22]). Bien qu'en pratique de telles données sont observées en un nombre fini de points, le défi dans ce contexte est d'inférer la structure des données en exploitant leur nature infini-dimensionnelle.

Pour atteindre ce but, nous disposons d'un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) où X_i est une fonction aléatoire à valeurs dans un espace fonctionnel $\mathcal{H}$, et Y_i est une variable aléatoire réelle. Typiquement, $\mathcal{H}$ pourra être un espace de Hilbert, comme par exemple l'espace $L^2([0\,,1])$ des fonctions de carré

intégrable sur $[0\,,1]$. L'objectif consiste alors à construire, à l'aide de cet échantillon, un estimateur $\hat{r}_n$ de la fonction de régression r définie, pour tout $x \in \mathcal{H}$, par

$$r(x) = \mathbb{E}\left[Y|X = x\right].$$

A ce jour, de nombreuses méthodes ont été étudiées, testées et comparées pour des observations évoluant dans des espaces de dimension finie (voir Györfi, Kohler, Krzyżak et Walk [15]). Bien que ces techniques puissent, sous certaines hypothèses, être étendues au domaine fonctionnel, elles se trouvent en général confrontées dans ce cas au fléau de la dimension (Abraham, Biau et Cadre [1]). Pour pallier cette difficulté, une possibilité est d'effectuer une étape préliminaire de réduction de dimension.

Parmi les nombreuses méthodes de réduction de dimension existantes, nous avons choisi d'utiliser une méthode de projection. Il s'agit en fait de projeter les données dans une base appropriée, et de ne considérer qu'un nombre réduit de coefficients. Dans des travaux récents, Biau, Bunea et Wegkamp [5] d'une part, et Berlinet, Biau et Rouvière [2] d'autre part ont utilisé respectivement des bases de Fourier et d'ondelettes dans une problématique de classification fonctionnelle supervisée. Fromont et Tuleau [12] proposent par ailleurs une étude sur la pertinence et le choix d'un terme de pénalité pour sélectionner le nombre de coefficients conservés. Nous proposons, dans le premier chapitre de cette thèse, d'étendre ces méthodes au cas de la régression fonctionnelle, sans toutefois privilégier le choix d'une base particulière.

L'idée générale est la suivante : on considère une base $\{\phi_j\}_{j=1}^{\infty}$ de $\mathcal{H}$, et on note $\mathcal{H}^{(d)}$ l'espace engendré par $\{\phi_j\}_{j=1}^{d}$. On projette ensuite les données sur $\mathcal{H}^{(d)}$ de la façon suivante :

$$X_i^{(d)} = \sum_{j=1}^{d} X_{i,j}\phi_j.$$

On calcule finalement l'estimateur $\hat{r}_n$ des k-plus proches voisins à partir de ces données projetées. Plus précisément, pour tout élément x dans $\mathcal{H}^{(d)}$, on

réordonne les données projetées

$$\left(X_{(1)}^{(d)}(x), Y_{(1)}(x)\right), \ldots, \left(X_{(n)}^{(d)}(x), Y_{(n)}(x)\right),$$

selon les distances croissantes $\|X_i^{(d)} - x\|$ entre $X_i^{(d)}$ et x. On calcule alors $\hat{r}_n(x)$ en prenant la moyenne des k plus proches voisins de x :

$$\hat{r}_n(x) = \sum_{i=1}^{k} Y_{(i)}^{(d)}(x).$$

A la fois k et d sont déterminés automatiquement par une méthode de "data splitting". Sous certaines hypothèses, nous obtenons (Théorème 1.2.1) la convergence faible de $\hat{r}_n$ vers r, c'est-à-dire

$$\mathbb{E}\int_{\mathcal{H}} \left(\hat{r}_n(x) - r(x)\right)^2 \rho_X(dx) \to 0 \text{ quand } n \to \infty,$$

où ρ_X représente la loi marginale des entrées X_i.

Ce travail à fait l'objet d'un article en anglais publié dans la revue *Statistics and Probability Letters*.

1.3 Classification fonctionnelle non supervisée

Comme nous l'avons souligné plus haut, l'enjeu du clustering est de repérer des groupes dans un ensemble d'observations. Il s'agit d'une méthode d'apprentissage non supervisé, utilisée dans des domaines très variés, comme par exemple l'intelligence artificielle, la sociologie, le marketing, la recherche médicale, ou encore les sciences politiques.

Nous nous intéressons dans cette partie à une technique de clustering par partitionnement, c'est-à-dire que nous allons classer les données en k groupes satisfaisant les deux conditions suivantes :

1. Chaque groupe contient au moins une donnée.
2. Chaque donnée appartient à un et un seul groupe.

La Figure 1.1 montre un exemple de points de $\mathbb{R}^2$ répartis en cinq groupes.

Pour les mêmes raisons que dans la première partie, nous considérons le cas où les données peuvent prendre la forme de courbes (ou surfaces) aléatoires. Cela nous ramène encore une fois dans le domaine de l'analyse de données fonctionnelles. La méthode de clustering que nous utilisons repose sur la technique dite de quantification, souvent utilisée en compression du signal (Graf et Luschgy [14], Linder [17]). Le principe est le suivant : étant donnés un espace normé $(\mathcal{H}, \|.\|)$ et un sous-ensemble fini C de $\mathcal{H}$, nommé alphabet, on représente chaque élément x de $\mathcal{H}$ par un élément (et un seul) $\hat{x}$ de C. La fonction, notée q, qui à x associe sa quantification $\hat{x} = q(x)$ est appelée quantificateur.

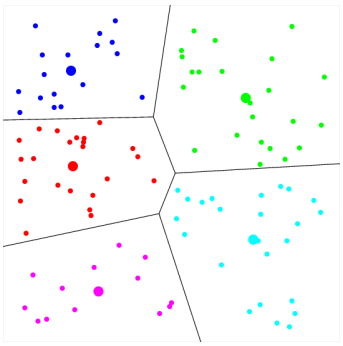

FIGURE 1.1 – Un exemple de répartition des données en 5 groupes.

Cette partie se décompose en deux chapitres. L'objectif du premier est de développer une méthodologie pour construire de manière automatique un quantificateur optimal. Cette notion d'optimalité 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 d'une part ou mesure d'erreur ponctuelle d'autre part. Dans ce travail de thèse, nous nous placerons dans le cadre d'applications nécessitant l'estimation de toute la courbe représentative de q (plutôt qu'une valeur en un point particulier) et nous ne considérerons donc que des critères globaux.

Parmi ceux-ci, les plus utilisés sont les normes L_p

$$\|x - q(x)\|_p = \begin{cases} \left(\int_{\mathcal{H}} \|x - q(x)\|^p \nu(dx) \right)^{1/p} & \text{si } 1 \leq p < \infty \\ \sup_{\mathcal{H}} \|x - q(x)\| & \text{si } p = \infty, \end{cases}$$

où ν désigne une loi de probabilité sur $\mathcal{H}$.

En particulier, la décomposition classique 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 trois raisons. D'abord, ce critère est calculable sous des hypothèses moins contraignantes que pour l'erreur L_2. En effet, il faut pour calculer l'erreur L_1 que les densités des observations soient intégrables, alors qu'elles doivent être de carré intégrable pour l'erreur L_2. Ensuite, l'erreur L_1 commise entre deux fonctions réelles peut être aisément visualisée : elle correspond en fait à l'aire comprise entre les courbes représentatives des deux fonctions. Enfin, les estimateurs calculés à partir de l'erreur L_1 sont robustes (Kemperman [16]). Ces raisons plaident en faveur du choix d'un critère L_1 pour mesurer l'erreur de la quantification. C'est donc avec ce critère que nous avons choisi de travailler dans ce chapitre.

Par conséquent, supposant que les données sont distribuées selon une loi μ sur $\mathcal{H}$, nous cherchons à déterminer un quantificateur qui minimise $\mathbb{E}\,\|X - q(X)\|$, où $X \sim \mu$. Ensuite, nous estimons ce quantificateur optimal à l'aide d'un échantillon $X_1, \ldots, X_n$ de variables aléatoires i.i.d. selon la même loi que celle de X. Nous cherchons pour cela à développer une procédure automatique pour minimiser $\sum_{i=1}^n \|X_i - q(X_i)\|$ sur l'ensemble de tous les quantificateurs possibles. Finalement, nous étudions et approfondissons en particulier deux méthodes : l'algorithme de Lloyd (Voir Gersho et Gray [13]), plus connu sous le nom de l'algorithme des k-means d'une part, et une méthode de minimisation sur les données introduite par Cadre [7] dans le cas où l'on ne

considère qu'un seul groupe d'autre part.

Sous certaines hypothèses sur $\mathcal{H}$ et la loi μ des observations, nous obtenons (Théorème 1.3.1) que notre estimateur q_n, défini par

$$q_n \in \arg\min_q \sum_{i=1}^{n} \|X_i - q(X_i)\|,$$

est convergent, dans le sens où

$$\lim_{n\to\infty} \mathbb{E}[\|X - q_n(X)\||X_1, \ldots, X_n] = \inf_q \mathbb{E}\|X - q(X)\| \qquad p.s.$$

Sous des hypothèses supplémentaires sur l'entropie métrique de $\mathcal{H}$, on obtient également (Théorème 1.3.4) une vitesse de convergence pour $\mathbb{E}\|X - q_n(X)\|$.

Nous terminons ce chapitre en validant nos méthodes sur des jeux de données réelles et simulées. En particulier nous utilisons un jeu de données réelles provenant de la base de données TIMIT (http://www-stat.stanford.edu/~tibs/ElemStatLearn/). Il s'agit d'enregistrements de différents phonèmes que nous essayons de discriminer. Ce travail a fait l'objet d'un article en anglais soumis pour publication.

Le second chapitre de cette partie est dédié à une étude sur des bancs d'anchois du Pérou. Ce travail a été mené en collaboration avec François Gerlotto, chercheur à l'Institut pour la Recherche et le Développement (IRD), au centre de recherche halieutique de Sète (IRD/IFREMER/UM2). Les technologies récentes de sonar multifaisceaux permettent d'obtenir des représentations tridimensionnelles d'un banc de poissons, à partir desquelles on peut à la fois reconstruire le banc et mesurer sa morphologie externe (dimensions, surface, volume, etc.), ainsi que sa structure interne (densité, hétérogénéité, etc.). Nous avons cherché à exploiter ces nouvelles mesures avec nos méthodes pour effectuer une discrimination des bancs observés en deux groupes. Nous nous sommes ensuite demandé quelle pouvait être la justification biologique de cette discrimination. L'analyse des groupes obtenus a ainsi permis d'établir

un nouvel indicateur de la relation des bancs avec leur environnement, et notamment avec la présence d'un prédateur. Ce travail a été présenté au groupe de travail ICES WGFAST 2009 (http://www.ismaran.it/tecpesca/news/ftfb_fast_2009/Default.aspx).

1.4 Estimation non paramétrique des ensembles de niveau pour la régression

Dans ce dernier chapitre, nous considérons le problème de la reconstruction des ensembles de niveau d'une fonction de régression r sur $\mathbb{R}^d$, c'est-à-dire des ensembles $\mathcal{L}(t) = \{x \in \mathbb{R}^d : r(x) > t\}$ où t est un niveau fixé. Nous ne considérons plus ici des données fonctionnelles mais des vecteurs de dimension d.

L'estimation des ensembles de niveau est un problème intéressant tant du point de vue théorique que pratique. L'une de ses principales applications est la mise au point de tests de fiabilité. Cependant, l'étude de ce problème à essentiellement porté sur l'estimation des ensembles de niveau d'une densité de probabilité (Tsybakov [25]). Parmi les principales applications possibles, on peut citer le clustering (Biau, Cadre et Pelletier [4]), la mise au point de tests de multimodalité (Müller et Stawitzki [18]), ou encore la reconnaissance de formes (Polonik [20]).

Ici nous nous intéressons à l'estimation des ensembles de niveau d'une fonction de régression, problème plus directement lié à l'analyse d'images. On trouvera dans Nowak et Willett [19] des exemples détaillés et précis.

Plusieurs méthodes sont utilisées dans ce domaine. On peut citer par exemple les approches fondées sur le principe de l'excès de masse (Cavalier [9]), ou encore les méthodes de type *cost sensitive* (Scott et Davenport [24]).

Pour notre part, nous avons choisi de considérer une approche de type *plug-in* (Cuevas, González-Manteiga et Rodríguez-Casal [10], Cadre [8]). Plus précisément, à partir d'un estimateur consistant $\hat{r}_n$ de r, nous estimons $\mathcal{L}(t)$ par $\mathcal{L}_n(t) = \{x \in \mathbb{R}^d : \hat{r}_n(x) > t\}$. Nous considérons plus particulièrement le cas où $\hat{r}_n$ est un estimateur à noyau de la fonction de régression (Bosq et Lecoutre [6], Prakasa Rao [21]). Cette approche permet d'obtenir des estimateurs facilement calculables, tout en restreignant les hypothèses nécessaires, notamment sur la forme des ensembles de niveau.

La qualité de l'approximation est mesurée en terme de différence symétrique entre $\mathcal{L}_n(t)$ et $\mathcal{L}(t)$ (voir Figure 1.2), définie par

$$\mathcal{L}_n(t)\Delta\mathcal{L}(t) = (\mathcal{L}_n(t) \cap \mathcal{L}(t)^C) \cup (\mathcal{L}_n(t)^C \cap \mathcal{L}(t)).$$

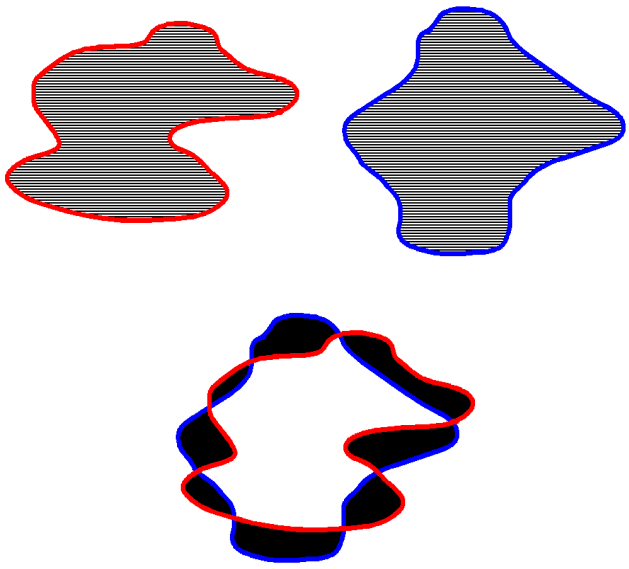

FIGURE 1.2 – Illustration de la différence symétrique (zone noire) entre deux ensembles A (en rouge) et B (en bleu).

1.4 Estimation non paramétrique des ensembles de niveau pour la régression

Ce critère d'erreur est souvent utilisée dans ce genre de problème car elle est facilement interprétable et visualisable.

Sous des hypothèses raisonnables de régularité sur la densité des observation et la fonction de régression r, nous obtenons (Théorème 1.2.2) que

$$\mathbb{E}\,\lambda(\mathcal{L}_n(t)\Delta\mathcal{L}(t)) \to 0,$$

avec une vitesse en $O(\sqrt{nh^d})$.

Nous terminerons ce chapitre par une étude pratique sur des données simulées.

Bibliographie de l'introduction

[1] C. Abraham, G. Biau, and B. Cadre. On the kernel rule for function classification. *Annals of the Institute of Statistical Mathematics*, 58 :619–633, 2006.

[2] A. Berlinet, G. Biau, and L. Rouvière. Functional supervised classification with wavelets. *Annales de l'I.S.U.P.*, 52(1-2) :61–80, 2008.

[3] G. Biau, F. Bunea, and M. H. Wegkamp. Functional classification in Hilbert spaces. *IEEE Trans. Inform. Theory*, 51(6) :2163–2172, 2005.

[4] G. Biau, B. Cadre, and B. Pelletier. A graph-based estimator of the number of clusters. *ESAIM. Probability and Statistics*, 11 :272–280, 2007.

[5] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54 :781–790, 2007.

[6] D. Bosq and J. P. Lecoutre. *Théorie de l'Estimation Fonctionnelle*. Ecole Nationale de la Statistique et de l'Administration Economique et Centre d'Etudes des Programmes Economiques. Economica, 1987.

[7] B. Cadre. Convergent estimators for the L_1-median of a Banach valued random variable. *Statistics*, 35(4) :509–521, 2001.

[8] B. Cadre. Kernel estimation of density level sets. *Journal of Multivariate Analysis*, 97(4) :999–1023, 2006.

[9] L. Cavalier. Nonparametric estimation of regression level sets. *Statistics*, 29(2) :131–160, 1997.

[10] A. Cuevas, W. González-Manteiga, and A. Rodríguez-Casal. Plug-in estimation of general level sets. *Australian and New Zealand Journal of Statistics*, 48(1) :7–19, 2006.

[11] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[12] M. Fromont and C. Tuleau. Functional classification with margin conditions. In *Learning Theory*, Lecture Notes in Computer Science, pages 94–108, 2006.

[13] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Norwell, MA, USA, 1991.

[14] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000.

[15] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

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

[17] T. Linder. Learning-theoretic methods in vector quantization. In *Principles of Nonparametric Learning (Udine, 2001)*, volume 434 of *CISM Courses and Lectures*, pages 163–210. Springer, Vienna, 2002.

[18] D. W. Müller and G. Sawitzki. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86(415) :738–746, 1991.

[19] R. D. Nowak and R. M. Willett. Minimax optimal level-set estimation. *IEEE Transactions on Image Processing*, 16(12) :2965–2979, 2007.

[20] W. Polonik. Measuring mass concentrations and estimating density contour clusters–an excess mass approach. *The Annals of Statistics*, 23(3) :855–881, 1995.

[21] B. L. S. Prakasa Rao. *Nonparametric Functional Estimation*. Academic Press, Orlando, 1983.

[22] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.

[23] L. Rouvière. *Estimation de Densité en Dimension Élevée et Classification de Courbes*. PhD thesis, Université Montpellier II, 2005.

[24] C. Scott and M. Davenport. Regression level set estimation via costsensitive classification. *IEEE Transaction on Signal Processing*, 55 :2752–2757, 2007.

[25] A. B. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3) :948–969, 1997.

[26] C. Tuleau. *Sélection de Variables pour la Discrimination en Grande Dimension, Classification deDonnées Fonctionnelles*. PhD thesis, University Paris XI, 2005.

[27] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

Première partie

Une approche de type k-plus proches voisins pour la régression fonctionnelle

Chapitre 1

A k-Nearest Neighbor approach for functional regression

Ce chapitre est composé d'un article en anglais, publié en 2008 dans la revue *Statistics and Probability Letter* (Volume 78, numéro 10, pages 1189 à 1193).

1.1 Introduction

Regression is the problem of predicting a variable from some observation. An observation is usually supposed to be a collection of numerical measurements represented by a d-dimensional vector. However, in many real-life problems, input data items are in the form of random functions (speech recordings, spectra, images) rather than standard vectors, and this casts the regression problem into the general class of functional data analysis. Even though in practice such observations are observed at discrete sampling points, the challenge in this context is to infer the data structure by exploiting the infinite-dimensional nature of the observations. The last few years have witnessed important developments in both the theory and practice of functional data analysis, and many traditional data analysis tools have been adapted to handle functional inputs. The book of Ramsay and Silverman [5] provides a comprehensive introduction to the area. For an updated list of references, we refer the reader to Cérou and Guyader [3], Rossi and Villa [6], and Tuleau [7].

In the present paper, we consider the functional regression setting, where the goal is to predict a scalar response Y from some infinite-dimensional observations X. More precisely, we will denote by (X, Y) a random pair taking values in $\mathcal{Z} = \mathcal{H} \times \mathbb{R}$, where $\mathcal{H}$ is an infinite dimensional separable Hilbert space. Throughout the document, we will denote by ρ the (unknown) distribution of (X, Y), and by ρ_X the marginal distribution of X. Based on n independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y), we introduce an estimator f_n of the regression function $f_\rho(x) = \mathbb{E}[Y|X = x]$ which is designed as follows : First, we reduce the dimension of $\mathcal{H}$ by considering the first d coefficients of an expansion of each observation in a orthonormal system of $\mathcal{H}$; Second, we perform k-nearest neighbor regression (see Györfi, Kohler, Krzyzak, and Walk [4])in $\mathbb{R}^d$. We select simultaneously the dimension d and the number of neighbors k by a data-splitting device. Our main result states weak consistency of the resulting estimator, thereby extending the general strategy introduced by Biau, Bunea, and Wegkamp [2] in the context of classification (i.e., when Y takes its values in a finite set).

The paper is organised as follows. We start in Section 2.1 by introducing some notation. Then, in Section 2.2, we present the construction of the estimator, and state its weak consistency. Proof are collected in Section 3.

1.2 Consistent functional regression

1.2.1 Notation

We let the symbols $\langle .|. \rangle$ and $\|.\|$ denote the inner product and the associated norm on $\mathcal{H}$, respectively, and we let $(\phi_j)_{j\geq 1}$ be a complete orthonormal system of $\mathcal{H}$ (Akhiezer and Glazman [1]). For each observation X_i, we set $X_{ij} = \langle X_i|\phi_j \rangle$. We know that

$$X_i = \sum_{j=1}^{\infty} X_{ij}\phi_j,$$

where the consistency holds in the $\mathcal{H}$ sense.

Introduce $\mathcal{H}^{(d)}$, the finite-dimensional vector space spanned by the functions $\{\phi_1, \phi_2, \ldots, \phi_d\}$, and let, for each X_i,

$$X_i^{(d)} = \sum_{j=1}^{d} X_{ij}\phi_j.$$

Finally, denote by f_ρ and $f_{\rho,d}$ the regression functions in $\mathcal{H}$ and $\mathcal{H}^{(d)}$, respectively, and by σ_ρ^2 and $\sigma_{\rho,d}^2$ their respective L^2 errors. More precisely, we have $f_\rho(x) = \mathbb{E}[Y|X = x]$, $\sigma_\rho^2 = \int_{\mathcal{Z}}(y - f_\rho(x))^2 d\rho(x, y)$, and the same in $\mathcal{H}^{(d)} \times \mathbb{R}$ for $f_{\rho,d}$ and $\sigma_{\rho,d}^2$. Throughout the document, we suppose that $\mathbb{E}(Y^2) < \infty$ $a.s.$, and all the integrals are to be understood over ρ or ρ_X.

1.2.2 k-nearest neighbor in $\mathcal{H}^{(d)}$

Let us first formally define our k-nearest neighbor type estimator. To this aim, we consider the sequence $(X_1^{(d)}, Y_1), \ldots, (X_n^{(d)}, Y_n)$ where the observations have been projected onto $\mathcal{H}^{(d)}$. For x in $\mathcal{H}^{(d)}$, we reorder the data :

$$\left(X_{(1)}^{(d)}(x), Y_{(1)}(x)\right), \ldots, \left(X_{(n)}^{(d)}(x), Y_{(n)}(x)\right),$$

according to the increasing Euclidean distances $\|X_i^{(d)} - x\|$ of the $X_i^{(d)}$ to x. In other words, $X_{(i)}^{(d)}(x)$ is the i-th nearest neighbor of x amongst $X_j^{(d)}$. If $\|X_i^{(d)} - x\| = \|X_j^{(d)} - x\|$, $X_i^{(d)}$ is declared closer to x if $i < j$. The k-nearest neighbor estimator of f_ρ is then defined (Györfi, Kohler, Krzyzak, and Walk [4]) as

$$f_{n,k,d}(x) = \frac{1}{k}\sum_{i=1}^{k} Y_{(i)}(x). \tag{1.1}$$

To select simultaneously the dimension d and the number of neighbors k, we suggest the following data-splitting device. First we split the data into a training set $\{(X_i, Y_i), i \in \mathcal{I}_\ell\}$ of length ℓ, and a validation set $\{(X_j, Y_j), j \in \mathcal{J}_m\}$ of length m, with $m + \ell = n$ (ℓ and m possibly function of n). For each $d \geq 1$,

$1 \leq k \leq \ell$, we construct a k-nearest neighbor estimator based on the training set. Second we use the validation set to select $\hat{d}$ and $\hat{k}$ as follows :

$$(\hat{d}, \hat{k}) \in \underset{d \geq 1, 1 \leq k \leq \ell}{\arg\min} \left[\frac{1}{m} \sum_{j \in \mathcal{J}_m} \left(Y_j - f_{\ell,k,d}(X_j^{(d)}) \right)^2 + \frac{\lambda_d}{\sqrt{m}} \right]. \tag{1.2}$$

Here, the term $\lambda_d/\sqrt{m}$ is a given penalty term which tends to infinity with d to prevent overfitting.

This method, which is computationnaly simple, leads to the estimator

$$\hat{f}_n(x) := f_{\ell,\hat{k},\hat{d}}(x^{(\hat{d})}), \tag{1.3}$$

which has an error

$$\mathcal{E}(\hat{f}_n) = \int_{\mathcal{Z}} \left(y - \hat{f}_n(x) \right)^2 d\rho(x,y) = \int_{\mathcal{H}} \left(\hat{f}_n(x) - f_\rho(x) \right)^2 d\rho_X(x) \ + \sigma_\rho^2 \, .$$

The estimator f_n satisfies the following oracle inequality :

Proposition 1.2.1 *Let M be a positive constant such that $Y^2 \leq M$ a.s., and suppose that*

$$\Delta := \sum_{d=1}^{\infty} e^{-2(\lambda_d/M)^2} < \infty.$$

Then there exists a constant $c > 0$, only depending on Δ and M, such that, for every integer $\ell > 1/\Delta$ and $m = n - \ell$,

$$\begin{aligned} &\mathbb{E} \int_{\mathcal{H}} \left(\hat{f}_n(x) - f_\rho(x) \right)^2 \\ &\leq \inf_{d \geq 1} \left[(\sigma_{\rho,d}^2 - \sigma_\rho^2) + \inf_{1 \leq k \leq \ell} \left(\mathbb{E} \int_{\mathcal{H}^{(d)}} \left(f_{\ell,k,d}(x) - f_{\rho,d}(x) \right)^2 \right) + \frac{\lambda_d}{\sqrt{m}} \right] + c \sqrt{\frac{\ln \ell}{m}} \, . \end{aligned} \tag{1.4}$$

The term $\sigma_{\rho,d}^2 - \sigma_\rho^2$ may be viewed as the price to be paid for using a finite dimensional approximation of the observations, and it converges to zero by Lemma 1.3.1 below. The term $\inf_{1 \leq k \leq \ell} \left(\mathbb{E} \int_{\mathcal{H}^{(d)}} (f_{\ell,k,d}(x) - f_{\rho,d}(x))^2 \right)$ converges also to zero by Lemma 1.3.2. Since the infimum is taken over all $d \geq 1$, weak convergence of $\hat{f}_n(x)$ to $f_\rho(x)$ is ensured.

Theorem 1.2.1 *Under the assumption of Proposition 1.2.1 and*

$$\lim_{n\to\infty} \ell = \infty, \lim_{\ell\to\infty} k = \infty, \lim_{\ell\to\infty} \frac{k}{\ell} = 0, \text{ and } \lim_{n\to\infty} \frac{\ln \ell}{m} = 0,$$

$\hat{f}_n$ *weakly converges to* f_ρ*, i.e.,*

$$\mathbb{E} \int_{\mathcal{H}} \left(\hat{f}_n(x) - f_\rho(x) \right)^2 \to 0 \text{ as } n \to \infty.$$

Pratically speaking, as discussed in Biau, Bunea, and Wegkamp [2], choosing the penalty in (1.2) is not an easy task. Indeed, an abusive penalisation of high dimensions can mask helpful information. For a more involved discussion about the penalty choice, and experimental results, we refer the reader to Tuleau [7], who shows that that adding a penalty term improves the stability of the selected dimension d. As an example, one can choose $\lambda_d = \sqrt{d}/n$.

1.3 Proof

Proof of Proposition 1.2.1 Let

$$L(k,d) = \mathbb{E}\left[\left(Y - f_{\ell,k,d}\left(X^{(d)}\right) \right)^2 \mid (X_i, Y_i), 1 \le i \le n \right],$$

and

$$\hat{L}(k,d) = \frac{1}{m} \sum_{j \in \mathcal{J}_m} \left(Y_j - f_{\ell,k,d}(X_j^{(d)}) \right)^2.$$

We have to minimize $\hat{L}(k,d) + \lambda_d/m$ in k and d.

Fix $\varepsilon > 0$. For every $d \ge 1$ and every k satisfying $1 \le k \le \ell$, we may write

$$\mathbb{P}\left\{ L(\hat{k},\hat{d}) - \hat{L}(k,d) > \frac{\lambda_d}{\sqrt{m}} + \varepsilon \right\} \le \mathbb{P}\left\{ L(\hat{k},\hat{d}) - \hat{L}(\hat{k},\hat{d}) > \frac{\lambda_{\hat{d}}}{\sqrt{m}} + \varepsilon \right\},$$

since, by definition of $(\hat{k},\hat{d})$,

$$\hat{L}(\hat{k},\hat{d}) + \frac{\lambda_{\hat{d}}}{\sqrt{m}} \le \hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}}\ .$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{L(\hat{k},\hat{d}) - \hat{L}(k,d) > \lambda_d/\sqrt{m} + \varepsilon\right\} \\
&\quad\leq \sum_{d=1}^{\infty}\sum_{k=1}^{\ell} \mathbb{P}\left\{L(k,d) - \hat{L}(k,d) > \lambda_d/\sqrt{m} + \varepsilon\right\} \\
&\quad\quad \text{(by the union bound)} \\
&\quad= \sum_{d=1}^{\infty}\sum_{k=1}^{\ell} \mathbb{E}\mathbb{P}\left\{L(k,d) - \hat{L}(k,d) > \lambda_d/\sqrt{m} + \varepsilon \mid (X_i,Y_i), i \in \mathcal{I}_\ell\right\} \\
&\quad\leq \sum_{d=1}^{\infty} \ell \exp\left\{ -2[(\lambda_d/\sqrt{m}) + \varepsilon]^2 \times (m/M^2)\right\} \\
&\quad\quad \text{(by Hoeffding's inequality)} \\
&\quad\leq \ell e^{-2m\varepsilon^2/M^2} \sum_{d=1}^{\infty} e^{-2(\lambda_d/M)^2} \\
&\quad= \Delta\ell e^{-2m\varepsilon^2/M^2},
\end{aligned}$$

where $\Delta = \sum_{d=1}^{\infty} e^{-2(\lambda_d/M)^2}$. Since, for every $d \geq 1$ and k with $1 \leq k \leq \ell$,

$$\mathbb{E}L(\hat{k},\hat{d}) \leq \mathbb{E}\hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}} + \int_0^{\infty} \mathbb{P}\left\{L(\hat{k},\hat{d}) - \hat{L}(k,d) > \frac{\lambda_d}{\sqrt{m}} + \varepsilon\right\} d\varepsilon \ ,$$

we obtain, for every $u > 0$,

$$\mathbb{E}L(\hat{k},\hat{d}) \leq \mathbb{E}\hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}} + u + \Delta\ell \int_u^{\infty} e^{-2m\varepsilon^2/M^2} d\varepsilon \ .$$

Note that

$$\begin{aligned}
\int_u^{\infty} e^{-2m\varepsilon^2/M^2} d\varepsilon &\leq \frac{1}{2}\int_u^{\infty} \left(2 + \frac{M^2}{2m\varepsilon}\right) e^{-2m\varepsilon^2/M^2} d\varepsilon \\
&= -\frac{1}{2}\left[\frac{M^2}{2m\varepsilon} e^{-2m\varepsilon^2/M^2}\right]_u^{\infty} \\
&= \frac{M^2}{4mu} e^{-2mu^2/M^2}.
\end{aligned}$$

Whence, choosing $u = M\sqrt{\ln(\Delta\ell)/2m}$, we obtain

$$\mathbb{E}L(\hat{k},\hat{d}) \leq \mathbb{E}\hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}} + M\sqrt{\frac{\ln(\Delta\ell)}{2m}} + \frac{M}{2\sqrt{2m\ln(\Delta\ell)}} \ .$$

Since k and d are arbitrary,

$$\mathbb{E}L(\hat{k},\hat{d}) \leq \inf_{d\geq 1, 1\leq k\leq \ell} \mathbb{E}\hat{L}(k,d) + \frac{\lambda_d}{\sqrt{m}} + M\sqrt{\frac{\ln(\Delta\ell)}{2m}} + \frac{M}{2\sqrt{2m\ln(\Delta\ell)}} \; .$$

The fact that $\mathbb{E}\hat{L}(k,d) = \mathbb{E}L(k,d)$ for each fixed k,d leads to the inequality (1.4). $\square$

Proof of Theorem 1.2.1 will rely on the following lemma.

Lemma 1.3.1 *We have*

$$\sigma^2_{\rho,d} - \sigma^2_\rho \to 0 \; as \; d \to \infty \; .$$

Proof of Lemma 1.3.1

$$\begin{aligned}
\sigma^2_{\rho,d} - \sigma^2_\rho &= \mathbb{E}\Big[Y - \mathbb{E}\big[Y|X^{(d)}\big]\Big]^2 - \mathbb{E}\Big[Y - \mathbb{E}\big[Y|X\big]\Big]^2 \\
&= \mathbb{E}\Big[Y - \mathbb{E}\big[Y|X\big]\Big]^2 + \mathbb{E}\Big[\mathbb{E}\big[Y|X\big] - \mathbb{E}\big[Y|X^{(d)}\big]\Big]^2 - \mathbb{E}\Big[Y - \mathbb{E}\big[Y|X\big]\Big]^2 \\
&= \mathbb{E}\Big[\mathbb{E}\big[Y|X\big] - \mathbb{E}\big[Y|X^{(d)}\big]\Big]^2 .
\end{aligned}$$

Since $\mathbb{E}[Y^2] < \infty$, the sequence $\Big(\mathbb{E}[Y|X^{(d)}]\Big)_{d\geq 1}$ is a L^2 bounded martingale, therefore we have

$$\mathbb{E}[Y|X^{(d)}] \to \mathbb{E}[Y|X] \text{ in the } L^2 \text{ sense as } d \to \infty. \qquad \square$$

Lemma 1.3.2 *Assume that $k \to \infty$ and $k/\ell \to 0$ as $\ell \to \infty$. Then, for any fixed d,*

$$\mathbb{E}\int_{\mathcal{H}^{(d)}} \Big(f_{\ell,k,d}(x) - f_{\rho,d}(x)\Big)^2 \to 0 \; as \; \ell \to \infty$$

Proof of Lemma 1.3.2 See Györfi, Kohler, Krzyzak, and Walk [4], Theorem 6.1, page 88. $\square$

We are now in a position to prove Theorem 1.2.1.

Proof of Theorem 1.2.1 Fix $\varepsilon > 0$. By Lemma 1.3.1, we know there exist d_0 such that $\sigma_{\rho,d} - \sigma_\rho^2 \leq \varepsilon$ for all $d \geq d_0$. Then, by Lemma 1.3.2, we have

$$\mathbb{E} \int_{\mathcal{H}^{(d_0)}} \Big(f_{\ell,k,d_0}(x) - f_{\rho,d_0}(x) \Big)^2 \to 0 \; as \; \ell \to \infty.$$

Finally by Proposition 1.2.1 we have :

$$\begin{aligned}
&\mathbb{E} \int_{\mathcal{H}} \Big(\hat{f}_n(x) - f_\rho(x) \Big) \\
&\leq \inf_{d \geq 1} \left[(\sigma_{\rho,d}^2 - \sigma_\rho^2) + \inf_{1 \leq k \leq \ell} \left(\mathbb{E} \int_{\mathcal{H}^{(d)}} \Big(f_{\ell,k,d}(x) - f_{\rho,d}(x) \Big)^2 \right) + \frac{\lambda_d}{\sqrt{m}} \right] + c \sqrt{\frac{\ln \ell}{m}} \\
&\leq (\sigma_{\rho,d_0}^2 - \sigma_\rho^2) + \inf_{1 \leq k \leq \ell} \left(\mathbb{E} \int_{\mathcal{H}^{(d_0)}} \Big(f_{\ell,k,d_0}(x) - f_{\rho,d_0}(x) \Big)^2 \right) + \frac{\lambda_{d_0}}{\sqrt{m}} + c \sqrt{\frac{\ln \ell}{m}} \\
&\leq \varepsilon + o(1), \text{ as } n \to \infty.
\end{aligned}$$

Since ε is arbitrary the convergence is ensured. $\square$

Bibliography

[1] N. I. Akhiezer and I. M. Glazman. *Theory of Linear Operators in Hilbert Space*. Frederick Ungar Publishing Co., New York, 1963.

[2] G. Biau, F. Bunea, and M. H. Wegkamp. Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory*, 51(6) :2163–2172, 2005.

[3] F. Cérou and A. Guyader. Nearest neighbor classification in infinite dimension. *ESAIM. Probability and Statistics*, 10 :340–355, 2006.

[4] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

[5] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, 2002.

[6] F. Rossi and N. Villa. Support vector machine for functional data classification. *Neurocomputing*, 69 :730–742, 2006.

[7] C. Tuleau. *Sélection de Variables pour la Discrimination en Grande Dimension, Classification de Données Fonctionnelles*. PhD thesis, Université Paris XI, 2005.

Deuxième partie

Classification fonctionnelle non supervisée

Chapitre 1

Quantification de courbes dans un espace de Banach

Ce chapitre à fait l'objet d'un article en anglais, soumis en 2009 dans la revue *Mathematical methods of statistics*. Cet article est donné en Annexe de cette thèse.

1.1 Introduction

Le clustering, ou l'art de repérer des groupes dans des données, est un sujet aux multiples ramifications, tant sur le plan théorique que pratique. Il y a environ 45 ans, des biologistes et des sociologues ont commencé à chercher des méthodes automatiques pour répartir leurs données en différents groupes. Au fil des ans de nombreux ouvrages sur le sujet ont vu le jour. Citons par exemple les ouvrages de Duda et Hart [10], Gan, Ma et Wu [14], Hartigan [17], Kogan [22], ou encore Mirkin [28]. Aujourd'hui, le clustering est utilisé dans des domaines très variés, comme par exemple l'intelligence artificielle, la sociologie, le marketing, la recherche médicale, ou encore les sciences politiques.

Dans de nombreux domaines de la statistique contemporaine, les données prennent la forme de courbes (ou surfaces) aléatoires. Ces courbes peuvent, par exemple, représenter l'intensité d'une source lumineuse, la quantité de CO_2 rejetée par une voiture, le taux de sucre dans le sang, ou encore le tracé

d'un électrocardiogramme. Les appareils de mesure ne captant que la valeur de la courbe à certains instants, les données disponibles ne sont en fait que des versions discrétisées de la courbe. Toutefois, le statisticien a intérêt à prendre en compte le caractère continu de ces observations (par exemple pour prendre en compte la régularité des fonctions). Il devient alors intéressant de ne plus considérer ces données 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. La littérature s'est récemment étoffée de résultats prenant en compte la nature fonctionnelle des données. Pour de plus amples références bibliographiques, nous renvoyons le lecteur à l'ouvrage de Ramsay et Silverman [30]. Dans ce travail, nous envisageons le problème du clustering dans un espace général, ce qui permet, en particulier, de prendre en compte des observations fonctionnelles, c'est-à-dire d'observations prenant leurs valeurs dans un espace de dimension infinie.

Nous nous intéresserons à un clustering par partitionnement. L'idée est de classer les données en k groupes satisfaisant les deux conditions suivantes :

1. Chaque groupe contient au moins une donnée ;
2. Chaque donnée appartient à un et un seul groupe.

La seconde condition impose que deux groupes ne peuvent avoir de données en commun et que les k groupes contiennent toutes les données. La Figure 1.1 montre un exemple de points répartis en cinq groupes.

Voici quelques exemples d'utilisation du clustering :

- On dispose des notes des élèves d'un lycée. Peut-on former des groupes ?
- À partir de données sur des bancs de poissons, peut-on différencier des types de bancs ?

Les données du deuxième exemple peuvent prendre la forme de courbes aléatoires, autrement dit des objets à valeur dans un espace de dimension infinie.

1.1 Introduction

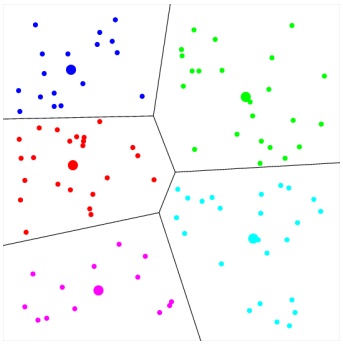

FIGURE 1.1 – Un exemple de répartition des données en 5 groupes.

La méthode de clustering que nous utilisons repose sur la technique dite de quantification, très utilisée en compression du signal (Graf et Luschgy [16], Linder [25]). Étant donné un espace normé $(\mathcal{H}, \|\|)$, on se donne un sous-ensemble fini $\mathcal{C}$ de $\mathcal{H}$, que l'on nomme alphabet et on représente chaque élément x de $\mathcal{H}$ par un élément (et un seul) $\hat{x}$ de cet alphabet. La fonction, notée q, qui à chaque donnée associe sa représentation (on a donc $q(x) = \hat{x}$) est appelée quantificateur. Notons de plus que nous avons décidé de travailler avec une taille d'alphabet k fixée.

Soit maintenant X une variable aléatoire à valeurs dans $\mathcal{H}$ de loi μ. Afin de mesurer la qualité de l'approximation effectuée en remplaçant X par sa quantification $\hat{X}$, on se donne une fonction $d : \mathcal{H} \times \mathcal{H} \to \mathbb{R}^+$ mesurable, que nous appellerons mesure de distorsion. La distorsion est alors définie par $\mathbb{E}\, d(X, \hat{X})$. L'objectif est d'optimiser la qualité de la quantification, c'est-à-dire de minimiser la distorsion $\mathbb{E}\, d(X, \hat{X})$ sur l'ensemble de tous les quantificateurs possibles.

Dans ce travail, nous avons choisi de considérer la mesure de distorsion donnée par la distance L_1, c'est-à-dire $d(x, y) = \|x - y\|$. Rappelons que, dans le cas où $k = 1$, un point y^* qui minimise $\mathbb{E}\|X - y\|$ est appelé une médiane. On peut montrer que l'ensemble des médianes est naturellement robuste face

à de petites perturbations de la loi de X (voir Kemperman [21]). Cette remarque motive le choix du critère L_1 dans le cas d'un problème à k-centres. En particulier, dans le cas pratique où la loi μ est inconnue et où nous disposons, en lieu et place de μ, d'un échantillon d'observations $(X_1, \ldots, X_n)$ indépendantes et identiquement distribuées (i.i.d.) de loi μ, cette robustesse naturelle permet d'atténuer l'effet de valeurs aberrantes dans l'échantillon.

Le document est organisé de la façon suivante. Nous commençons dans la Section 2 par définir le cadre général de notre travail. Nous montrons ensuite que les quantificateurs optimaux (au sens de la distorsion) existent, et qu'ils sont obtenus dans une famille de quantificateurs particuliers, dits de type plus proches voisins. Nous abordons ensuite dans la Section 3 le problème statistique en présentant différents estimateurs des ces quantificateurs optimaux, ainsi que les algorithmes qui permettent de les calculer. Enfin nous présentons dans la dernière section une étude pratique sur des données simulées et réelles.

1.2 Quantification dans un espace de Banach général

1.2.1 Cadre général

Dans toute la suite, $(\mathcal{H}, \|.\|)$ désigne un espace de Banach réflexif (c'est-à-dire un espace pour lequel la boule unité est faiblement compacte) séparable (voir Dunford et Schwartz [12]), et X représente une variable aléatoire de loi μ à valeurs dans $\mathcal{H}$ telle que $\mathbb{E}\|X\| < \infty$. Insistons sur le fait que nous n'imposons aucune restriction a priori sur la dimension de $\mathcal{H}$.

Définition 1.2.1 *Soit $\mathcal{C} = \{y_i\}_{i=1}^k$ un alphabet, c'est-à-dire un ensemble de k points distincts de $\mathcal{H}$, appelés centres. On appelle k-quantificateur toute application borélienne $q : \mathcal{H} \to \mathcal{C}$.*

Posons, pour tout (x, y) dans $\mathcal{H} \times \mathcal{H}$:

$$d(x, y) = \|x - y\|.$$

Afin de mesurer l'erreur commise en représentant X par $q(X)$, nous introduisons la distorsion $D(\mu, q)$ définie par

$$D(\mu, q) = \mathbb{E}\, d(X, q(X)) = \int_{\mathcal{H}} d(x, q(x))\mu(dx).$$

Notons d'emblée que $D(\mu, q) < \infty$, car $\mathbb{E}\|X\| < \infty$. Pour une taille d'alphabet k fixée, nous allons chercher à minimiser $D(\mu, q)$ sur l'ensemble $\mathcal{Q}_k$ de tous les k-quantificateurs. Pour cela, nous définissons

$$D_k^*(\mu) = \inf_{q \in \mathcal{Q}_k} D(\mu, q).$$

Lorsque cet infimum est atteint pour un quantificateur particulier q^*, nous dirons que q^* est un quantificateur optimal, ce qui signifie que $D(\mu, q^*) = D_k^*(\mu)$.

On peut remarquer que tout quantificateur est caractérisé par :

1. Son alphabet $\mathcal{C} = \{y_i\}_{i=1}^k$ d'une part ;
2. Une partition de $\mathcal{H}$ en cellules $S_i = \{x \in \mathcal{H} : q(x) = y_i\},\ i = 1, \ldots, k$ d'autre part,

via la règle

$$q(x) = y_i \iff x \in S_i.$$

Dans la suite nous définirons donc un quantificateur par son alphabet et ses cellules.

1.2.2 Quantizeurs de type plus proches voisins

Partition de Voronoï et centroïdes

Définition 1.2.2 *Une partition $\{S_i\}_{i=1}^k$, associée à un alphabet $\mathcal{C} = \{y_i\}_{i=1}^k$, est dite partition de Voronoï si elle est définie comme suit :*

$$S_1 = \{x \in \mathcal{H} : \|x - y_1\| \leq \|x - y_j\|,\ j = 1, \ldots, k\},$$

et, pour $i = 2, \ldots, k$,

$$S_i = \{x \in \mathcal{H} : \|x - y_i\| \leq \|x - y_j\|,\ j = 1, \ldots, k\} \setminus \bigcup_{k=1}^{i-1} S_k.$$

La Figure 1.2 présente un exemple de partition de Voronoï dans $\mathbb{R}^2$ pour la distance euclidienne.

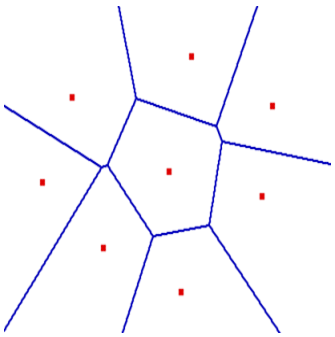

FIGURE 1.2 – Un exemple de partition de Voronoï dans $\mathbb{R}^2$ pour la distance euclidienne.

Remarque : Dans la définition ci dessus, on enlève $\bigcup_{k=1}^{i-1} S_k$ à S_i pour éviter les ambiguïtés sur les frontières des cellules.

Définition 1.2.3 *On dit qu'un quantificateur d'alphabet* $\mathcal{C} = \{y_i\}_{i=1}^k$ *est de type plus proches voisins s'il admet pour partition la partition de Voronoï associée à* $\mathcal{C}$.

Nous allons maintenant montrer qu'il suffit de considérer des quantificateurs de type plus proches voisins.

Lemme 1.2.1 (Condition des plus proches voisins) *Soit* q *un quantificateur d'alphabet* $\mathcal{C} = \{y_i\}_{i=1}^k$ *et de partition associée* $\{S_i\}_{i=1}^k$. *Soit* q' *le quantificateur de même alphabet et de partition associée la partition de Voronoï. Alors*

$$D(\mu, q') \leq D(\mu, q).$$

Preuve du Lemme 1.2.1 On peut remarquer que :

$$\|x - q'(x)\| = \min_{y \in \mathcal{C}} \|x - y_i\|.$$

Ainsi nous pouvons écrire

$$\begin{aligned} D(\mu, q) &= \mathbb{E}\|X - q(X)\| \\ &= \sum_{j=1}^{k} \int_{S_j} \|x - y_j\| \mu(dx) \\ &\geq \sum_{j=1}^{k} \int_{S_j} \min_{y \in \mathcal{C}} \|x - y\| \mu(dx) \\ &= \int_{\mathcal{H}} \min_{y \in \mathcal{C}} \|x - y\| \mu(dx) \\ &= \mathbb{E} \|X - q'(X)\| = D(\mu, q'). \qquad \square \end{aligned}$$

Nous pouvons en déduire immédiatement :

Corollaire 1.2.1 *Si un quantificateur q^* optimal existe, il appartient aux quantificateur de type plus proches voisins. On a de plus*

$$D_k^*(\mu) = \inf_{\mathcal{C} \subset \mathcal{H} \,:\, |\mathcal{C}| = k} \mathbb{E} \min_{y \in \mathcal{C}} \|X - y\|.$$

Nous montrerons dans la section suivante que cet infimum est atteint, c'est-à-dire qu'il existe un alphabet $\mathcal{C}^* = \{y_i^*\}_{i=1}^k$ tel que

$$\mathbb{E} \min_{y^* \in \mathcal{C}^*} \|X - y^*\| = \inf_{\mathcal{C} \subset \mathcal{H} : |\mathcal{C}| = k} \mathbb{E} \min_{y \in \mathcal{C}} \|X - y\|.$$

Le lemme suivant permet, pour une partition donnée, de choisir un alphabet optimal. Rappelons au préalable que $\mathcal{H}$ étant un espace de Banach réflexif séparable, toute loi de probabilité sur $\mathcal{H}$ admet au moins une médiane (Kemperman [21]). Nous pouvons donc énoncer :

Lemme 1.2.2 (Condition des centroïdes) *Soit q un quantificateur de partition $\{S_i\}_{i=1}^k$ telle que $\mu(S_i) > 0$ pour $i = 1, \ldots, k$. Si q' est un quantificateur admettant les mêmes cellules et dont les centres $y_1', \ldots, y_k'$ sont des médianes des lois de X conditionnées par $[X \in S_1], \ldots, [X \in S_k]$, c'est-à-dire,*

$$y_i' \in \arg\min_{y \in \mathcal{H}} \mathbb{E} \left[\|X - y\| \mid X \in S_i \right], \; i = 1, \ldots, k,$$

alors

$$D(\mu, q') \leq D(\mu, q).$$

Preuve du Lemme 1.2.2 On a

$$\begin{aligned} D(\mu, q) &= \mathbb{E} \, \|X - q(X)\| \\ &= \sum_{i=1}^{k} \mathbb{E} \, [\|X - y_i\| | X \in S_i] \mu(S_i) \\ &\geq \sum_{i=1}^{k} \mathbb{E} \, [\|X - y_i'\| | X \in S_i] \mu(S_i) \\ &= \mathbb{E} \, \|X - q'(X)\| = D(\mu, q'). \qquad \square \end{aligned}$$

Ces deux lemmes vont nous permettre d'établir une méthode simple pour construire un "bon" quantificateur (au sens de la distorsion).

Une méthodologie simple : l'algorithme de Lloyd

Du Lemme 1.2.1 et du Lemme 1.2.2 nous pouvons facilement déduire un algorithme simple permettant de construire un k-quantificateur, l'algorithme de Lloyd (Gersho et Gray [15], Chapitre 6). La base de cet algorithme est l'itération de Lloyd (voir Figure 1.3) qui permet, à partir d'un alphabet donné, d'en construire un meilleur de la manière suivante.

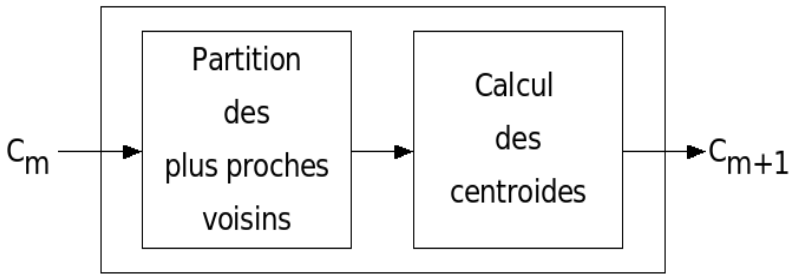

FIGURE 1.3 – Itération de Lloyd.

L'algorithme procède de la façon suivante :

1. Étant donné un alphabet C_m, on lui associe la partition de Voronoï correspondante ;

2. On utilise ensuite la condition des centroïdes pour construire C_{m+1}, l'alphabet optimal pour la partition calculée juste avant.

Grâce aux deux conditions d'optimalité (celle des plus proches voisins et celle des centroïdes), chaque itération diminue ou laisse inchangée la distorsion. L'algorithme de Lloyd peut alors être visualisé de manière concise comme dans le Tableau 1.1. La méthodologie générale est résumée dans la Figure 1.4.

Tableau 1: L'algorithme de Lloyd
Etape 1. Soit C_1 un alphabet initial et posons m = 1.
Etape 2. Étant donné l'alphabet C_m, on utilise l'itération de Lloyd pour construire l'alphabet C_{m+1},
Etape 3. On calcule D_{m+1} la distorsion de C_{m+1}. Si D_m- D_{m+1} est plus petit qu'un seuil fixé l'algorithme s'arrête. Sinon on pose m = m + 1 et on retourne a l'étape 2.

TABLE 1.1 – Algorithme de Lloyd.

Malheureusement cet algorithme présente deux défauts. Le premier est qu'il dépend de l'alphabet initial choisi (il faut éviter en particulier de choisir des alphabets initiaux concentrés dans une petite partie de l'espace $\mathcal{H}$). Le second est que, même si D_m est effectivement affaiblie à chaque itération, elle ne converge pas forcément vers la distorsion optimale. Nous verrons par la suite (Section 3) d'autre méthodes permettant de construire des quantificateurs optimaux.

Remarque : En pratique μ est inconnue et nous présenterons dans la Section 3 une version de cet algorithme utilisant la mesure empirique μ_n en lieu et

place de μ.

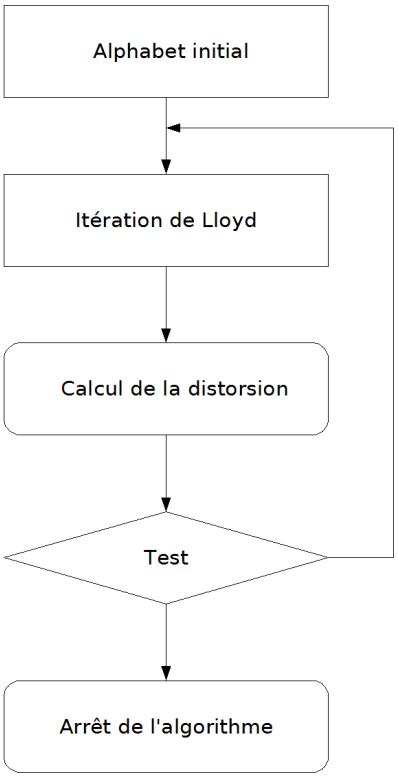

FIGURE 1.4 – Diagramme de l'algorithme de Lloyd.

1.2.3 Minimisation de la distorsion pour les quantificateurs de type plus proches voisins

Nous allons montrer dans cette sous-section que le problème de la minimisation de $D(\mu, q)$ admet au moins une solution. En vertu du Lemme 1.2.1, nous ne considérerons à présent que des quantificateurs de type plus proches voisins. Nous n'aurons ainsi besoin pour les caractériser que de leur alphabet.

Notations, définitions et résultats préliminaires

Notons $\mathbf{y}_k$ un élément de $\mathcal{H}^k$ de coordonnées $(y_1, \ldots, y_k)$. Ne considérant que des partitions de Voronoï, nous pouvons écrire, pour un quantificateur donné

q d'alphabet $\mathbf{y}_k$,

$$D(\mu, q) = D(\mu, \mathbf{y}_k).$$

Montrer l'existence de solutions à notre problème de minimisation est alors équivalent à montrer que la fonction $D(\mu, .)$ admet au moins un minimum $\mathbf{y}_k^*$, ce qui sera assuré par le Théorème 1.2.1 ci-après. Rappelons au préalable quelques définitions importantes.

Définition 1.2.4 *Une fonction $\phi : \mathcal{H} \to \bar{\mathbb{R}}$ est dite semi-continue inférieurement pour la topologie faible (en abrégé faiblement s.c.i.) si elle satisfait à l'une des conditions équivalentes suivantes :*

(i) $\forall t \in \mathbb{R}, \{u \in \mathcal{H} : \phi(u) \leq t\}$ *est fermé pour la topologie faible.*

(ii) $\forall \bar{u} \in \mathcal{H}, \liminf_{u \xrightarrow{w} \bar{u}} \phi(u) \geq \phi(\bar{u})$ *(où $\xrightarrow{w}$ note la convergence faible dans $\mathcal{H}$).*

Pour une preuve de l'équivalence entre les deux conditions de la Définition 1.2.4, et plus de précisions sur les notions de semi-continuité inférieure et topologie faible, le lecteur est invité à consulter le livre de Ekeland et Teman [13]. On y trouve aussi la démonstration des propriétés suivantes.

Propriétés 1.2.1 *Avec les notations de la Définition 1.2.4 :*
(i) Si ϕ est continue convexe, alors elle est faiblement s.c.i.
(ii) Si ϕ est faiblement s.c.i. sur un compact B pour la topologie faible, elle possède un minimum sur B.

Nous noterons enfin B_R la boule fermée dans $(\mathcal{H}, \|.\|)$ de centre 0 et de rayon R.

Existence d'un alphabet optimal

Nous sommes maintenant en mesure d'énoncer et prouver le résultat principal de cette section :

Théorème 1.2.1 *Supposons que $\mathcal{H}$ est un espace de Banach réflexif et séparable. La fonction $D(\mu, .)$ admet au moins un minimum, autrement dit il existe au moins un alphabet optimal.*

Afin de prouver le Théorème 1.2.1, nous allons montrer qu'il suffit de minimiser $D(\mu, .)$ sur une boule fermée bornée.

Lemme 1.2.3 *Il existe $A > 0$ et $\ell \leq k$ tels que*

$$\inf_{\mathbf{y}_k \in \mathcal{H}^k} D(\mu, \mathbf{y}_k) = \inf_{\mathbf{y}_\ell \in B_A^\ell} D(\mu, \mathbf{y}_\ell).$$

Preuve du Lemme 1.2.3 Rappelons que $\mathbb{E}\,\|X\| < \infty$. On suppose que le support de μ n'est pas réduit à un point, de sorte que $D_1^*(\mu) > D_2^*(\mu)$. Soit alors ℓ l'unique entier $2 \leq \ell \leq k$ (on écarte le cas trivial où $k = 1$) tel que

$$D_k^*(\mu) = \ldots = D_\ell^*(\mu) < D_{\ell-1}^*(\mu)$$

(si le support de μ contient au moins k points, alors $\ell = k$). Pour $\varepsilon > 0$ tel que

$$\varepsilon < \frac{1}{2}\left(D_{\ell-1}^*(\mu) - D_\ell^*(\mu)\right),$$

considérons $0 < r < R$ tel que

$$(R - r)\mu(B_r) > D_\ell^*(\mu) + \varepsilon, \text{ et } 2\int_{B_{2R}^C} \|x\|\mu(dx) < \varepsilon. \tag{1.1}$$

Soit enfin $\mathbf{y}_\ell$ vérifiant $D(\mu, \mathbf{y}_\ell) < D_\ell^*(\mu) + \varepsilon$ avec, sans perte de généralité, $\|y_1\| \leq \ldots \leq \|y_\ell\|$. On a alors que $\|y_1\| \leq R$. En effet, dans le cas contraire, nous aurions $\min_{i=1,\ldots,\ell} \|x - y_i\| \geq (R - r)$ pour tout x dans B_r, et donc

$$D_\ell^*(\mu) + \varepsilon > \int_{B_r} \min_{i=1,\ldots,\ell} \|x - y_i\|\mu(dx) \geq (R - r)\mu(B_r),$$

ce qui contredit (1.1).

Nous allons maintenant montrer que, pour tout $j = 2, \ldots, \ell$, nous avons $\|y_j\| \leq 5R$. Supposons au contraire que $\|y_\ell\| > 5R$. Alors pour tout x dans $\mathcal{H}$,

$$\|x - y_1\| \leq \|x - y_\ell\| \mathbf{1}_{x \in B_{2R}} + 2\|x\| \mathbf{1}_{x \in B_{2R}^C}. \tag{1.2}$$

Soit $\{S_i\}_{i=1}^{\ell}$ la partition de Voronoï associée à $\{y_i\}_{i=1}^{\ell}$. Alors

$$\begin{aligned} D(\mu, \mathbf{y}_{\ell-1}) &= \sum_{j=1}^{\ell} \int_{S_j} \min_{i=1,\ldots,\ell-1} \|x - y_i\| \mu(dx) \\ &\leq \sum_{j=1}^{\ell-1} \int_{S_j} \|x - y_j\| \mu(dx) + \int_{S_\ell} \|x - y_1\| \mu(dx) \\ &\leq \sum_{j=1}^{\ell} \int_{S_j} \|x - y_j\| \mu(dx) + 2 \int_{B_{2R}^C} \|x\| \mu(dx) \\ &\text{(d'après (1.2))} \\ &\leq D(\mu, \mathbf{y}_\ell) + \varepsilon \leq D_\ell^*(\mu) + 2\varepsilon \\ &< D_{\ell-1}^*(\mu) \\ &\text{(d'après(1.1)),} \end{aligned}$$

ce qui est impossible. Nous obtenons donc

$$D(\mu, \mathbf{y}_\ell) < D_\ell^*(\mu) + \varepsilon \implies \mathbf{y}_\ell \in (B(5R))^\ell.$$

On conclut en posant $A = 5R$ que

$$\inf_{\mathbf{y}_k \in \mathcal{H}^k} D(\mu, \mathbf{y}) = D_k^*(\mu) = D_\ell^*(\mu) = \inf_{\mathbf{y}_\ell \in B_A^\ell} D(\mu, \mathbf{y}_\ell). \qquad \square$$

Pour tout x dans $\mathcal{H}$, définissons les fonctions $g_{i,x} : \mathcal{H}^k \to \mathbb{R}$ et $g_x : \mathcal{H}^k \to \mathbb{R}$ par :

$$g_{i,x}(\mathbf{y}_k) = \|x - y_i\|,$$

et

$$g_x(\mathbf{y}_k) = \min_{i=1,\ldots,k} g_{i,x}(\mathbf{y}_k).$$

Lemme 1.2.4 *Pour tout x dans $\mathcal{H}$, la fonction g_x est faiblement s.c.i. sur $\mathcal{H}^k$.*

Preuve du Lemme 1.2.4 Pour chaque x dans $\mathcal{H}$, les fonctions $g_{i,x}$ sont continues et convexes, elles sont donc faiblement s.c.i. d'après la Propriété 1.2.1 (i). Pour tout t dans $\mathbb{R}$, les ensembles

$$\{\mathbf{y}_k \in \mathcal{H}^k : g_{i,x}(\mathbf{y}_k) \leq t\}$$

sont donc faiblement fermés. On peut en déduire que

$$\{\mathbf{y}_k \in \mathcal{H}^k : g_x(\mathbf{y}_k) \leq t\} = \bigcup_{i=1}^{k} \{\mathbf{y}_k \in \mathcal{H}^k : g_{i,x}(\mathbf{y}_k) \leq t\}$$

est faiblement fermé, d'où le résultat en utilisant (i) dans la Définition 1.2.4. □

Corollaire 1.2.2 *La fonction $D(\mu, .)$ est faiblement s.c.i. sur $\mathcal{H}^k$.*

Preuve du Corollaire 1.2.2 Pour chaque $\mathbf{y}_k^* \in \mathcal{H}^k$, nous pouvons écrire :

$$\begin{aligned}
\liminf_{\mathbf{y}_k \xrightarrow{w} \mathbf{y}_k^*} D(\mu, \mathbf{y}_k) &= \liminf_{\mathbf{y}_k \xrightarrow{w} \mathbf{y}_k^*} \int_{\mathcal{H}} g_x(\mathbf{y}_k)\mu(dx) \\
&\geq \int_{\mathcal{H}} \liminf_{\mathbf{y}_k \xrightarrow{w} \mathbf{y}_k^*} g_x(\mathbf{y}_k)\mu(dx) \\
&\quad \text{(d'après le Lemme de Fatou)} \\
&\geq \int_{\mathcal{H}} g_x(\mathbf{y}_k^*)\mu(dx) \\
&\quad \text{(d'après le Lemme 1.2.4 et le } (ii) \text{ de la Définition 1.2.4)} \\
&= D(\mu, \mathbf{y}_k^*),
\end{aligned}$$

ce qui montre bien que $D(\mu, .)$ satisfait la condition (ii) de la Définition 1.2.4. □

Nous sommes maintenant en mesure de prouver le Théorème 1.2.1.

Preuve du Théorème 1.2.1 D'après le Lemme 1.2.3, il existe $R > 0$ tel que l'infimum de $D(\mu, .)$ sur $\mathcal{H}^k$ est aussi l'infimum de $D(\mu, .)$ sur B_R^k. Or, d'une part B_R^k est compact pour la topologie faible car $\mathcal{H}$ est réflexif (Dunford et Schwartz [12]), et d'autre part $D(\mu, .)$ est faiblement s.c.i. d'après le Corollaire 1.2.2. Donc, d'après la Propriété 1.2.1 (ii), la fonction $D(\mu, .)$ atteint son infimum sur B_R^k. □

1.3 Le problème statistique

Dans un contexte statistique, la loi de X est inconnue et nous disposons seulement à la place d'un n-échantillon $\{X_i\}_{i=1}^n$, où les X_i sont des variables aléatoires i.i.d. de même loi que X. Il est alors possible de définir μ_n la mesure empirique par

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \in A},$$

pour tout ensemble mesurable $A \subset \mathcal{H}$. Cette mesure empirique est une bonne approximation de μ quand n est grand, et nous allons nous en servir pour rendre utilisable en pratique les résultats de la Section 2.

1.3.1 Un premier estimateur convergent

Construction

Nous allons chercher à construire un estimateur

$$q_n = q_n(., X_1, \ldots, X_n)$$

dont la distorsion s'approche de la distorsion optimale D^*. L'idée consiste à de construire des quantificateurs empiriques dont les distorsions s'approchent asymptotiquement (avec n) de la distorsion optimale pour X. Dans la mesure où q_n est aléatoire, sa distorsion est donnée (avec un léger abus de notations) par

$$D(\mu, q_n) = \mathbb{E}\left[\|X - q_n(X)\| \mid X_1 \ldots, X_n\right].$$

Insistons bien sur le fait que $D(\mu, q_n)$ est une variable aléatoire.

Définition 1.3.1 *Un quantificateur empirique q_n^* est dit empiriquement optimal si*

$$q_n^* \in \arg\min_{q \in \mathcal{Q}_k} \sum_{i=1}^{n} \|X_i - q(X_i)\|, \tag{1.3}$$

autrement dit, si

$$D(\mu_n, q_n^*) = D_k^*(\mu_n).$$

Nous avons en fait simplement repris la définition d'un quantificateur optimal, en remplaçant μ par μ_n. Nous savons par conséquent grâce au Théorème 1.2.1 que, pour tout n, $D(\mu_n, .)$ admet (au moins) un minimum, ce qui veut dire qu'il existe toujours au moins un quantificateur empiriquement optimal q_n^*. Comme pour les quantificateurs optimaux théoriques, le Lemme 1.2.1 nous permet de ne minimiser $D(\mu_n, q_n)$ que dans la classe des quantificateurs de type plus proches voisins. Dans toute la suite, tous les quantificateurs empiriques considérés seront donc également de type plus proches voisins.

Convergence

Nous allons montrer que les quantificateurs empiriques optimaux obtenus à partir de la formule (1.3) sont consistants, au sens où la suite des distorsions $D(\mu, q_n^*)$ converge presque sûrement vers la distorsion optimale $D(\mu, q^*)$ lorsque la taille n de l'échantillon croît vers l'infini.

Théorème 1.3.1 *Pour tout $k \geq 1$, une suite de k-quantificateurs empiriques optimaux $(q_n^*)_{n\geq 1}$ satisfait*

$$\lim_{n\to\infty} D(\mu, q_n^*) = D_k^*(\mu) \qquad p.s.$$

et

$$\lim_{n\to\infty} D(\mu_n, q_n^*) = D_k^*(\mu) \qquad p.s.$$

L'idée est la suivante : si pour de grandes valeurs de n la mesure empirique μ_n estime bien la mesure μ, alors les quantificateurs optimaux pour μ_n devraient constituer de bonnes estimations des quantificateurs optimaux pour μ. Pour formaliser cette approche nous avons besoin d'une mesure appropriée de la proximité entre deux mesures de probabilité. À cet effet, étant données μ et ν deux mesures de probabilité sur $\mathcal{H}$ de moments d'ordre 1 fini, nous définissons la distance de L_1-Wasserstein entre μ et ν (cf. Dudley [11], Section 11.8) par

$$\rho(\mu, \nu) = \inf_{X\sim\mu, Y\sim\nu} \mathbb{E}\|X - Y\|,$$

où l'infimum est pris sur toutes les paires de variables aléatoires (X, Y) à valeurs dans $\mathcal{P}(\mu, \nu)$, où $\mathcal{P}(\mu, \nu)$ est l'espace des couples aléatoires (X, Y)

tels que X soit de loi μ ($X \sim \mu$) et Y de loi ν ($Y \sim \nu$). Nous pouvons remarquer immédiatement que, pour un couple $(X, Y) \in \mathcal{P}(\mu, \nu)$, nous avons

$$\rho(\mu, \nu) \leq \mathbb{E}\,\|X\| + \mathbb{E}\,\|Y\| < \infty.$$

Lemme 1.3.1 *L'infimum définissant $\rho(\mu, \nu)$ est en fait un minimum. De plus, ρ est une distance sur l'espace des lois de probabilité sur $\mathcal{H}$ de moment d'ordre 1 fini.*

Preuve du Lemme 1.3.1 Voir Dudley [11], pages 329-333.

Le lemme suivant justifie le choix de ρ en montrant que deux quantificateurs fondés sur des mesures proches au sens de ρ sont proches au sens de leur distorsion.

Lemme 1.3.2 *On a,*

$$\sup_{q \in \mathcal{Q}_k} |D(\mu, q) - D(\nu, q)| \leq \rho(\mu, \nu).$$

De plus,

$$|D_k^*(\mu) - D_k^*(\nu)| \leq \rho(\mu, \nu).$$

Preuve du Lemme 1.3.2 Soient $\mathcal{C} = \{y_i\}_{i=1}^k$ l'alphabet de q, $X \sim \mu$, et $Y \sim \nu$ atteignant le minimum définissant $\rho(\mu, \nu)$. Nous avons alors

$$\begin{aligned} D(\mu, q) &= \mathbb{E} \min_{i=1,\ldots,k} \|X - y_i\| \\ &\leq \mathbb{E}\|X - Y\| + \mathbb{E} \min_{i=1,\ldots,k} \|Y - y_i\| \\ &= \rho(\mu, \nu) + D(\nu, q). \end{aligned}$$

L'inégalité $D(\nu, q) - D(\mu, q) \leq \rho(\mu, \nu)$ se démontre de manière identique.

Prouvons maintenant le second point. Soit q^* un k-quantificateur optimal pour ν. Nous avons, par la première inégalité du lemme,

$$\begin{aligned} D_k^*(\mu) - D_k^*(\nu) &= D_k^*(\mu) - D(\nu, q^*) \\ &\leq D(\mu, q^*) - D(\nu, q^*) \\ &\leq \rho(\mu, \nu). \end{aligned}$$

L'inégalité $D_k^*(\nu) - D_k^*(\mu) \leq \rho(\mu, \nu)$ se démontre de manière similaire en considérant un k-quantificateur optimal pour μ. $\square$

Ce lemme fournit donc une majoration de l'écart entre la distorsion du quantificateur empirique optimal et la distorsion optimale.

Corollaire 1.3.1 *Tout k-quantificateur q_n^* empiriquement optimal vérifie*

$$D(\mu, q_n^*) - D_k^*(\mu) \leq 2\rho(\mu, \mu_n).$$

Preuve du Corollaire 1.3.1 Soit q^* un k-quantificateur optimal pour μ. On a alors

$$\begin{aligned} D(\mu, q_n^*) - D_k^*(\mu) &= D(\mu, q_n^*) - D(\mu, q^*) \\ &= D(\mu, q_n^*) - D(\mu_n, q_n^*) + D(\mu_n, q_n^*) - D(\mu, q^*) \\ &\leq D(\mu, q_n^*) - D(\mu_n, q_n^*) + D(\mu_n, q^*) - D(\mu, q^*) \\ &\leq 2\rho(\mu, \mu_n), \end{aligned}$$

la dernière inégalité découlant du Lemme 1.3.2. $\square$

Il nous reste un dernier résultat à prouver avant de pouvoir démontrer le Théorème 1.3.1. Nous utiliserons la notation $\Rightarrow$ pour désigner la convergence étroite vers μ (voir Billingsley [4]).

Lemme 1.3.3 *Étant données une probabilité ν et une suite de mesures de probabilité $(\nu_n)_{n\in\mathbb{N}^*}$, ν et ν_n de moment d'ordre 1 fini pour tout n, on a*

$$\left(\lim_{n\to\infty} \rho(\nu, \nu_n) = 0\right) \Leftrightarrow \left(\nu_n \Rightarrow \nu \text{ et } \int \|x\| \nu_n(dx) \to \int \|x\| \nu(dx)\right).$$

Preuve du Lemme 1.3.3 Ce résultat, classique en dimension finie (voir Graf et Luschgy [16], page 57), s'étend facilement au cas où $\mathcal{H}$ est un espace de Banach réflexif séparable. En voici la preuve.

- Supposons $\lim_{n\to\infty} \rho(\nu, \nu_n) = 0$.

Pour tout n, soit (Y, Y_n) un couple aléatoire dans $\mathcal{P}(\nu, \nu_n)$ tel que $\rho(\nu, \nu_n) = \mathbb{E}\|Y - Y_n\|$. On a alors $\lim_{n\to\infty} \mathbb{E}\|Y - Y_n\| = 0$, ce qui implique

$$\nu_n \Rightarrow \nu \text{ et } \int \|x\|\nu_n(dx) \to \int \|x\|\nu(dx).$$

- Réciproquement,supposons $\nu_n \Rightarrow \nu$.

Alors le Théorème de Skorohod (voir Dudley [11], Théorème 11.7.2) affirme qu'il existe, pour tout n, une paire $(Y, Y_n) \in \mathcal{P}(\nu, \nu_n)$ telle que $Y_n \to Y$ $p.s.$ Comme

$$2\|Y_n\| + 2\|Y\| - \|Y - Y_n\| \geq \|Y_n\| + \|Y\| \geq 0,$$

le Lemme de Fatou donne :

$$\begin{aligned} &\liminf_{n\to\infty} \mathbb{E}\left\{2\|Y_n\| + 2\|Y\| - \|Y - Y_n\|\right\} \\ &\quad \geq \mathbb{E}\left\{\liminf_{n\to\infty}(2\|Y_n\| + 2\|Y\| - \|Y - Y_n\|)\right\} \\ &\quad = 4\mathbb{E}\|Y\|. \end{aligned}$$

Supposons de plus que $\mathbb{E}\|Y_n\| \to \mathbb{E}\|Y\|$. Nous obtenons que $\mathbb{E}\|Y_n - Y\| \to 0$, ce qui implique $\rho(\nu, \nu_n) \to 0$. □

Preuve du Théorème 1.3.1 Nous savons que $D(\mu, q_n^*) \geq D_k^*(\mu)$ pour tout n. Par le Lemme 1.3.2 et le Corollaire 1.3.1, il suffit de prouver que

$$\lim_{n\to\infty} \rho(\mu, \mu_n) = 0 \; p.s. \tag{1.4}$$

Nous savons par le Lemme 1.3.3 que (1.4) est vérifiée si, presque sûrement, $\mu_n \Rightarrow \mu$ et $\int \|x\|\mu_n(dx) \to \int \|x\|\mu(dx)$, ce qui est obtenu en utilisant le Théorème de Varadarajan (voir Dudley [11], Théorème 11.4.1) et la loi forte des grands nombres. □

Vitesse

La plupart des résultats présents dans la littérature concernent le cas où $\mathcal{H} = \mathbb{R}^d$ et où la distorsion est fondée sur une distance L_2 (Pollard [29],

Linder [25], Linder, Lugosi, et Zeger [26]). Par exemple, il est montré dans [25] que s'il existe $T > 0$ tel que $\mathbb{P}[\|X\| \leq T] = 1$, alors

$$\mathbb{E}\, D(\mu, q_n^*) - D^*(\mu) \leq CT^2 \sqrt{\frac{k(d+1)\ln(k(d+1))}{n}},$$

où $C > 0$ est une constante universelle.

Plus récemment, Biau, Devroye et Lugosi [3] ont montré que si $\mathcal{H}$ est un espace de Hilbert et si la distorsion est donnée par une distance L_2, alors

$$\mathbb{E}\, D(\mu, q_n^*) - D^*(\mu) \leq C\frac{k}{\sqrt{n}},$$

où $C > 0$ est une constante universelle.

Dans ce qui suit, notre but est d'établir une vitesse de convergence dans un espace de Banach pour une distorsion donnée par une distante L_1, suivant une méthodologie utilisée par Bolley, Guillin, et Villani [6] et Malrieu [27]. Nous avons besoin pour cela d'introduire les notions suivantes :

Définition 1.3.2 *Soient (E, d) un espace métrique et $p \in [1; \infty[$*

1. *La distance de L_p-Wasserstein ρ_p est définie par :*

$$\rho_p(\phi, \xi) = \inf_{X \sim \phi, Y \sim \xi} (\mathbb{E}\, d(X, Y)^p)^{\frac{1}{p}},$$

   où ϕ et ξ sont deux probabilités sur E.

2. *La probabilité ϕ sur E satisfait une inégalité de transport $T_p(\lambda)$ si il existe $\lambda > 0$ tel que, pour toute probabilité ξ sur E,*

$$\rho_p^p(\phi, \xi) \leq \sqrt{\frac{2}{\lambda} H(\xi|\phi)},$$

   où $H(\xi|\phi) = \displaystyle\int \frac{d\xi}{d\phi} \log\left(\frac{d\xi}{d\phi}\right) d\phi$ est l'information de Kullback entre ξ et ϕ (voir Kullback et Leibler [23]).

Remarques :

- La distance de L_p-Wasserstein, également nommée distance de L_p-Kantorovich, est connue pour être appropriée dans le contexte de la quantification (Graf et Luschgy, Section 3 [16]) ;
- Pour ce choix de distance, et dans le but d'obtenir des vitesses de convergence, les inégalités de transport (ou inégalités de Talagrand) sont des outils adaptés (Ledoux [24]).

En pratique, il est difficile de déterminer directement si une mesure μ vérifie une inégalité de transport $T_p(\lambda)$. Dans le cas $p = 1$ les choses sont plus simples, comme le montre le Théorème ci-dessous, dont une preuve est calquée sur celle du Théorème 1.1 de Bolley, Guillin, et Villani [6], consacré au cas où $\mathcal{H} = \mathbb{R}^d$.

Théorème 1.3.2 *Une probabilité ϕ sur $\mathcal{H}$ satisfait une inégalité de transport $T_1(\lambda)$ si et seulement si pour tout $\alpha < \lambda/2$*

$$\int e^{\alpha\|x-y\|^2} d\mu(x) < \infty$$

pour un certain (et donc pour tout) y dans $\mathcal{H}$.

Dans la suite, nous ne considérons que le cas $p = 1$, et on note $\rho = \rho_1$. Par ailleurs, pour tout ensemble $\Lambda \subset \mathcal{H}$, on définit $\mathcal{P}(\Lambda)$ l'ensemble des probabilités sur Λ, muni de la métrique induite par ρ. On note également $\mathcal{N}(r, \Lambda)$ le plus petit nombre de boules de rayon r nécessaire pour recouvrir $\mathcal{P}(\Lambda)$, c'est-à-dire

$$\begin{aligned}&\mathcal{N}(r, \Lambda)\\ &\quad = \inf\left\{ n \in \mathbb{N} \text{ t.q. } \exists\, x_1, \ldots, x_n \in \mathcal{P}(\Lambda) : \textstyle\bigcup_{i=1}^n B_{\mathcal{P}(\Lambda)}(x_i, r) \supset \mathcal{P}(\Lambda) \right\},\end{aligned}$$

où $B_{\mathcal{P}(\Lambda)}(x_i, r)$ est la boule de $\mathcal{P}(\Lambda)$ centrée en x_i et de rayon r. La quantité $\ln(\mathcal{N}(r, \Lambda))$ est appelée l'entropie de $\mathcal{P}(\Lambda)$ (Van der Vaart et Wellner [32]).

De la même manière, soit $N(r, \Lambda)$ le plus petit nombre de boules de rayon r (par rapport à la métrique sur $\mathcal{H}$) nécessaire pour recouvrir Λ.

Dans le but d'établir une vitesse de convergence pour $D(\mu, q_n^*)$, on introduit les hypothèses suivantes :

H1 : Il existe $\lambda > 0$ tel que μ satisfait une inégalité $T_1(\lambda)$;

H2 : Toute boule fermée bornée B de $\mathcal{H}$ est précompacte. Autrement dit, pour tout $r > 0$, $N(r, B)$ est fini ;

Notons que **H1** est vérifiée pour les solutions des équations différentielles du type

$$dX_t = b(X_t)dt + s(X_t)dW_t,$$

où $t \in [0, T]$, $T < \infty$, et $b(.)$, $s(.)$ satisfont certaines hypothèses de régularité (Djellout, Guillin et Wu [9], Corollaire 4.1). **H2** est vérifiée, par exemple, si $\mathcal{H}$ est un espace de Sobolev sur un domaine compact de $\mathbb{R}^d$ (Cucker et Smale [8], exemple 3).

Pour $R > 0$, on rappelle que B_R représente la boule de $\mathcal{H}$ centrée sur l'origine et de rayon R. On note par ailleurs sans ambiguïté $\mathcal{N}(r, R) = \mathcal{N}(r, B_R)$ et $N(r, R) = N(r, B_R)$. En utilisant l'hypothèse **H2** et le Théorème A.1 de Bolley, Guillin, et Villani [6] on déduit qu'il existe une constante C strictement positive pour laquelle

$$\mathcal{N}(r, R) \leq \left(\frac{CR}{r}\right)^{N(r/2,R)}, \forall r, R > 0. \tag{1.5}$$

Théorème 1.3.3 *Supposons que $\mathcal{H}$ est un espace de Banach réflexif et séparable, et que* **H1** *et* **H2** *sont vérifiées. Pour tout $\lambda' < \lambda$ et $\varepsilon > 0$, il existe trois constantes K, γ, et R_1 strictement positives telles que pour $R = R_1 \max\left(1, \varepsilon^2, \ln\left(1/\varepsilon^2\right)\right)^{1/2}$ et $n \geq K \ln\left(\mathcal{N}(\gamma\varepsilon, R)\right)/\varepsilon^2$ on a,*

$$\mathbb{P}\left[\rho(\mu, \mu_n) \geq \varepsilon\right] \leq e^{-(\lambda'/2)n\varepsilon^2}.$$

En utilisant le Corollaire 1.3.1 on obtient immédiatement :

Corollaire 1.3.2 *Supposons que $\mathcal{H}$ est un espace de Banach réflexif et séparable, et que* **H1** *et* **H2** *sont vérifiées. Pour tout $\lambda' < \lambda$ et $\varepsilon > 0$, il existe trois constantes K, γ, et R_1 strictement positives telles que pour $R = R_1 \max\left(1, \varepsilon^2, \ln\left(1/\varepsilon^2\right)\right)^{1/2}$ et $n \geq K \ln\left(\mathcal{N}(\gamma\varepsilon, R)\right)/\varepsilon^2$ on a,*

$$\mathbb{P}\left[D(\mu, q_n^*) - D(\mu, q^*) \geq \varepsilon\right] \leq e^{-(\lambda'/8)n\varepsilon^2}.$$

Soit $\mathcal{R}$ la fonction de $\mathbb{R}_+^*$ dans $\mathbb{R}_+^*$ définie par $\mathcal{R}(x) = R_1 \max\left(1, x^2, \ln\left(1/\varepsilon^2\right)\right)^{1/2}$. On note $\mathcal{M}$ la fonction de $\mathbb{R}_+^*$ dans $\mathbb{R}_+^*$ définie par

$$\mathcal{M}(x) = K \ln\left(\mathcal{N}(\gamma x, \mathcal{R}(x))\right)/x^2. \tag{1.6}$$

Le Théorème 1.3.4 ci-dessous donne la vitesse de convergence désirée.

Théorème 1.3.4 *Supposons que $\mathcal{H}$ est un espace de Banach réflexif et séparable, et que* **H1** *et* **H2** *sont vérifiées. Si il existe une constante b strictement positive telle que $\mathcal{M}$ est inversible sur $]0, b]$ on a,*

$$\mathbb{E}\left(D\left(\mu, q_n^*\right) - D\left(\mu, q^*\right)\right) \leq C_0 \max(\mathcal{M}^{-1}(n), n^{-1/2}),$$

où $C_0 > 0$ est une constante.

Preuve du Théorème 1.3.4 Soit $\varepsilon > 0$ suffisamment petit. On a

$$\mathbb{P}\left[D(\mu, q_n^*) - D(\mu, q^*) > \varepsilon\right] \leq e^{-(\lambda'/8)n\varepsilon^2},$$

dès que $n \geq \mathcal{M}(\varepsilon)$, c'est-à-dire dès que $\varepsilon \geq \mathcal{M}^{-1}(n)$. On peut alors écrire :

$$\begin{aligned}
\mathbb{E}(D(\mu, q_n^*) - D(\mu, q^*)) &= \int_0^{+\infty} \mathbb{P}\left[D(\mu, q_n^*) - D(\mu, q^*) > \varepsilon\right] d\varepsilon \\
&= \int_0^{\mathcal{M}^{-1}(n)} \mathbb{P}\left[D(\mu, q_n^*) - D(\mu, q^*) > \varepsilon\right] d\varepsilon \\
&\quad + \int_{\mathcal{M}^{-1}(n)}^{+\infty} \mathbb{P}\left[D(\mu, q_n^*) - D(\mu, q^*) > \varepsilon\right] d\varepsilon \\
&\leq \mathcal{M}^{-1}(n) + \int_0^{+\infty} e^{-(\lambda'/8)n\varepsilon^2} d\varepsilon \\
&\leq C_0 \max(\mathcal{M}^{-1}(n), n^{-1/2}),
\end{aligned}$$

d'où le théorème. □

Notons que nous n'imposons aucune restriction sur le support de μ. En particulier nous n'avons pas besoin que μ soit à support borné. c'est un point intéressant car une telle hypothèse n'est pas vérifiée, par exemple, par les distributions des processus de diffusion classiques, pourtant souvent utilisés en modélisation stochastique.

Avant de démontrer le Théorème 1.3.3, nous allons donner les vitesses de convergences obtenues pour deux espaces fonctionnels différents.

Exemples

E1 : Espaces de Sobolev

Soit $\mathcal{X}$ un domaine compact de $\mathbb{R}^d$ ayant une frontière lisse. On considère l'espace $C^\infty(\mathcal{X})$ des fonctions infiniment différentiables sur $\mathcal{X}$. Pour tout s dans $\mathbb{N}$ on peut définir un produit scalaire dans $C^\infty(\mathcal{X})$ par :

$$\langle f|\ g\rangle_s = \int_{\mathcal{X}} \sum_{|\alpha|\leq s} D^\alpha f(x) D^\alpha g(x) dx\ , \tag{1.7}$$

où, pour $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, $\mathrm{D}^\alpha f = \dfrac{\partial^\alpha f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$, et $|\alpha| = \alpha_1 + \ldots + \alpha_d$. On note $\|\|\|_s$ la norme induite par le produit scalaire (1.7). L'espace de Sobolev $H^s(\mathcal{X})$ est le complété de $C^\infty(\mathcal{X})$ par rapport à la norme $\|\|\|_s$.

Le Théorème de plongement de Sobolev (voir Adams [1]) affirme que, pour $s > d/2$, l'inclusion canonique

$$J_s : H^s(\mathcal{X}) \hookrightarrow C(\mathcal{X})$$

est bien définie et est bornée, où $C(\mathcal{X})$ désigne l'espace des fonctions réelles continues définies sur $\mathcal{X}$. Par ailleurs le Théorème de Rellich (voir Adams

[1]) nous dit de plus que l'image par J_s de tout ensemble borné de $H^s(\mathcal{X})$ est compacte.
En combinant l'entropie métrique des boules B_R de $J_s(H^s(\mathcal{X}))$ (voir Cucker et Smale [8]) avec le Corollaire 1.3.2, obtient finalement que

$$\mathbb{P}\left[D(\mu, q_n^*) - D(\mu, q^*) > \varepsilon\right] \leq e^{-(\lambda'/8)n\varepsilon^2},$$

dès lors que

$$n \geq \ \ln\left(\ln\left(\frac{C_1}{\varepsilon}\right)\right)\frac{C_2}{\varepsilon^2}\exp\left(\left(\frac{C_3}{\varepsilon}\right)^{d/s}\right),$$

où C_1, C_2 et C_3 sont des constantes strictement positives, en particulier dès lors que

$$n \geq \exp\left(\left(\frac{C}{\varepsilon}\right)^{d/s}\right),$$

où C est une constante strictement positive. Tous calculs faits, on obtient avec le Théorème 1.3.4 l'existence d'une constante C_0 strictement positive telle que

$$\mathbb{E}(D(\mu, q_n^*) - D(\mu, q^*)) \leq \frac{C_0}{(\ln n)^{s/d}}.$$

E2 : **Espaces de Hilbert à noyaux reproduisants**

Pour plus de précisions sur les RKHS [1] nous renvoyons le lecteur au livre de Berlinet et Thomas-Agnan [2] et à l'article de Cucker et Smale [8]. On reprend les notations de l'exemple précédent.

Soient K un noyau de Mercer, *i.e.* une fonction de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{R}$ continue, symétrique et définie positive, ν une mesure sur $\mathcal{X}$ et $L_K : \mathcal{L}^2_\nu(\mathcal{X}) \longrightarrow C(\mathcal{X})$ l'opérateur linéaire défini par

$$(L_K(f))(x) = \int K(x,t)f(t)d\nu(t)\ . \tag{1.8}$$

Alors L_K est bien défini, positif et compact. On sait qu'il existe un unique espace de Hilbert $\mathcal{H}_\mathcal{K}$ de fonctions continues sur $\mathcal{X}$ et indépendant de ν tel

1. On utilise ici l'acronyme anglais RKHS pour Reproducing Hilbert Kernel Spaces.

que l'opérateur linéaire $L_K^{1/2}$ soit un isomorphisme entre $\mathcal{L}^2_\nu(\mathcal{X})$ et $\mathcal{H}_\mathcal{K}$, où $L_K^{1/2}$ et tel que $L_K^{1/2} \text{o} L_K^{1/2} = L_K$. Ainsi on a le diagramme suivant :

$$\begin{array}{ccc} \mathcal{L}^2_\nu(\mathcal{X}) & \xrightarrow{L_{K,C}^{1/2}} & C(\mathcal{X}) \\ & \searrow^{L_K^{1/2}} \quad \approx & \uparrow I_K \\ & & \mathcal{H}_K, \end{array}$$

où nous avons écrit $L_{K,C}^{1/2}$ pour souligner le fait que la cible est $C(\mathcal{X})$, et où I_K dénote l'inclusion triviale. Si K est de classe C^∞, l'image par I_K d'une boule fermée bornée de $\mathcal{H}_\mathcal{K}$ est compacte. Dans ce cas, en utilisant l'entropie métrique de $I_K(\mathcal{H}_K)$ (voir Cucker et Smale [8]) et le Corollaire 1.3.2, on obtient que pour tout $h > 2d$, il existe une constante C_h telle que

$$\mathbb{P}\left[D(\mu, q_n^*) - D(\mu, q^*) > \varepsilon\right] \leq e^{-(\lambda'/8)n\varepsilon^2},$$

dès lors que

$$n \geq \ \ln\left(\ln\left(\frac{C_1}{\varepsilon}\right)\right) \frac{C_2}{\varepsilon^2} \exp\left(\left(\frac{C_h}{\varepsilon}\right)^{2d/h}\right),$$

où C_1, C_2 et C_h sont des constantes strictement positives. De la même façon que dans l'exemple précédent, on obtient l'existence d'une constante C_0 strictement positive telle que

$$\mathbb{E}(D(\mu, q_n^*) - D(\mu, q^*)) \leq \frac{C_0}{(\ln n)^{2d/h}}.$$

Preuve du Théorème 1.3.3

Dans toute ce paragraphe, on suppose que les hypothèses **H1** et **H2** sont vérifiées. La démonstration s'effectue en trois étapes :

1. On commence par ramener les mesures μ et μ_n à des versions tronquées sur une boule B_R ;
2. On recouvre ensuite l'espace $\mathcal{P}(B_R)$ des mesures de probabilité sur B_R par des petites boules de rayon r ;

3. Enfin, on optimise les différents paramètres introduits dans le raisonnement.

Les trois lemmes suivants correspondent à ces trois étapes.

Soit $R > 0$. Considérons la mesure tronquée μ_R de support B_R définie, pour tout borélien A de $\mathcal{H}$ par :

$$\mu_R[A] = \frac{\mu[A \cap B_R]}{\mu[B_R]} = \mu(A|B_R).$$

On considère à présent les variables indépendantes $\{X_i\}_{i=1}^n$ de loi μ et $\{Y_i\}_{i=1}^n$ de loi μ_R. Posons ensuite, pour $i \leq n$:

$$X_i^R = \begin{cases} X_i \text{ si } \|X_i\| \leq R \\ Y_i \text{ si } \|X_i\| > R. \end{cases}$$

Soit δ_x la mesure de Dirac au point x. Les mesures empiriques μ_n et μ_n^R sont définies par

$$\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i} \text{ et } \mu_n^R = \frac{1}{n}\sum_{i=1}^n \delta_{X_i^R}.$$

On note de plus $E_\alpha = \int \exp(\alpha\|x\|^2)\mu(dx)$. Comme μ est supposée satisfaire une inégalité $T_1(\lambda)$, on a, pour $\alpha < \lambda/2$, $E_\alpha < \infty$.

Lemme 1.3.4 *Soient $\eta \in]0,1[$, $\varepsilon, \theta > 0$, $\alpha_1 \in]0, \lambda/2[$, et $\alpha \in]\alpha_1, \lambda/2[$. Alors on a pour tout $R > \max\left(\sqrt{1/2\alpha}, 2\theta/\alpha_1\right)$,*

$$\begin{aligned} \mathbb{P}\left[\rho\left(\mu, \mu_n\right) > \varepsilon\right] \;\leq\; & \mathbb{P}\left[\rho\left(\mu^R, \mu_n^R\right) > \eta\varepsilon - 2E_\alpha R e^{-\alpha R^2}\right] \\ & + \exp\left(-n\left[\theta\left(1-\eta\right)\varepsilon - E_\alpha e^{(\alpha_1-\alpha)R^2}\right]\right). \end{aligned}$$

Preuve du Lemme 1.3.4

Soit $\varepsilon > 0$ fixé, nous allons majorer la quantité $\mathbb{P}[\rho(\mu, \mu_n) > \varepsilon]$ en fonction de μ_R et de la mesure empirique associée. Supposons que μ satisfait une inégalité $T_1(\lambda)$. En suivant les arguments de la preuve du Théorème 1.1 dans [6] on établit que, pour tout $\alpha < \lambda/2$ et $R \geq \sqrt{1/2\alpha}$:

$$\rho(\mu, \mu_R) \leq 2E_\alpha R e^{-\alpha R^2}. \tag{1.9}$$

De plus, les mesures μ_n et μ_n^R vérifient

$$\rho(\mu_n, \mu_n^R) \leq \frac{1}{n}\sum_{i=1}^{n} \|X_i^R - X_i\| \leq \frac{1}{n}\sum_{i=1}^{n} Z_i,$$

où $Z_i = 2\|X_i\|\mathbf{1}_{\|X_i\|>R}$ $(i = 1, \ldots, n)$. On en déduit, comme dans la preuve du Théorème 1.1 dans [6], que si ε, θ sont strictement positifs et $\alpha < \lambda/2$ on a,

$$\mathbb{P}\left[\rho(\mu_n, \mu_n^R) > \varepsilon\right] \leq \exp\left(-n\left[\theta\varepsilon - E_\alpha e^{(\alpha_1-\alpha)R^2}\right]\right). \tag{1.10}$$

Le Lemme 1.3.4 est une conséquence immédiate des équations (1.9), (1.10), et de l'inégalité triangulaire pour ρ. $\square$

Lemme 1.3.5 *Étant donnés $\theta, \alpha, \alpha_1, \lambda_1 > 0$, tels que $\lambda_1 < \lambda$, $\alpha \in]\alpha_1, \lambda/2[$, et $\zeta > 1$, il existe des constantes strictement positives $\delta_1, \lambda_2 < \lambda_1$, K_1 et K_2 telles que pour tout $R > \zeta \max\left(\sqrt{1/2\alpha}, 2\theta/\alpha_1\right)$, et $\varepsilon > 0$,*

$$\begin{aligned}\mathbb{P}\left[\rho(\mu, \mu_n) > \varepsilon\right] \;\leq\; & \mathcal{N}(\delta_1\varepsilon/2, R)\exp\left(-n\left[\frac{\lambda_2}{2}\varepsilon^2 - K_1R^2e^{-\alpha R^2}\right]\right) \\ & + \exp\left(-n\left[K_2\zeta\varepsilon - K_3e^{(\alpha_1-\alpha)R^2}\right]\right),\end{aligned}$$

où K_3 est une constante strictement positive ne dépendant que de θ et α_1.

Preuve du Lemme 1.3.5 Nous allons commencer par montrer que μ^R satisfait une inégalité de transport $T_1(\lambda)$ modifiée. Soit Λ un borélien de $\mathcal{P}(B_R)$. En suivant les arguments de la preuve du Théorème 1.1 de [6] on obtient

$$\mathbb{P}[\mu_n^R \in \Lambda] \leq \exp\left(-n \inf_{\nu\in\Lambda} H(\nu|\mu^R)\right). \tag{1.11}$$

Soient à présent $\delta > 0$ et A un sous-ensemble mesurable de $\mathcal{P}(B_R)$. On note $\mathcal{N}^A$ le plus petit nombre de boules de rayon $\delta/2$ pour la distance ρ nécessaire pour recouvrir A. Les boules de ce recouvrement sont notées

B_i, $i = 1, \ldots, \mathcal{N}^A$. Chacune de ces boules est convexe, contenue dans le δ-voisinage A_δ de A pour la distance ρ (c'est-à-dire l'ensemble des éléments de $\mathcal{H}$ qui sont à une distance d'au plus δ de A). De plus, comme la norme de l'espace $\mathcal{H}$ rend les boules fermées bornées totalement bornées, les boules B_i sont totalement bornées (pour la distance ρ). On déduit facilement de l'équation (1.11) que :

$$\mathbb{P}[\mu_n^R \in A] \leq \mathcal{N}^A \exp\left(-n \inf_{\nu \in A_\delta} H(\nu|\mu^R)\right). \tag{1.12}$$

Posons maintenant

$$A = \left\{\nu \in \mathcal{P}(B_R) : \rho(\nu, \mu^R) \geq \eta\varepsilon - 2E_\alpha R e^{-\alpha R^2}\right\}.$$

D'après l'inégalité élémentaire

$$\forall a \in]0,1[, \exists C > 0 \text{ t.q. } \forall x, y \in \mathbb{R}, (x-y)^2 \geq (1-a)x^2 - Cy^2, \tag{1.13}$$

on a pour toute probabilité ν sur $\mathcal{H}$:

$$\forall \lambda_1 < \lambda, \exists K > 0 : H(\nu|\mu^R) \geq \frac{\lambda_1}{2}\rho^2(\mu^R, \nu)^2 - KR^2 e^{-\alpha R^2}.$$

On a donc

$$\forall \nu \in A_\delta, \quad H(\nu|\mu^R) \geq \frac{\lambda_1}{2}\rho^2(\mu^R, \nu)^2 - KR^2 e^{-\alpha R^2} \geq \frac{\lambda_1}{2}m^2 - KR^2 e^{-\alpha R^2},$$

où

$$m = \max\left(\eta\varepsilon - 2E_\alpha R e^{-\alpha R^2} - \delta, 0\right).$$

L'équation (1.12) nous permet alors de conclure que

$$\mathbb{P}\left[\rho(\mu^R, \mu_n^R) \geq \eta\varepsilon - 2E_\alpha R e^{-\alpha R^2}\right] \leq \mathcal{N}^{\mathcal{A}} \exp\left(-n\left[\frac{\lambda_1}{2}m^2 - KR^2 e^{-\alpha R^2}\right]\right). \tag{1.14}$$

D'après (1.13), il existe δ_1, η_1 et K_1 strictement positifs ne dépendant que de α, λ_1, et λ_2 tels que

$$\frac{\lambda_1}{2}m^2 - KR^2 e^{-\alpha R^2} \geq \frac{\lambda_2}{2}\varepsilon^2 - K_1 R^2 e^{-\alpha R^2},$$

où $\delta = \delta_1 \varepsilon$. Cela donne, avec (1.14),

$$\mathbb{P}\left[\rho(\mu^R, \mu_n^R) \geq \eta\varepsilon - 2E_\alpha R e^{-\alpha R^2}\right] \leq \mathcal{N}^{\mathcal{A}} \exp\left(-n\left[\frac{\lambda_2}{2}\varepsilon^2 - K_1 R^2 e^{-\alpha R^2}\right]\right) \tag{1.15}$$

Pour majorer $\mathcal{N}^A$, on remarque que, comme A est contenu dans $\mathcal{P}(\mathcal{B}_\mathcal{R})$, on a

$$\mathcal{N}^A \leq \mathcal{N}(\delta/2, R) = \mathcal{N}(\delta_1\varepsilon/2, R).$$

Le Lemme 1.3.5 s'obtient alors en combinant le Lemme 1.3.4 et (1.15). □

Le Lemme 1.3.6 ci-dessous simplifie la borne obtenue par les Lemmes 1.3.4 et 1.3.5 ci dessus.

Lemme 1.3.6 *Soient $\lambda' < \lambda$, $\alpha < \lambda/2$, et $\alpha' < \alpha$. Il existe $\delta_1 > 0$ tel que pour tout $\varepsilon > 0$:*

$$\mathbb{P}\left[\rho(\mu, \mu_n) > \varepsilon\right] \leq \exp\left(-\frac{\lambda'}{2} n\varepsilon^2\right) + \exp\left(-\alpha' n\varepsilon^2\right),$$

dès lors que

$$R^2 \geq R_2 \max\left(1, \varepsilon^2, \ln\left(\frac{1}{\varepsilon^2}\right)\right) \text{ et } n \geq K_4 \frac{\ln\left(\mathcal{N}(\delta_1\varepsilon/2, R)\right)}{\varepsilon^2}, \tag{1.16}$$

où R_2 et K_4 sont des constantes ne dépendant de μ qu'au travers de λ et α.

Preuve du Lemme 1.3.6 On a d'une part, sous les hypothèses et notations du Lemme 1.3.5, que pour tout $\lambda' < \lambda_2$,

$$\ln\left(\mathcal{N}(\delta_1\varepsilon/2, R) \exp\left(-n\left[\frac{\lambda_2}{2}\varepsilon^2 - K_1 R^2 e^{-\alpha R^2}\right]\right)\right) \leq \frac{-n\lambda'\varepsilon^2}{2} \tag{1.17}$$

dès lors que R, $R/\ln(1/\varepsilon^2)$ et $n\varepsilon^2/\ln(\mathcal{N}(\delta_1\varepsilon/2, R))$ sont suffisamment grands (il suffit pour obtenir cela de suivre les arguments de la preuve du Théorème 1.1 dans [6]).

Posons d'autre part $\alpha' < \alpha_2 < \alpha_1$. Nous pouvons choisir ζ tel que $K_2\zeta = \alpha_2\varepsilon$, et obtenir ainsi

$$\exp\left(-n\left[K_2\zeta\varepsilon - K_3 e^{(\alpha_1-\alpha)R^2}\right]\right) = \exp\left(-n\left[\alpha_2\varepsilon^2 - K_3 e^{(\alpha_1-\alpha)R^2}\right]\right),$$

qui peut être majoré par

$$\exp\left(-\alpha' n\varepsilon^2\right), \tag{1.18}$$

pour R et $R^2/\ln(1/\varepsilon^2)$ assez grands. Il suffit alors de combiner les équations (1.17) et (1.18) pour obtenir le lemme. $\square$

Pour prouver le Théorème 1.3.3, il suffit maintenant de remarquer que pour $K < \min((\lambda'/2), \alpha')$ et n assez grand, on a

$$\exp\left(-\frac{\lambda'}{2} n\varepsilon^2\right) + \exp\left(-\alpha' n\varepsilon^2\right) \leq \exp\left(-Kn\varepsilon^2\right).$$

Discussion

L'hypothèse **H2** se justifie du point de vue de la modélisation, comme le montrent les deux exemples étudiés précédemment. La méthode que nous avons utilisée, se basant sur la distance de Wasserstein, permet de ne pas faire intervenir la taille de l'alphabet choisi, contrairement au résultat de Linder [25], dans lequel la constante du Théorème 1.3.4 se comporte en $\sqrt{k\ln(k)}$. Précisons aussi que, contrairement au résultat de Linder, le Théorème 1.3.3 s'obtient sans imposer que μ est à support borné (une telle hypothèse n'est pas vérifiée par les lois des processus de diffusion classiques, pourtant abondamment utilisés en modélisation stochastique).

Le principal inconvénient de l'estimateur que nous venons d'étudier réside dans le fait que le calcul des quantificateurs empiriques optimaux se révèle être d'une trop grande complexité (Linder [25]).

L'algorithme de Lloyd sur les données

Nous avons présenté dans la Section 2 un algorithme afin de choisir un bon quantificateur de la mesure μ. Nous allons ici utiliser cet algorithme avec la mesure empirique μ_n.

Le principe de l'algorithme reste le même (Tableau 1.2), mais nous utilisons μ_n à la place de μ pour l'itération de Lloyd. Ce qui donne donc :

1. Étant donné un alphabet C_m, soit $\{S_i\}_{i=1}^k$ la partition de Voronoï correspondante.
2. On utilise la condition des centroïdes afin de construire $C_{m+1} = \{y_i^{(m)}\}_{i=1}^k$, l'alphabet optimal pour la partition calculée juste avant, c'est-à-dire,

$$y_i^{(m)} \in \arg\min_{y \in \mathcal{H}} \left\{ \sum_{\ell=1}^{n} \|x_\ell - y\| \mathbf{1}_{x_\ell \in S_i^{(m)}} \right\}, \ i = 1, \dots k.$$

Tableau 1: L'algorithme de Lloyd

Etape 1. Soit C_1 un alphabet initial et posons m = 1.

Etape 2. Étant donné l'alphabet C_m, on utilise l'itération de Lloyd pour construire l'alphabet C_{m+1},

Etape 3. On calcule D_{m+1} la distorsion de C_{m+1}. Si D_m- D_{m+1} est plus petit qu'un seuil fixé l'algorithme s'arrête. Sinon on pose m = m + 1 et on retourne a l'étape 2.

TABLE 1.2 – Algorithme de Lloyd.

Cependant cet algorithme n'est toujours pas satisfaisant car les médianes empiriques sont très difficiles à calculer en pratique. Afin de pallier cet inconvénient, une solution possible consiste à considérer des "medoïdes" à la place des centroïdes (voir Kaufman et Rousseuw [20]). Il s'agit en fait de centroïdes calculés parmi les éléments de l'échantillon. Plus précisément, l'étape 2 de l'itération de Lloyd est remplacée par

$$y_i^{(m)} \in \arg\min_{x_1,\dots,x_n} \left\{ \sum_{\ell=1}^{n} \|x_\ell - x_k\| \mathbf{1}_{x_\ell \in S_i^{(m)}} \right\}, \ i = 1, \dots k.$$

Dans un cas comme dans l'autre, le quantificateur sélectionné n'est pas optimal vis à vis de la distorsion.

Cette idée de minimiser sur les données est reprise et formalisée dans la section qui suit.

1.3.2 Une méthodologie différente : minimisation sur les données

Construction

Dans ce paragraphe, nous allons présenter un nouveau quantificateur empirique, qui est à la fois calculable et optimal vis à vis de la distorsion. L'idée est de minimiser la distorsion, en imposant que les centres choisis le soient dans l'échantillon. Replaçons nous dans le cadre général où $(\mathcal{H}, \|.\|)$ est un espace de Banach réflexif et séparable. Comme dans la Section 3, la notation $D(\mu, \mathbf{y}_k)$ est utilisée en lieu et place de la notation $D(\mu, q)$. Notons également

$$D(\mu_n, \mathbf{y}_k) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,k} \|X_i - y_j\|.$$

Nous pouvons alors définir notre estimateur $\mathbf{y}_{k,n}^*$ (qui est un k-uplet) par

$$\mathbf{y}_{k,n}^* \in \operatorname*{arg\,min}_{\mathbf{z} \in \{X_1,\dots,X_n\}^k} D(\mu_n, \mathbf{z}).$$

En d'autres termes nous cherchons dans l'échantillon un alphabet qui minimise la distorsion, c'est-à-dire que nous reprenons l'idée des medoïdes évoquée pour l'algorithme de Lloyd sur les données. On généralise ainsi la méthode proposée par Cadre [7], qui considère le cas $k = 1$. Notons $\|.\|_k$ la norme utilisée sur $\mathcal{H}^k$ (par exemple, pour tout élément $\mathbf{z} = \{z_1, \dots, z_k\} \in \mathcal{H}^k$, $\|\mathbf{z}\|_k = \max_{i=1,\dots,k} \|z_i\|$), et $B_{\mathcal{H}^k}(\mathbf{z}, r)$ la boule fermée de $\mathcal{H}^k$ centrée en $\mathbf{z}$ et de rayon r.

Convergence presque sûre

Théorème 1.3.5 *Supposons que pour un alphabet $\mathbf{y}_k^*$ optimal pour μ, on a*

$$\forall \varepsilon > 0,\ \mathbb{P}\left[(X_1, \dots, X_k) \in B_{\mathcal{H}^k}(\mathbf{y}_k^*, \varepsilon)\right] > 0. \tag{1.19}$$

Alors,

$$\lim_{n\to\infty} D(\mu, \mathbf{y}^*_{k,n}) = D^*_k(\mu) \ p.s.$$

Remarque : La condition du Théorème 1.3.5 ci-dessus requiert simplement que la probabilité d'avoir k données dans le voisinage de $\mathbf{y}^*_k$ soit non nulle. La nécessité de cette condition est évidente. En effet, supposons qu'il existe $\varepsilon > 0$ tel que pour tout alphabet optimal (pour μ)) $\mathbf{y}^*_k$, $\mathbb{P}[(X_1, \ldots, X_k) \notin B_{\mathcal{H}^k}(\mathbf{y}^*_k, \varepsilon)] = 1$. Alors, par construction, $D(\mu, \mathbf{y}^*_{k,n})$ ne peut converger vers $D^*_k(\mu)$.

Preuve du Théorème 1.3.5 D'après le Lemme 1.3.2 on a

$$D(\mu_n, \mathbf{y}^*_{k,n}) - D(\mu, \mathbf{y}^*_{k,n}) \leq \rho(\mu, \mu_n).$$

On a donc avec le Lemme 1.3.3 et le Théorème de Varadarajan [11] :

$$D(\mu_n, \mathbf{y}^*_{k,n}) - D(\mu, \mathbf{y}^*_{k,n}) \to 0 \text{ p.s. quand } n \to \infty. \tag{1.20}$$

Soient $p \leq n$ et $\mathbf{z} \in \{X_1, \ldots, X_p\}^k$. Comme $D(\mu_n, \mathbf{y}^*_{k,n}) \leq D(\mu_n, \mathbf{z})$ et $D(\mu_n, \mathbf{z}) \to D(\mu, \mathbf{z})$ p.s. d'après la loi des grands nombres, nous avons

$$\limsup_n D(\mu_n, \mathbf{y}^*_{k,n}) \leq D(\mu, \mathbf{z}) \text{ p.s.}$$

D'après (1.20), on obtient pour tout $p \geq 1$:

$$\limsup_n D(\mu, \mathbf{y}^*_{k,n}) \leq \min_{\mathbf{z}\in\{X_1,\ldots,X_p\}^k} D(\mu, \mathbf{z}). \tag{1.21}$$

Calculons maintenant la limite, lorsque $p \to \infty$, du terme de droite de l'équation (1.21). Posons, pour $\varepsilon > 0$ et $p \geq 1$,

$$\begin{aligned} N(p,\varepsilon) \ = \ & \Big[\exists \mathbf{z}^* \in \operatorname*{arg\,min}_{\mathbf{z}\in\{X_1,\ldots,X_p\}^k} D(\mu, \mathbf{z}) \cap B_{\mathcal{H}^k}(\mathbf{y}^*_k, \varepsilon), \\ & D(\mu, \mathbf{z}^*) \geq D(\mu, \mathbf{y}^*_k) + 2\varepsilon\Big]. \end{aligned}$$

Comme nous avons pour tous $\mathbf{y}_k, \mathbf{y}'_k \in \mathcal{H}^k$,

$$|D(\mu, \mathbf{y}_k) - D(\mu, \mathbf{y}'_k)| \leq \|\mathbf{y}_k - \mathbf{y}'_k\|_k,$$

on obtient

$$N(p,\varepsilon) \quad \subset [D(\mu,\mathbf{y}_k^*) \geq D(\mu,\mathbf{y}_k^*) + \varepsilon] = \emptyset.$$

Ainsi, dès lors que $p \geq k$,

$$\begin{aligned}
&\mathbb{P}\Big[\min_{\mathbf{z}\in\{X_1,\ldots,X_p\}^k} D(\mu,\mathbf{z}) - D(\mu,\mathbf{y}_k^*) > 2\varepsilon\Big] \\
&\leq \mathbb{P}\Big[N\left(p,\varepsilon\right)\Big] + \mathbb{P}\Big[\forall\, \mathbf{z} \in \{X_1,\ldots,X_p\}^k,\ \mathbf{z} \notin B_{\mathcal{H}^k}(\mathbf{y}_k^*,\varepsilon)\Big] \\
&\leq \mathbb{P}\Big[\left(X_1,\ldots,X_k\right) \notin B_{\mathcal{H}^k}\left(\mathbf{y}_k^*,\varepsilon\right)\Big]^{\lfloor p/k \rfloor} \\
&= \Big[1 - \mathbb{P}\big[(X_1,\ldots,X_k) \in B_{\mathcal{H}^k}(\mathbf{y}_k^*,\varepsilon)\big]\Big]^{\lfloor p/k \rfloor}
\end{aligned} \tag{1.22}$$

où $\lfloor . \rfloor$ désigne la fonction partie entière. En appliquant le Lemme de Borel-Cantelli, nous pouvons conclure que

$$\lim_{p\to\infty} \min_{\mathbf{z}\in\{X_1,\ldots,X_p\}^k} D(\mu,\mathbf{z}) = D(\mu,\mathbf{y}_k^*) \text{ p.s.},$$

ce qui, combiné avec (1.21), nous donne le Théorème. □

Convergence dans L_1

Théorème 1.3.6 *Sous les hypothèse* **H1** *et* **H2**, *on a :*

$$\lim_{n\to\infty} \mathbb{E}\, D(\mu,\mathbf{y}_{k,n}^*) = D_k^*(\mu)$$

Preuve du Théorème 1.3.6 : On peut écrire d'une part

$$\begin{aligned}
&D(\mu,\mathbf{y}_{k,n}^*) - D_k^*(\mu) \\
&= D(\mu,\mathbf{y}_{k,n}^*) - D(\mu_n,\mathbf{y}_{k,n}^*) + D(\mu_n,\mathbf{y}_{k,n}^*) - D_k^*(\mu) \\
&\leq |D(\mu,\mathbf{y}_{k,n}^*) - D(\mu_n,\mathbf{y}_{k,n}^*)| + |D(\mu_n,\mathbf{y}_{k,n}^*) - D_k^*(\mu)| \\
&\leq \rho(\mu,\mu_n) + |D(\mu_n,\mathbf{y}_{k,n}^*) - D_k^*(\mu)|.
\end{aligned}$$

d'autre part on a

$$\lim_{n\to\infty} D(\mu_n,\mathbf{y}_{k,n}^*) = D_k^*(\mu) \text{ p.s. et } \lim_{n\to\infty} \rho(\mu,\mu_n) = 0.$$

De plus,

$$\begin{aligned} D(\mu_n, \mathbf{y}^*_{k,n}) &= \min_{\mathbf{z}\in\{X_1,\ldots,X_n\}^k} \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\ldots,k} \|X_i - \mathbf{z}_j\| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \|X_i - X_1\| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \|X_i\| + \|X_1\|. \end{aligned}$$

Comme $\frac{1}{n}\sum_{i=1}^{n}\|X_i\|$ est équi-intégrable, $\frac{1}{n}\sum_{i=1}^{n}\|X_i\| + \|X_1\|$ l'est aussi. Le théorème de convergence dominée de Lebesgue nous permet donc d'écrire

$$\lim_{n\to\infty} \mathbb{E}|D(\mu, \mathbf{y}^*_{k,n}) - D^*_k(\mu)| = 0,$$

d'où le théorème. □

Vitesse de convergence

Le but de ce paragraphe est de montrer que $D(\mu, \mathbf{y}^*_{n,k})$ converge vers $D^*_k(\mu)$ à la même vitesse que $D(\mu, q^*_n)$. On rappelle que la fonction $\mathcal{M}$ est définie par la relation (1.6). Pour tout $\mathbf{y}_k \in \mathcal{H}^k$ et $\varepsilon > 0$, on note :

$$f(\mathbf{y}_k, \varepsilon) = \mathbb{P}\big[(X_1, \ldots, X_k) \in B_{\mathcal{H}^k}(\mathbf{y}_k, \varepsilon)\big].$$

On introduit par ailleurs les hypothèses suivantes :

H3 Il existe une fonction décroissante $V : \mathbb{N}^* \to \mathbb{R}^*_+$ et des constantes strictement positives u, v, C telles que

$$\max\left(\int_0^u (1 - f(\mathbf{y}^*_k, \varepsilon))^{\lfloor n/k \rfloor}\, d\varepsilon, \int_v^{+\infty} (1 - f(\mathbf{y}^*_k, \varepsilon))^{\lfloor n/k \rfloor}\, d\varepsilon\right) \leq V(n);$$

H4 Il existe $c_1 > 0$ tel que $f(\mathbf{y}_k, \varepsilon) \geq 1 - \exp(-\varepsilon^2)$ pour $\varepsilon \in]0, c_1]$;

H5 Il existe $c_2 > 0$ tel que $f(\mathbf{y}_k, \varepsilon) \geq 1 - \exp(-\varepsilon^2)$ pour $\varepsilon \in [c_2, +\infty[$;

Théorème 1.3.7 *On suppose que $\mathcal{H}$ est un espace de Banach réflexif et séparable, et que* **H1** *et* **H2** *sont satisfaites. Soit $\mathbf{y}^*_k$ un alphabet optimal pour*

μ vérifiant **H3**. *Si $\mathcal{M}$ est inversible sur un intervalle $]0, b]$, Il existe $C_0 > 0$ une constante telle que, pour n assez grand,*

$$\mathbb{E}\left(D\left(\mu, \mathbf{y}_{k,n}^*\right) - D_k^*(\mu)\right) \leq C_0 \max(\mathcal{M}^{-1}(n), V(n), \lfloor n/k \rfloor^{-1/2}).$$

Corollaire 1.3.3 *Sous les hypothèses du Théorème 1.3.7, si* **H4** *et* **H5** *sont satisfaites, alors*

$$\mathbb{E}\left(D\left(\mu, \mathbf{y}_{k,n}^*\right) - D_k^*(\mu)\right) \leq C_0 \max(\mathcal{M}^{-1}(n), \lfloor n/k \rfloor^{-1/2}).$$

Remarques :

- L'hypothèse **H3** impose que la probabilité que des données soient présentes dans un voisinage d'un alphabet optimal pour μ grandisse suffisamment avec n. C'est une hypothèse incontournable dans la preuve du Théorème 1.3.7.
- Les hypothèses **H4** et **H5** impliquent l'hypothèse **H3**;
- L'hypothèse **H5** est satisfaite si μ admet un support borné;
- Le Corollaire 1.3.3 montre en fait que $D\left(\mu, \mathbf{y}_{k,n}^*\right)$ converge vers $D_k^*(\mu)$ à la même vitesse que $D(\mu, q_n^*)$.

Preuve du Théorème 1.3.7 On a la décomposition

$$\begin{aligned} D(\mu, \mathbf{y}_{k,n}^*) - D_k^*(\mu) &= D(\mu, \mathbf{y}_{k,n}^*) - D(\mu_n, \mathbf{y}_{k,n}^*) \\ &+ D(\mu_n, \mathbf{y}_{k,n}^*) - \min_{\mathbf{z} \in \{X_1, \ldots, X_n\}^k} D(\mu, \mathbf{z}) \\ &+ \min_{\mathbf{z} \in \{X_1, \ldots, X_n\}^k} D(\mu, \mathbf{z}) - D_k^*(\mu). \end{aligned}$$

D'après le Lemme 1.3.2,

$$D(\mu, \mathbf{y}_{k,n}^*) - D(\mu_n, \mathbf{y}_{k,n}^*) \leq \rho(\mu, \mu_n)$$

et

$$D(\mu_n, \mathbf{y}_{k,n}^*) - \min_{\mathbf{z} \in \{X_1, \ldots, X_n\}^k} D(\mu, \mathbf{z}) \leq \rho(\mu, \mu_n).$$

Par suite,

$$D(\mu, \mathbf{y}_{k,n}^*) - D_k^*(\mu) \leq 2\rho(\mu, \mu_n) + \min_{\mathbf{z} \in \{X_1, \ldots, X_n\}^k} D(\mu, \mathbf{z}) - D_k^*(\mu). \qquad (1.23)$$

De plus, d'après l'inégalité (1.22), on a pour $n \geq k$:

$$\mathbb{P}\left[\min_{\mathbf{z}\in\{X_1,\ldots,X_n\}^k} D(\mu,\mathbf{z}) - D_k^*(\mu) \geq 2\varepsilon\right] \leq \Big[1 - f(\mathbf{y}_k^*,\varepsilon)\Big]^{\lfloor n/k\rfloor}.$$

On en déduit :

$$\begin{aligned}
&\mathbb{E}\min_{\mathbf{z}\in\{X_1,\ldots,X_n\}^k} D(\mu,\mathbf{z}) - D_k^*(\mu)\\
&= \int_0^{+\infty} \mathbb{P}\Big[\min_{\mathbf{z}\in\{X_1,\ldots,X_n\}^k} D(\mu,\mathbf{z}) - D_k^*(\mu) \geq \varepsilon\Big] d\varepsilon\\
&\leq 2\int_0^{+\infty} \Big(1 - f(\mathbf{y}_k^*,\varepsilon)\Big)^{\lfloor n/k\rfloor} d\varepsilon\\
&\leq 2\left(\int_{[0,u]\cup[v,\infty[} \Big(1 - f(\mathbf{y}_k^*,\varepsilon)\Big)^{\lfloor n/k\rfloor} d\varepsilon + \int_u^v \Big(1 - f(\mathbf{y}_k^*,\varepsilon)\Big)^{\lfloor n/k\rfloor} d\varepsilon\right)\\
&\leq 2\left(2V(n) + \int_u^v \Big[1 - f(\mathbf{y}_k^*,\varepsilon)\Big]^{\lfloor n/k\rfloor} d\varepsilon\right)\\
&\text{(d'après l'hypothèse } \mathbf{H3})\\
&\leq 2\left(2V(n) + (v-u)\Gamma^{\lfloor n/k\rfloor}\right)\\
&\leq C\max\left(\lfloor n/k\rfloor^{-1/2}, V(n)\right) \text{ pour } n \text{ assez grand,}
\end{aligned}$$

où $\Gamma < 1$ et C sont des constantes strictement positives. Il suffit alors d'utiliser l'inégalité (1.23), le Théorème 1.3.3 et le Théorème 1.3.4 pour conclure la preuve. □

Preuve du Corollaire 1.3.3 Il suffit de reprendre la preuve du Théorème 1.3.7 et de remarquer que si **H4** et **H5** sont vérifiés, alors

$$\begin{aligned}
&\mathbb{E}\min_{\mathbf{z}\in\{X_1,\ldots,X_n\}^k} D(\mu,\mathbf{z}) - D_k^*(\mu)\\
&\leq 2\left(\int_{[0,u]\cup[v,\infty[} \Big(\exp(-\varepsilon^2)\Big)^{\lfloor n/k\rfloor} d\varepsilon + (v-u)\Gamma^{\lfloor n/k\rfloor}\right)\\
&\leq \frac{C}{\sqrt{\lfloor n/k\rfloor}}. \qquad \square
\end{aligned}$$

Algorithmes

Afin de calculer en pratique l'alphabet $\mathbf{y}_{k,n}^*$, nous avons construit un algorithme simple, entièrement basé sur les données, que nous avons appelé

algorithme Alter. Son mode de fonctionnement est présenté dans le Tableau 1.3.

Algorithme Alter
Etape 1 On créé une bijection entre $N_k=\{1,2,...,n*(n-1)*...*(n-k+1)\}$ et tous les k-uples $y_k=\{x_1,...x_k\}$. **Etape 2** A tout élement n_k de N_k correspond un unique y_k, et on note D_k la distortion associée à y_k. **Etape 3** on initialise $D^*=D_1$, puis a chaque fois que $D_{k+1}<D_k$ on actualise $D^*=D_{k+1}$.

TABLE 1.3 – Modus operandi de l'algorithme Alter.

Cet algorithme permet d'éliminer les deux défauts de l'algorithme de Lloyd : Il ne dépend pas de conditions initiales et il converge vers la distorsion optimale. Cependant, le nombre de quantificateurs à tester grandit très vite avec n. En effet, pour une taille d'alphabet k, il faut calculer des distorsions pour $n \times (n-1) \times \ldots \times (n-k+1)$ quantificateurs. Cette complexité en $o(n^k)$ rend cet algorithme inutilisable pour de grandes valeurs de n et k.

Dans le but de résoudre ce problème de complexité, nous avons défini l'itération Alter-fast de la façon suivante :

1. Choisir au hasard $n_1 < n$ données dans le jeu de données complet (n_1 devant être petit) ;
2. Appliquer l'algorithme Alter sur ces n_1 data (Les distorsions empiriques étant calculées en utilisant le jeu de données complet) ;
3. Stoquer l'alphabet sélectionné.

Nous avons ensuite construit une version accélérée de l'algorithme Alter, que nous nommons algorithme Alter-fast :

1. Lancer n_2 fois l'itération Alter-fast (n_2 devant être grand) ;
2. Sélectionner, parmi les n_2 alphabets obtenus, celui qui minimise la distorsion empirique associée (calculée en utilisant le jeu de donnée tout entier).

De la même manière que l'algorithme de Lloyd sur les données pour l'algorithme de Lloyd, l'algorithme Alter-fast fournit une alternative utilisable en pratique de l'algorithme Alter. En effet, sa complexité en $o(n_2 \times n_1^k)$ est bien plus raisonnable. Nous verrons dans la Section 1.4.2 que cet algorithme fonctionne presque aussi bien que l'algorithme Alter sur un jeu de données réelles.

1.4 Applications

Dans cette dernière section nous proposons d'expérimenter les différentes méthodes sur plusieurs jeux de données (simulées ou réelles), afin de mettre en évidence la pertinence de nos choix théoriques.

1.4.1 Étude sur des données simulées

Un exemple simple

Examinons tout d'abord un exemple simple. On se place dans $\mathbb{R}^2$, où l'on construit deux groupes de données bien distincts. Dans la Figure 1.5, le graphique de gauche représente les deux groupes de données simulées (en rouge et bleu), et celui de droite montre la séparation obtenue par l'algorithme de Lloyd [1] (version medoïdes) pour la distance L_1. La classification obtenue est correcte : toutes les données sont bien réparties.

1. Nous avons reprogrammé cet algorithme, mais il en existe un prêt à l'emploi dans R qui s'intitule PAM (Partitioning Around Medoids) dans le package Cluster.

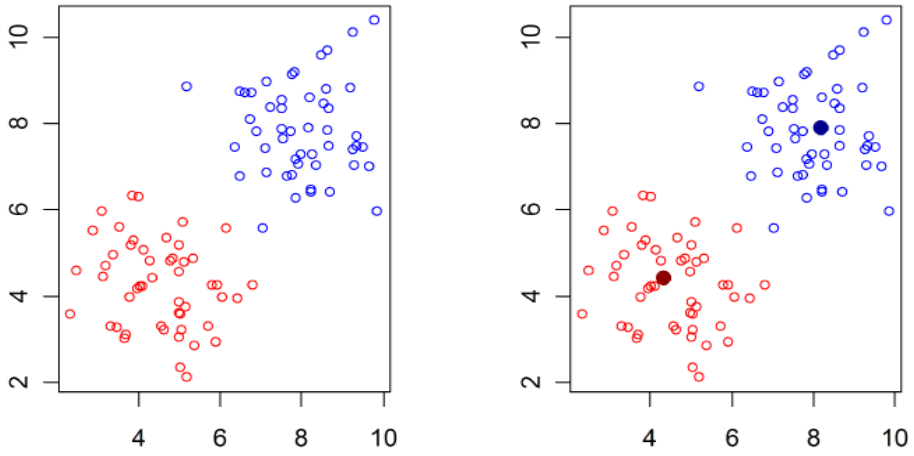

FIGURE 1.5 – Un exemple simple de discrimination en deux groupes dans $\mathbb{R}^2$.

Un exemple plus complexe

Absence de données aberrantes

Considérons maintenant un jeu de données un peu plus complexe. Dans un premier temps nos simulations sont effectuées de la façon suivante : chacun des $n = 100$ individus est un vecteur dans $\mathbb{R}^{5000}$ dont chaque coordonnée suit une loi normale. Pour les 60 premiers individus, chaque coordonnée suit une loi $\mathcal{N}(a, 1)$, et pour les 40 derniers, chaque coordonnée suit une loi $\mathcal{N}(b, 1)$ (avec a et b deux réels). Nous comparons ensuite les groupes déterminés par nos méthodes aux vrais groupes. Le Tableau 1.4 rassemble les taux de bonne classification des individus pour les différents algorithmes (Lloyd version medoïdes et Alter) pour des distorsions définies à partir des distances L_1 et L_2. La Figure 1.6, construite sur le même modèle que la Figure 1.5, représente les projetés des n individus sur les deux premières coordonnées.

En terme de taux de bonne classification comme en terme de variance, l'algorithme Alter donne toujours de meilleurs résultats que l'algorithme de Lloyd. Pour une même méthodologie, les résultats en distance L_2 sont sensiblement

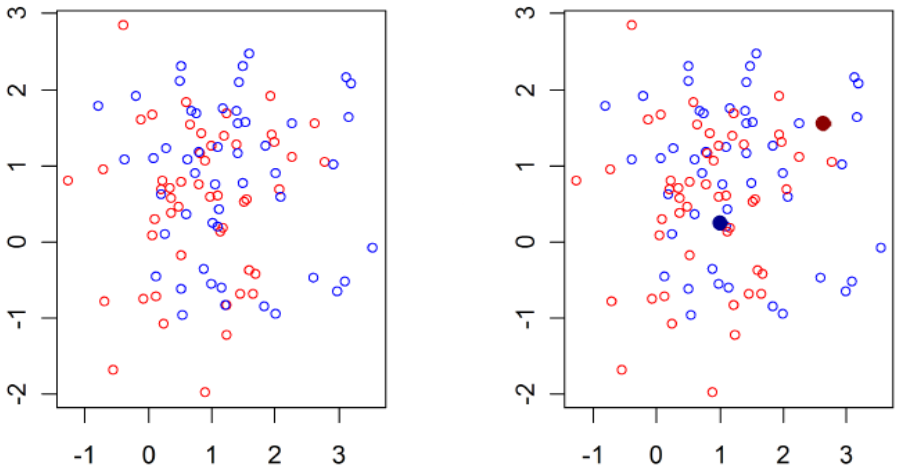

FIGURE 1.6 – Illustration graphique de la discrimination des données simulées de dimension 5000.

Vecteurs comparés	Lloyd	Alter
$\mathcal{N}(0.6,1)$ et $\mathcal{N}(1,1)$ (L_1)	0.856 (var=0.0466)	0.999 (var=10^{-6})
$\mathcal{N}(0.7,1)$ et $\mathcal{N}(1,1)$ (L_1)	0.823 (var=0.0405)	0.975 (var=0.0004)
$\mathcal{N}(0.8,1)$ et $\mathcal{N}(1,1)$ (L_1)	0.715 (var=0.0193)	0.804 (var=0.01)
$\mathcal{N}(0.6,1)$ et $\mathcal{N}(1,1)$ (L_2)	0.845 (var=0.0452)	0.999 (var=10^{-6})
$\mathcal{N}(0.7,1)$ et $\mathcal{N}(1,1)$ (L_2)	0.830 (var=0.0407)	0.974 (var=0.0003)
$\mathcal{N}(0.8,1)$ et $\mathcal{N}(1,1)$ (L_2)	0.719 (var=0.0171)	0.786 (var=0.0097)

TABLE 1.4 – Taux de bonne classification pour les différentes méthodes.

les mêmes que ceux en distance L_1.

Présence de données aberrantes

L'intérêt de la distance L_1 réside principalement dans sa robustesse aux valeurs extrêmes. Pour mettre cette propriété en évidence, nous allons perturber l'échantillon précédent en rajoutant un individu (dans $\mathbb{R}^{5000}$) dont chaque coordonnée suit une loi $\mathcal{N}(200,1)$. Les résultats sont rassemblés dans le Tableau 1.5.

La distance L_1 donne à présent de meilleurs résultats que la distance L_2,

Vecteurs comparés	Lloyd	Alter
$\mathcal{N}(0.6,1)$ et $\mathcal{N}(1,1)$ et un $\mathcal{N}(200,1)$ (L_1)	0.842 (var=0.0474)	0.999 (var=10^{-6})
$\mathcal{N}(0.7,1)$ et $\mathcal{N}(1,1)$ et un $\mathcal{N}(200,1)$ (L_1)	0.822 (var=0.0395)	0.979 (var=0.0039)
$\mathcal{N}(0.8,1)$ et $\mathcal{N}(1,1)$ et un $\mathcal{N}(200,1)$ (L_1)	0.705 (var=0.0201)	0.807 (var=0.0117)
$\mathcal{N}(0.6,1)$ et $\mathcal{N}(1,1)$ et un $\mathcal{N}(200,1)$ (L_2)	0.827 (var=0.0466)	0.998 (var=10^{-6})
$\mathcal{N}(0.7,1)$ et $\mathcal{N}(1,1)$ et un $\mathcal{N}(200,1)$ (L_2)	0.812 (var= 0.0463)	0.935 (var=0.0044)
$\mathcal{N}(0.8,1)$ et $\mathcal{N}(1,1)$ et un $\mathcal{N}(200,1)$ (L_2)	0.698 (var=0.0173)	0.752 (var=0.0098)

TABLE 1.5 – Influence d'une valeur extrême.

notamment lorsque les groupes sont proches. Plus précisément, les performances pour la distance L_1 restent les mêmes que pour l'échantillon non perturbé, alors que celles pour la distance L_2 se dégradent considérablement.

Afin de compléter cette étude, nous avons perturbé un peu plus l'échantillon en ajoutant un individu construit à partir d'une $\mathcal{N}(200,1)$ supplémentaire. Nous avons rassemblé les résultats dans le Tableau 1.6. À nouveau les performances pour la distance L_2 se dégradent alors que celles pour la distance L_1 restent stables.

Vecteurs comparés	Lloyd	Alter
$\mathcal{N}(0.6,1)$ et $\mathcal{N}(1,1)$ et deux $\mathcal{N}(200,1)$ (L_1)	0.831(var=0.0472)	0.996 (var=10^{-6})
$\mathcal{N}(0.7,1)$ et $\mathcal{N}(1,1)$ et deux $\mathcal{N}(200,1)$ (L_1)	0.825(var=0.0393)	0.973 (var=0.0004)
$\mathcal{N}(0.8,1)$ et $\mathcal{N}(1,1)$ et deux $\mathcal{N}(200,1)$ (L_1)	0.714 (var=0.0213)	0.808 (var=0.0135)
$\mathcal{N}(0.6,1)$ et $\mathcal{N}(1,1)$ et deux $\mathcal{N}(200,1)$ (L_2)	0.827 (var=0.0448)	0.998 (var=10^{-6})
$\mathcal{N}(0.7,1)$ et $\mathcal{N}(1,1)$ et deux $\mathcal{N}(200,1)$ (L_2)	0.812 (var=0.0409)	0.931 (var=0.0003)
$\mathcal{N}(0.8,1)$ et $\mathcal{N}(1,1)$ et deux $\mathcal{N}(200,1)$ (L_2)	0.698 (var=0.0162)	0.746 (var=0.0012)

TABLE 1.6 – Influence de valeurs extrêmes.

Un exemple de données fonctionnelles : le processus d'Ornstein-Uhlenbeck

A présent, nous allons utiliser les méthodes présentées dans ce chapitre pour discriminer des trajectoires de processus d'Ornstein-Uhlenbeck. Pour une introduction détaillée sur le sujet, nous renvoyons le lecteur au livre de Iacus [18], Section 1.13.

Construction d'un processus d'Ornstein-Uhlenbeck

Un processus d'Ornstein-Uhlenbeck, où processus de Vasicek, est l'unique solution de l'équation différentielle stochastique

$$dX_t = (\theta_1 - \theta_2 X_t)dt + \theta_3 dW_t, \quad X_0 = x_0,$$

où W_t est le mouvement brownien, $\theta_3 \in \mathbb{R}_+$ et $\theta_1, \theta_2, x_0 \in \mathbb{R}$. Le cas $\theta_1 = 0$ a tout d'abord été présenté par Ornstein et Uhlenbeck en 1930 [31]. Vasicek a ensuite proposé en 1977 une généralisation au cas θ_1 quelconque pour modéliser des taux d'intérêts.

Pour $\theta_2 > 0$, on obtient un processus de retour à la moyenne, ce qui signifie que le processus tend à osciller autour d'une certaine valeur d'équilibre. Une autre propriété intéressante de ce processus et que, contrairement au mouvement Brownien, il admet une variance finie pour tout $t \geq 0$.

Une autre paramétrisation du processus de Ornstein-Uhlenbeck, plus communément utilisée en finance, est

$$dX_t = \theta(\mu - X_t)dt + \sigma dW_t, \quad X_0 = x_0, \tag{1.24}$$

où σ représente la volatilité, μ est la valeur d'équilibre vers laquelle tend le processus, et θ est la vitesse de retour à la moyenne.

En plus des applications classiques en finance, les processus d'Ornstein-Uhlenbeck sont aujourd'hui utilisés en biostatistique pour approcher des pro-

cessus de branchement, par exemple pour modéliser des flux de pollen (Istas [19]).

Classification de processus d'Ornstein-Uhlenbeck simulés

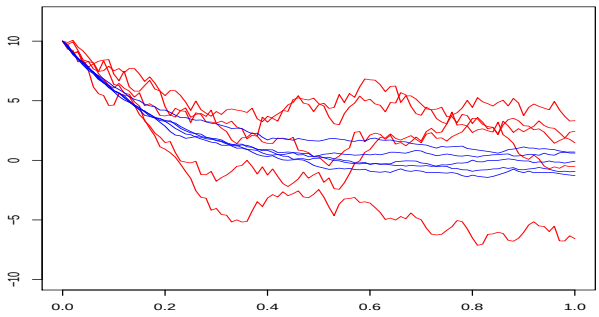

FIGURE 1.7 – Exemple de trajectoires à classifier

On considère l'équation (1.24), avec $\mu = 0$ et $\sigma, \theta > 0$. Par ailleurs on prend $t \in [0, 1]$, et on fixe $\theta = 5, x_0 = 10$. Nous allons simuler des trajectoires à partir de deux processus calculés avec des valeurs de σ différentes (Figure 1.7). Plus précisément, on simule 75 trajectoires d'un processus X_t calculé avec $\sigma = \sigma_1$, et 75 trajectoires d'un processus Y_t calculé avec $\sigma = \sigma_2$. Les simulations sont réalisées en utilisant la méthode présentée dans la Section 1.13.1 de [18]. Comme dans le paragraphe précédent, on compare les résultats obtenus par les algorithmes Lloyd et Alter (Tableau 1.7).

Processus comparés	Lloyd	Alter
$\sigma_1 = 1$ et $\sigma_2 = 6$ (L_1)	0.608 (var=0.0038)	0.655 (var=0.0006)
$\sigma_1 = 1$ et $\sigma_2 = 8$ (L_1)	0.598 (var=0.0035)	0.662 (var=0.0008)
$\sigma_1 = 1$ et $\sigma_2 = 6$ (L_2)	0.597 (var=0.0023)	0.655 (var=0.001)
$\sigma_1 = 1$ et $\sigma_2 = 8$ (L_2)	0.622 (var=0.0036)	0.655 (var=0.001)

TABLE 1.7 – Taux de bonne classification de processus d'Ornstein-Uhlenbeck.

Comme dans le paragraphe précédent, on s'aperçoit que les résultats sont meilleurs pour Alter que pour Lloyd. De plus les résultats sont à nouveau équivalents pour les distance L_1 et L_2. Nous allons cette fois encore perturber les échantillons en ajoutant une trajectoire d'un processus calculé avec $\sigma = 20$. Les résultats sont rassemblés dans le Tableau 1.8.

Processus comparés	Lloyd	Alter
$\sigma_1 = 1$ et $\sigma_2 = 6$ (L_1)	0.601 (var=0.0029)	0.649 (var=0.0007)
$\sigma_1 = 1$ et $\sigma_2 = 8$ (L_1)	0.586 (var=0.0038)	0.657 (var=0.0011)
$\sigma_1 = 1$ et $\sigma_2 = 6$ (L_2)	0.563 (var=0.0028)	0.626 (var=0.0007)
$\sigma_1 = 1$ et $\sigma_2 = 8$ (L_2)	0.591 (var=0.0033)	0.619 (var=0.0012)

TABLE 1.8 – Influence d'une valeur extrême.

On peut noter que les résultats pour la distance L_1 sont stables alors que ceux pour la distance L_2 se dégradent.

1.4.2 Études sur des données réelles

Nous utilisons une partie de la base de données TIMIT **(http ://www-stat.stanford.edu/ tibs/ElemStatLearn/)**. Les données correspondent à des enregistrement de phonèmes de durée 32 ms. Nous voulons répartir ces enregistrements en 5 groupes, correspondant aux 5 phonèmes suivantes : "sh" comme dans "she" (872 données), "dcl" comme dans "dark" (757 données), "iy" comme la voyelle dans "she" (1163 données), "aa" comme la voyelle dans "dark" (695 données) et "ao" comme la première voyelle dans "water" (1022 données). Chaque son est enregistré à 16 kHz et nous ne retenons que les 256 premières fréquences (voir Figure 1.4.2). Nous disposons ainsi de 4509 vecteurs de taille 256. Au vu du nombre de données et de groupes, nous n'avons pu employer l'algorithme Alter, et nous avons à la place utilisé l'algorithme Alter-fast présenté dans le Paragraphe 1.4. (cf. Tableau 1.9). On note que les résultats de l'algorithme Alter-fast sont meilleurs que ceux de l'algorithme de Lloyd.

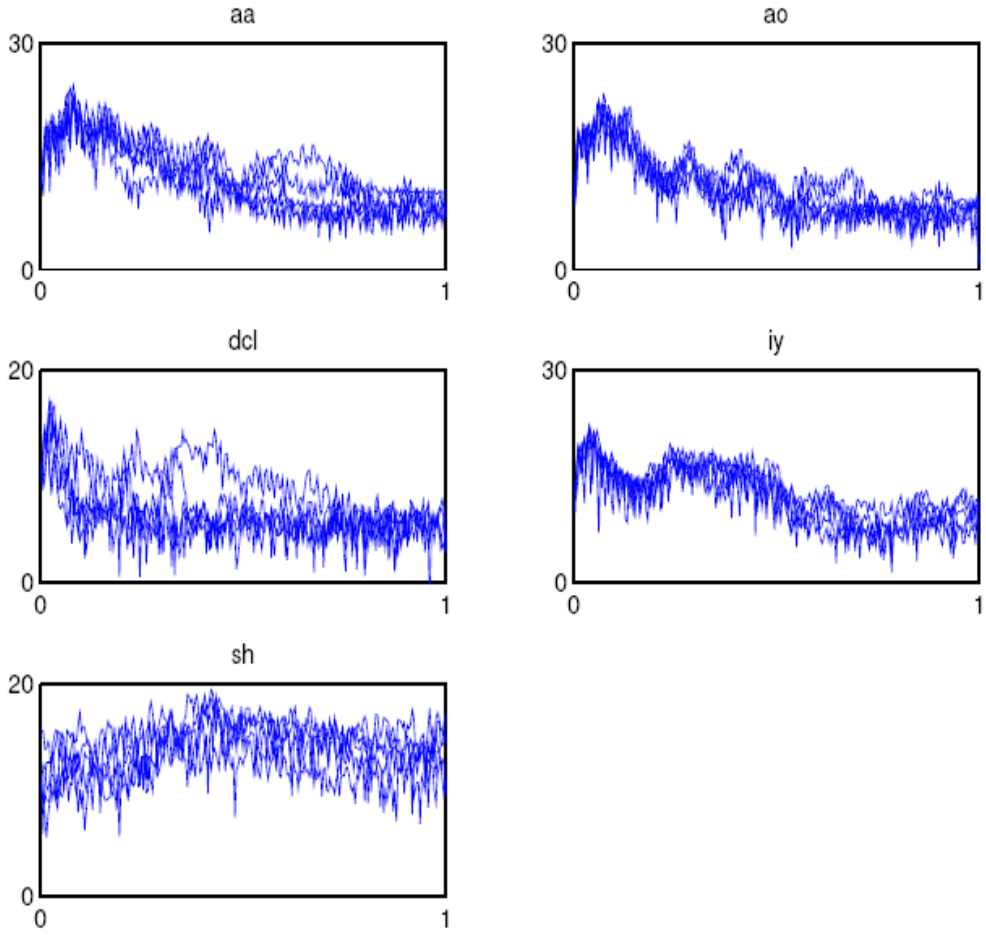

FIGURE 1.8 – Un exemple de chacun des 5 phonèmes.

Algorithme	Taux de bonne classification
Lloyd (L_1)	0.80 (var=0.0047)
Alter-fast (L_1)	0.84 (var=0.00014)

TABLE 1.9 – Taux de bonne classification des cinq phonèmes.

Les phonèmes “ao” et “aa” sont particulièrement difficiles à classer. Pour illustrer ce phénomène, nous avons sélectionné dans la base de données ces deux types de courbes en essayant de les séparer à l'aide nos algorithmes. Nous avons comparé les performances des algorithmes Lloyd, Alter et Alter-fast (cf. Tableau 1.10). Comme prévu, les résultats obtenus ne sont pas satisfaisants. On remarque cependant que l'algorithme Alter donne des résultats bien plus intéressants que l'algorithme de Lloyd, et que l'algorithme Alter-fast offre un compromis appréciable. Notons de plus que nos résultats sont significativement meilleurs que ceux obtenus par Bleakley [5] (Chapitre 2). Il n'y a pas ici de variance pour l'algorithme Alter. Cela est du au fait que nous ne disposons ici que d'un seul jeu de données, et pas d'un grand nombre d'échantillon simulés. Les variances des algorithmes Lloyd et Alter-fast ne représentent que la variabilité due à leur dépendance à leur différent paramètres (centres initiaux pour Lloyd, sous populations sélectionnées pour Alter-fast).

Algorithme	Taux de bonne classification
Lloyd (L_1)	0.64 (var=0.0031)
Alter (L_1)	0.71
Alter-fast (L_1)	0.68 (var=0.00015)
Max. bin. kernel [5]	0.61
Min. bin. kernel [5]	0.63
Pseu.-Markov NN [5]	0.57

TABLE 1.10 – Taux de bonne classification des phonèmes “aa” et “ao”.

Enfin, nous avons fait une étude similaire en supprimant les phonèmes “ao” de la base de données (cf. Tableau 1.11). Les résultats sont sensiblement meilleurs qu'avec la base de données dans son intégralité.

Algorithme	Taux de bonne classification
Lloyd (L_1)	0.87 (var=0.0032)
Alter-fast (L_1)	0.90 (var=0.0001)

TABLE 1.11 – Taux de bonne classification sans la phonème “ao”.

Bibliographie

[1] R. A. Adams. *Sobolev Spaces*. Pure and applied mathematics. Elsevier, 1975.

[2] C. Berlinet, Alain et Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academics Publishers, 2004.

[3] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54 :781–790, 2007.

[4] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, third edition, 1995. A Wiley-Interscience Publication.

[5] K. Bleakley. *Quelques Contributions à l'Analyse Statistique et à la Classification des Graphes et des Courbes. Applications à l'Immunobiologie et à la Reconstruction des Réseaux Biologiques*. PhD thesis, Université Montpellier II, 2007.

[6] F. Bolley, A. Guillin, and C. Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3-4) :541–593, 2007.

[7] B. Cadre. Convergent estimators for the L_1-median of a Banach valued random variable. *Statistics*, 35(4) :509–521, 2001.

[8] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1) :1–49, 2002.

[9] H. Djellout, A. Guillin, and L. Wu. Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *Annals of Probability*, 32(3B) :2702–2732, 2004.

[10] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, New York, second edition, 2001.

[11] R. M. Dudley. *Real Analysis and Probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

[12] N. Dunford and J. T. Schwartz. *Linear Operators. Part I.* Wiley Classics Library. John Wiley & Sons Inc., New York, 1988. General theory, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1958 original, A Wiley-Interscience Publication.

[13] I. Ekeland and R. Temam. *Analyse Convexe et Problèmes Variationnels.* Dunod, 1974. Collection Études Mathématiques.

[14] G. Gan, C. Ma, and J. Wu. *Data Clustering*, volume 20 of *ASA-SIAM Series on Statistics and Applied Probability.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

[15] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression.* Kluwer Academic Publishers, Norwell, MA, USA, 1991.

[16] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics.* Springer-Verlag, Berlin, 2000.

[17] J. A. Hartigan. *Clustering Algorithms.* John Wiley & Sons, New York-London-Sydney, 1975. Wiley Series in Probability and Mathematical Statistics.

[18] S. M. Iacus. *Simulation and Inference for Stochastic Differential Equations : With R Examples.* Springer Series in Statistics. Springer, New York, 2008.

[19] J. Istas. *Mathematical Modeling for the Life Sciences.* Springer-Verlag, 2005.

[20] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data : An Introduction to Cluster Analysis.* Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1990.

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

[22] J. Kogan. *Introduction to Clustering Large and High-Dimensional Data.* Cambridge University Press, Cambridge, 1954.

[23] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22 :49–86, 1951.

[24] M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs.* American Mathematical Society, 2001.

[25] T. Linder. Learning-theoretic methods in vector quantization. In *Principles of Nonparametric Learning (Udine, 2001)*, volume 434 of *CISM Courses and Lectures*, pages 163–210. Springer, Vienna, 2002.

[26] T. Linder, G. Lugosi, and K. Zeger. Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Transactions on Information Theory*, 40 :1728–1740, 1994.

[27] F. Malrieu. Logarithmic sobolev inequalities for some nonlinear pde's. *Stochastic Processes and their Applications*, 95(1) :109–132, 2001.

[28] B. Mirkin. *Mathematical Classification and Clustering*, volume 11 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, 1996.

[29] D. Pollard. A central limit theorem for k-means clustering. *The Annals of Probability*, 10 :919–926, 1982.

[30] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.

[31] G. E. Uhlenbeck and L. S. Ornstein. On the theory of the brownian motion. *Physical Review*, 36(5) :823–841, 1930.

[32] A. W. Van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

Chapitre 2

Application à la typologie des bancs d'anchois au Pérou

2.1 Introduction

L'étude des populations de poissons exploités implique que soient connues et mesurées leurs relations et interactions avec l'environnement. Ces interactions sont caractérisées par des indicateurs, construits à partir des informations qui peuvent être extraites du milieu. Une des caractéristiques particulières des poissons pélagiques côtiers dont l'anchois du Pérou (Engraulis ringens) est un exemple remarquable, est de former des bancs importants, regroupant plusieurs dizaines ou centaines de milliers d'individus (Fréon et Misund [4]). Ces bancs représentent l'instrument d'interaction des poissons avec leur milieu immédiat (Bertrand et al. [2]). On doit donc pouvoir, en observant ces bancs, obtenir des indicateurs de leurs relations avec leur environnement et de l'état de la population (Reid [10]). Ces indicateurs ont mené à la construction de typologies (Petitgas et Levenez [9]), et ont démontré leur efficacité. On a pu par exemple montrer que les proportions des différents types de bancs de chinchards du Pacifique (Trachurus murphyi) changeaient lors des périodes "El Nino" (Barbieri, Córdova et Catasti [1]).

Jusqu'à la fin des années 90 la seule information spatiale possible provenait des sondeurs verticaux (Simmonds et MacLennan [11]). La conception de

sonars multifaisceaux à hautes fréquences a ensuite permis de fournir des images tridimensionnelles d'un banc (Gerlotto, Soria et Fréon [7]), à partir desquelles il est possible de reconstruire le banc et de mesurer sa morphologie externe (dimensions, surface, volume, etc.) et sa structure interne (densité, hétérogénéité, etc.) (Gerlotto et Paramo [6]). Ces informations ont permis d'améliorer la compréhension du fonctionnement d'un banc et de ses relations avec son environnement.

Il a été possible en particulier de mesurer l'effet de l'évitement des bancs face à un navire de recherche (Soria, Fréon, Gerlotto et Paramo [12]), de décrire dans le détail l'organisation du banc et d'émettre l'hypothèse d'une structuration auto-organisée (Gerlotto et Paramo [6]), de décrire les effets de la prédation sur un banc, et d'observer les instruments de communication dont dispose le poisson pour y faire face (Gerlotto, Bertrand, Bez et Gutiérrez [5]). Néanmoins, jusqu'à maintenant, il n'existait pas de méthode systématique pour caractériser les bancs et en faire une typologie en trois dimensions. Il devenait donc intéressant de vérifier s'il était possible de mettre au point un instrument automatique permettant de construire une typologie à partir des représentations tridimensionnelles reconstituées à partir des échos sonar. Au vu du nombre important de points de discrétisations de l'image, il nous a paru intéressant de répondre à cette question en appliquant la méthode mise au point dans le chapitre précédent. L'échantillon de bancs utilisé pour cette étude provient d'une campagne expérimentale, effectuée au Pérou en novembre 2004. Par la suite, il était important de vérifier si une éventuelle distinction de groupes pouvait être reliée à des caractéristiques biologiques ou écologiques différentes.

2.2 Matériel et méthode

2.2.1 Le sonar

Nous avons utilisé un sonar RESON 6012 "Seabat" adapté spécifiquement à l'observation des bancs de poissons (Fernandes, Gerlotto, Soria et Sim-

monds [3]). Il s'agit d'un sonar multifaisceaux, fréquence 455 kHz, durée d'impulsion 0.064 ms, formé de 60 faisceaux contigus de 1.5° qui permettent d'observer un champ de 90° à une portée fixée à 100 m dans notre expérience. L'angle d'observation dans la dimension perpendiculaire au plan observé est de 22°. Le sonar a été implanté sur le navire océanographique Olaya (navire de l'IMARPE, Pérou), le plan principal d'observation étant vertical sur 90°, ce qui permet d'observer depuis la surface de la mer jusqu'à la verticale du bateau (Figure 2.1). Le navire se déplaçant à la vitesse de 8 noeuds, la troisième dimension est obtenue à l'aide de plans successifs en faisant l'hypothèse que les mouvements du banc sont négligeables par rapport à la vitesse du navire (un banc d'anchois se déplace en moyenne à une vitesse de 0.25 m/s, à comparer aux 4 m/s du navire). Les plans successifs (taux de répétitions des émissions : 3.5 par seconde) permettent alors de reconstruire la distribution des poissons dans la dimension parallèle à la route du navire.

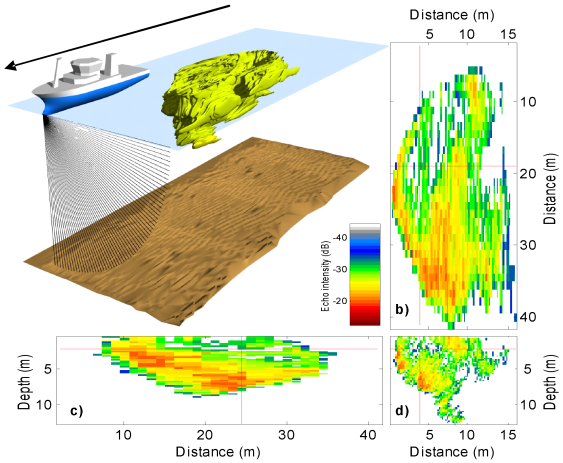

FIGURE 2.1 – Description de la méthode d'observation avec un sonar multifaisceaux Seabat. Les trois figures à droite et sous le schéma représentent trois coupes du banc représenté en trois dimensions dans l'image centrale, dans les trois plans d'observation (coupes au niveau de la plus grande extension du banc dans le plan considéré). Fernandes, communication personnelle.

Les informations principales que nous avons utilisées sont fournies par l'analyse des images tridimensionnelles. Ce sont les suivantes : longueur L (dimension horizontale parallèle à la route du navire) ; largeur l (dimension horizontale perpendiculaire à la précédente) ; hauteur h (dimension verticale) ; surface S ; volume V ; rugosité R ($R = S/V$) ; volume des vacuoles V_v en m^3 (une vacuole est un ensemble de "voxels" [1] de densité nulle) ; surface de l'enveloppe des vacuoles S_v en m^2 ; densité moyenne d.

2.2.2 La campagne

Les données ont été extraites d'une campagne expérimentale de l'*Instituto del Mar del Peru* (IMARPE) réalisée en novembre 2004 (Bertrand, Gerlotto, Soria et Gutiérrez [2]). Le trajet a consisté en deux séries de répétitions, durant 24 heures, d'un carré de deux milles nautiques de coté au large de Pucusana et de Paracas, le long de la côte péruvienne (Figure 2.2). Un échantillon de 96 bancs a été extrait dans la journée du 18 novembre (position moyenne : 12° 36' S, 76° 54' W). Nous avons sélectionné uniquement des données de jour, les bancs se désagrégeant la nuit.

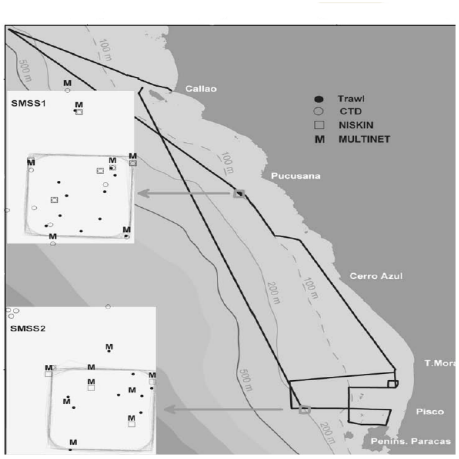

FIGURE 2.2 – Lieu de la campagne.Bertrand et al.[2]

1. le terme Voxel est la contraction de "volumetric Pixel" et désigne donc un pixel en trois dimensions. Un Voxel est ici de dimension $0.25m \times 0.25m \times 0.8m$

2.2.3 La méthode

Nous disposons donc ici de représentations tridimensionnelles des bancs d'anchois. Plus précisément, les données se présentent sous la forme de matrices à quatre colonnes. Les trois premières renseignent les coordonnées sur les trois axes, et la quatrième la densité de l'écho sonar associé aux différentes coordonnées.

Pour discriminer les bancs de poissons, nous avons utilisé l'algorithme Alter présenté dans le chapitre précédent (Section 1.3.2, page 68). Pour éviter un biais relatif à la position des bancs par rapport au bateau, nous avons centré tous les bancs sur la même position. Pour pouvoir appliquer notre algorithme il nous a fallu faire en sorte que toutes les données aient le même nombre de lignes. Pour ce faire, nous avons pris comme référence le banc le plus grand. Ensuite, pour tous les autre bancs, nous avons rajouté des 0 aux coordonnées n'ayant pas de valeur. En faisant ceci, nous avons pu introduire un nouveau biais en séparant artificiellement les gros bancs des petits. Pour l'éviter, nous avons sélectionné une petite portion de chaque banc pour effectuer la discrimination.

Une fois toutes ces opérations effectuées, nous avons cherché s'il était possible d'effectuer une discrimination pertinente en deux groupes. Les résultats sont donnés dans la section suivante.

2.3 Résultats

2.3.1 La discrimination

Deux classes ont pu être extraites à partir de l'application de la méthode à la base de données. Une "classe 1", formée de 83 bancs ; et une "classe 2", formée de 13 bancs. La première action consiste à vérifier si certains paramètres des bancs sont significativement différents dans les deux classes. Nous avons réalisé des tests de Student et Kolmogorov-Smirnov sur les bancs

pour les 9 variables présentées à la fin de la Section 2.2.1 (Tableaux 2.1 et 2.2). Ils montrent que les caractéristiques des variables sont différentes dans les deux groupes. On peut noter toutefois que la rugosité et la densité moyenne sont les variables les moins différentes (bien qu'elles le soient significativement).

Variable	p-value	Moyenne classe 1	Moyenne classe 2
Longueur	$p < 0.001$	29.095	75.77
Largeur	$p < 0.001$	17.321	49.93
Hauteur	$p < 0.001$	11.917	27.12
Volume	$p < 0.001$	1054.482	12110.28
Surface	$p < 0.001$	3775.738	33293.49
rugosité	$p < 0.025$	3.975	2.95
Surface vacuoles	$p < 0.001$	6.649	2 .73
Volume vacuoles	$p < 0.001$	2.985	0.91
densité	$p < 0.025$	69.014	85.78

TABLE 2.1 – Test de Kolmogorov-Smirnov (non paramétrique) de comparaison de deux séries. N(classe 1) = 83 ; N(classe 2) = 13. La surface et le volume des vacuoles sont ramenés respectivement à une surface et un volume de 1 m^2 et $1m^3$ du banc.

Variable	Moy. classe 1	Moy. classe 2	t-value	df	p
Longueur	29.095	75.77	$p < 0.001$	94	0.000000
Largeur	17.321	49.93	$p < 0.001$	94	0.000000
Hauteur	11.917	27.12	$p < 0.001$	94	0.000000
Volume	1054.482	12110.28	$p < 0.001$	94	0.000000
Surface	3775.738	33293.49	$p < 0.001$	94	0.000000
rugosité	3.975	2.95	$p < 0.025$	94	0.004721
Surface vacuoles	6.649	2 .73	$p < 0.001$	94	0.000050
Volume vacuoles	2.985	0.91	$p < 0.001$	94	0.000012
densité	69.014	85.78	$p < 0.025$	94	0.073960

TABLE 2.2 – Valeurs du test de Student pour les principaux paramètres. Mêmes remarques que pour le tableau 1.

2.3.2 Caractéristiques des types

Nous avons vu dans le paragraphe précédent qu'il existait bien deux types de bancs. Il est intéressant de les décrire plus en détails.

- La classe 2 se caractérise par des dimensions supérieures à celles de la classe 1. Rappelons que la classification ayant été réalisée sur une portion égale dans chaque banc, ces dimensions ne sont pas intervenues dans la discrimination des deux classes. Par ailleurs la classe 2 regroupe des bancs plus denses que ceux de la classe 1.
- A l'inverse les caractéristiques structurales (volume et surface des vacuoles, rugosité) sont plus importantes dans la classe 1 que dans la classe 2.

Dans tous les cas, à l'exception du volume, les dimensions des variables se chevauchent entre les deux classes (Tableau 2.3).

Var.	Moy.cl.1	Moy.cl.2	Min.cl.1	Min.cl.2	Max.cl.1	Max.cl.2
Long.	29.095	75.77	10.0600	45.67	72.78	128.3
Larg.	17.321	49.93	5.4700	25.79	45.83	101.9
Haut.	11.917	27.12	3.9300	15.20	36.19	42.4
Vol.	1054.482	12110.28	29.4780	5060.13	3751.67	25433.8
Surf.	3775.738	33293.49	219.3430	12940.65	19065.22	101493.7
Rug.	3.975	2.95	1.4450	1.21	7.45	4.9
S.vac.	6.649	2.73	0.0000	0.77	15.44	4.4
V.vac.	2.983	0.91	0.0000	0.36	7.32	1.9
Dens.	69.014	85.78	25.0821	57.45	170.59	171.4

TABLE 2.3 – Comparaison des dimensions maximales et minimales des classes de bancs pour toutes les variables étudiées. On voit que seul le volume est une variable qui ne se chevauche pas entre les deux classes.

2.3.3 Relations avec l'environnement

Les groupes étant identifiés et validés, il faut maintenant tester si des relations avec l'environnement peuvent être mises en évidence.

La première observation que l'on peut faire est que l'on retrouve les deux groupes mélangés tout au long de l'expérience. Nous présentons sur la Figure 2.3 la distribution des bancs par classe au long de l'expérience. On peut voir qu'il n'y a aucune indication de relation entre la classe et l'heure (et donc entre la classe et la position du navire), les bancs de la classe 2 (la moins nombreuse) se trouvant toujours présents avec des bancs de la classe 1.

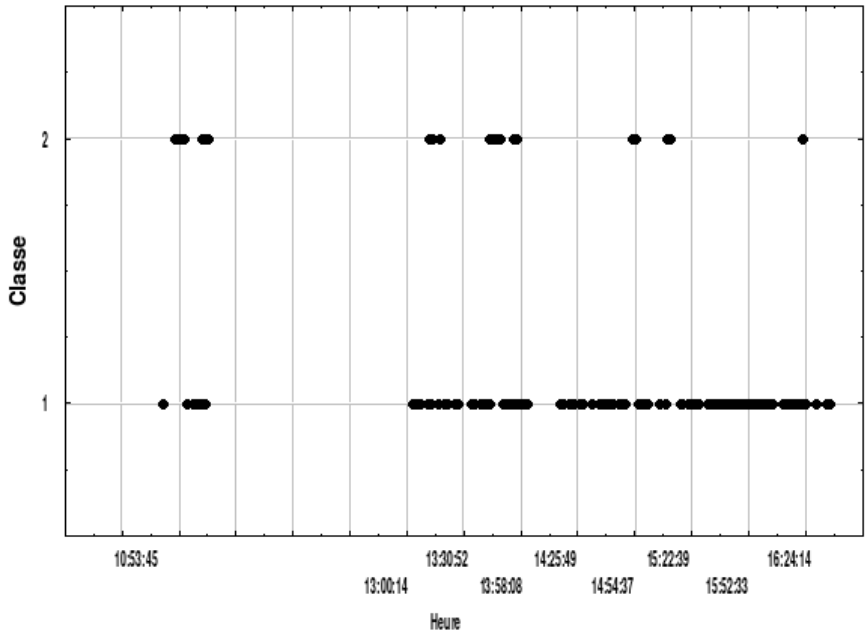

FIGURE 2.3 – Distribution des bancs des classes 1 (en bas) et 2 (en haut) en fonction de l'heure, et donc de la position du navire.

Il existe par ailleurs une caractéristique environnementale de dimension très réduite mais dont l'effet potentiel n'est pas négligeable : les ondes internes (Figure 2.4). Elles forment des fronts visibles en surface, ce qui nous a permis de savoir pendant l'enregistrement des données si un banc se trouvait dans une onde interne ou non.

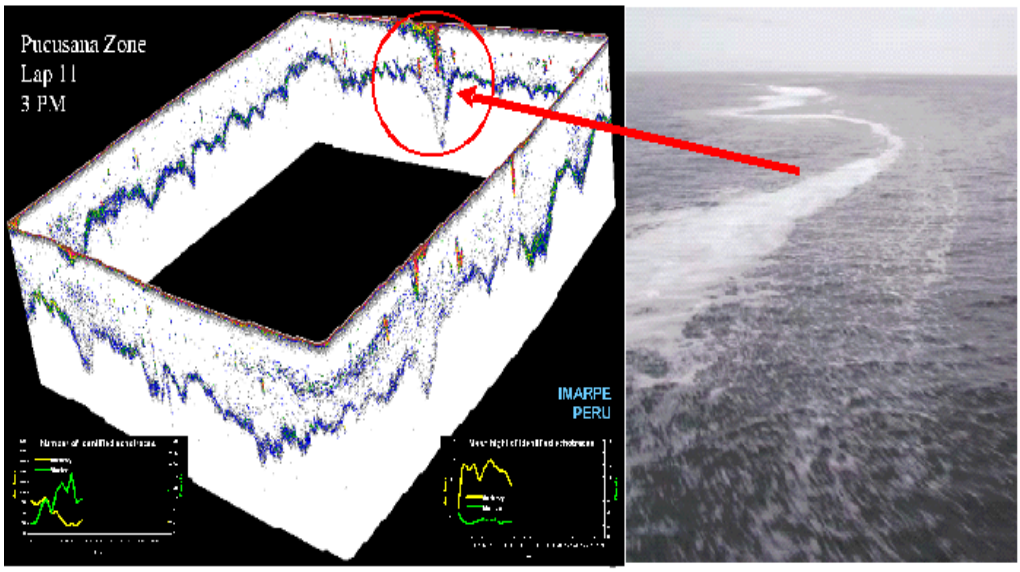

FIGURE 2.4 – Exemple d'échogramme enregistré pendant la campagne de novembre 2004 avec un sondeur vertical (SIMRAD, EK60, 120 kHz). L'échelle du sondeur est de 100 mètres et les côtés du carré prospecté font 2 milles nautique de longueur. Le front visible en surface (à droite) est produit par une onde interne (cercle rouge, à gauche). Bertrand et al. [2]

Parmi les bancs étudiés dans ce travail, 13 ont été observés dans une zone avec front, c'est-à-dire dans une onde interne (voir Tableau 2.4). Tout en

	Avec front	Sans front	Total
Effectif classe 2	1	12	13
Effectif classe 1	12	73	85

TABLE 2.4 – Tableau de contingence des bancs des classes 1 et 2 dans les zones avec et sans fronts.

sachant que la faible proportion de bancs concernés par des ondes internes rend les tests peu puissants, nous avons effectué un test exact de Fisher pour déterminer si les distributions des classes 1 et 2 étaient significativement différentes en fonction des ondes internes. Les résultats obtenus montrent que, sous l'hypothèse d'indépendance des facteurs "front" et "classe", la répartition observée (Tableau 2.4) est la plus probable (conditionnellement

aux marges observées), avec une probabilité de 0,31. Il apparaît donc que l'on ne peut relier de façon significative les facteurs classe et onde interne.

2.4 Discussion

Nous avons vu que les deux classes présentent des caractéristiques très différentes, tant en morphologie qu'en structure, et que ces différences ne pouvaient être corrélées aux caractéristiques environnementales disponibles.

Dans ces conditions une autre explication possible réside dans des différences dues à des caractéristiques comportementales individuelles. Ce point a été étudié dans une autre expérience réalisée durant la même campagne et au même endroit, en déployant le sonar dans le plan horizontal et en maintenant le navire en dérive, afin d'observer la dynamique des bancs. Cette expérience a été conçue pour observer les réactions des bancs à l'attaque par des lions de mer (Gerlotto, Bertrand, Bez et Gutiérrez [5]). Notons que les bancs observés en deux dimensions (coupe horizontale) et en dynamique (navire en dérive) montrent des évolutions rapides de leurs structures quand ils sont attaqués par des lions de mer (Gerlotto, Bertrand, Bez et Gutiérrez [5]). Ces changements aboutissent en général à la construction de bancs plus grands et plus homogènes (Figure 2.5). La diminution des espaces vides (vacuoles) étant l'expression d'une plus grande homogénéité, nous pouvons faire l'hypothèse que la classe 2 nous montre les bancs ayant été soumis à une prédation. Il faudra toutefois réaliser un test dans le milieu pour le confirmer.

Le filtre mis au point permet donc a priori de définir la part des bancs sous prédation dans une zone donnée sans avoir à observer les prédateurs. Le fait que les deux groupes ne soient pas entièrement distincts dans leurs différentes dimensions (sauf le volume) montre tout l'intérêt de la discrimination : elle repérerait les bancs attaqués quelle que soit leur taille. Nous aurions alors là un instrument pouvant nous permettre une classification automatique des bancs soumis à prédation, indépendamment de leur taille. Un tel indicateur

présente un grand intérêt, car il devrait permettre de proposer des pistes pour les calculs de mortalité par prédation (qui représente une partie importante de la mortalité naturelle) sur laquelle nous ne disposons à l'heure actuelle de pratiquement aucune information directe (Laurec et Le Guen [8]). L'estimation de cet aspect de la mortalité naturelle est importante. En effet, les modèles de dynamique des populations s'appuient sur cette valeur de mortalité naturelle dont l'observation est généralement fondée sur des données théoriques, voire intuitives.

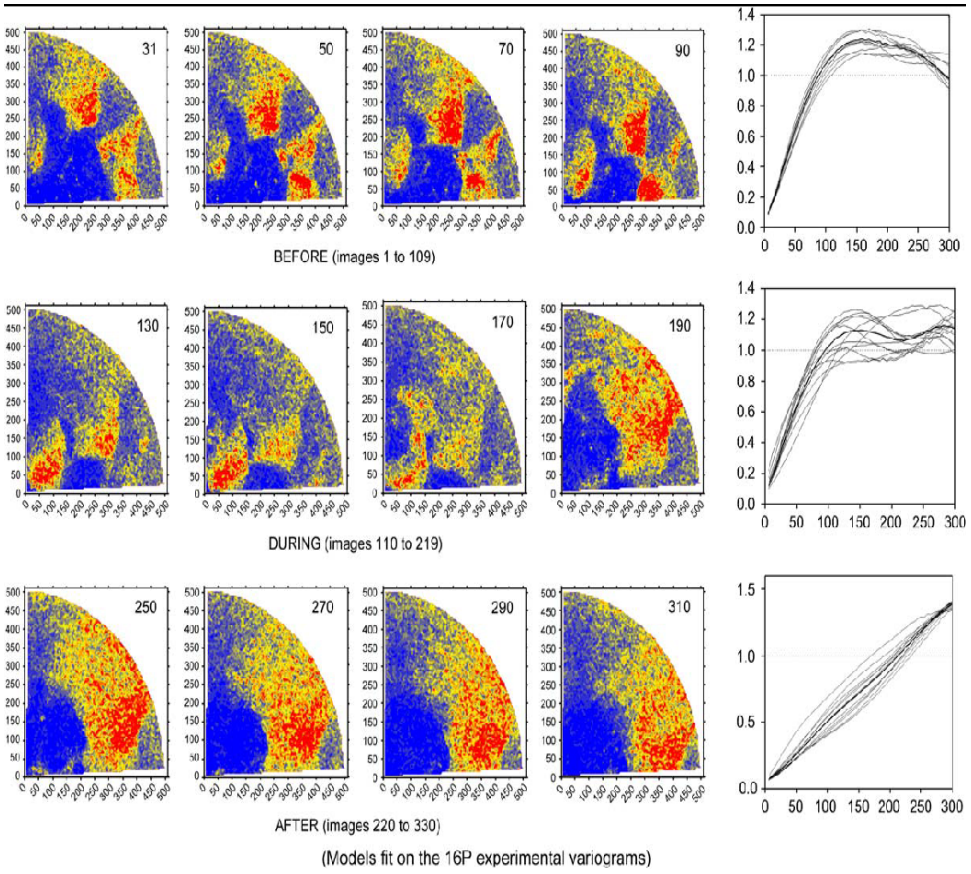

FIGURE 2.5 – Différences de structures avant, pendant et après une attaque par lions de mer. Les images représentent des séries de séquences avant, pendant et après la prédation (images horizontales prises durant une dérive du navire océanographique Olaya) ; les cadres (à droite) rassemblent les variogrammes standardisés pour les images prises durant les trois phases. Gerlotto, Bertrand, Bez et Gutiérrez [5]

Bibliographie

[1] M. A. Barbieri, J. Córdova, and V. Catasti. Distribución espacial del jurel (*Trachurus Murphy*) y su relación con las variables ambientales en la zona central de chile en el período 1997-2004. XXV Congreso de Ciencias del Mar y XI Congreso Latinoamericano de Ciencias del Mar, 2005.

[2] A. Bertrand, F. Gerlotto, S. Bertrand, M. Gutiérrez, L. Alza, A. Chipollini, E. Díaz, P. Espinoza, J. Ledesma, R. Quesquén, S. Peraltilla, and F. Chavez. Schooling behaviour and environmental forcing in relation to anchoveta distributiojn : an analysis across multiple spatial scales. *Progress in Oceanography*, 79 :264–277, 2008.

[3] P. Fernandes, F. Gerlotto, M. Soria, and E. Simmonds. Into the next dimension : three dimensional acoustic observations of fish schools aggregations. *ICES Conference Meeting*, J33, 1998.

[4] P. Fréon and O. Misund. *Dynamics of Pelagic-Fish Distribution and Behaviour : Effects on Fisheries and Stock Assessment*. Blackwell, Oxford, 1999.

[5] F. Gerlotto, S. Bertrand, N. Bez, and M. Gutiérrez. Waves of agitation inside anchovy schools observed with multibeam sonar : a way to transmit information in response to predation. *ICES Journal of Marine Science*, 53 :1405–1417, 2006.

[6] F. Gerlotto and J. Paramo. The three-dimensional morphology and internal structure of clupeid schools as observed using vertical scanning multibeam sonar. *Aquatic Living Resources*, 16 :113–122, 2003.

[7] F. Gerlotto, M. Soria, and P. Fréon. From two dimensions to three : the use of multibeam sonar for a new approach in fisheries acoustics. *Canadian Journal of Fisheries and Aquatic Sciences*, 56 :6–12, 1999.

[8] A. Laurec and J. C. Le Guen. *Dynamique des Populations Marines Exploitées, Tome 1 : Concepts et Méthodes*. Number 45 in Rapports Scientifiques et Techniques. Publications du Centre National pour l'Exploitation des Océans, 1981.

[9] P. Petitgas and J. J. Levenez. Spatial organization of pelagic fish : echogram structure, spatio-temporal condition, and biomass in senegalese waters. *ICES Journal of Marine Science*, 53 :147–153, 2008.

[10] D. G. Reid. Report on echo trace classification. *ICES Cooperative Research Reports*, 238, 2000.

[11] J. Simmonds and D. N. MacLennan. *Fisheries Acoustics : Theory and Practice, 2nd Edition*. Wiley-Blackwell, 2005.

[12] M. Soria, P. Fréon, F. Gerlotto, and J. Paramo. Analysis of vessel influence on spatial behaviour of fish schools using a multi-beam sonar and consequences for biomass estimates by echo-sounder. *ICES Journal of Marine Science*, 53 :453–458, 1996.

Troisième partie

Estimation non paramétrique des ensembles de niveau pour la régression

Chapitre 1

Estimation non paramétrique des ensembles de niveau pour la régression

1.1 Introduction

Dans ce chapitre, nous considérons le problème de l'estimation des ensembles de niveau de la fonction de régression. Plus précisément, étant donné un couple aléatoire (X, Y) à valeurs dans $\Lambda \times J$, $\Lambda \subset \mathbb{R}^d$ et $J \subset \mathbb{R}$ étant supposés bornés, nous allons chercher à estimer les ensembles de niveau de la fonction de régression r de Y sur X, définie pour tout $x \in \Lambda$ par

$$r(x) = \mathbb{E}\left[Y|X = x\right].$$

Pour $t > 0$, un ensemble de niveau pour r est défini par

$$\mathcal{L}(t) = \{x \in \Lambda : r(x) > t\}.$$

On suppose disposer d'un échantillon i.i.d $\Big((X_1, Y_1), \ldots, (X_n, Y_n)\Big)$ de même loi que (X, Y). On considère alors un estimateur de type plug-in de $\mathcal{L}(t)$. Plus précisément, à partir d'un estimateur consistant $\hat{r}_n$ de r, on estime $\mathcal{L}(t)$ par

$$\mathcal{L}_n(t) = \{x \in \Lambda : \hat{r}_n(x) > t\}.$$

On utilise ainsi la méthode présentée par Cadre [3] dans le cas de la fonction de densité.

L'essentiel des travaux de recherche sur le thème de l'estimation d'ensembles de niveau concerne la fonction de densité. On peut citer par exemple les travaux de Cadre [3], Cuevas et Fraiman [7], Hartigan [10], Polonik [14], Tsybakov [17], Walther [18]. Ce foisonnement de travaux sur le sujet est motivée par le grand nombre d'applications possibles. L'estimation de ces ensembles de niveau est notamment utile en estimation du mode (Müller et Stawitzki [12], Polonik [14]), ou encore en clustering (Biau, Cadre et Pelletier [1], Cuevas, Febrero et Fraiman [6, 5]). En particulier, Biau, Cadre et Pelletier [1] utilisent un estimateur des ensembles de niveau de la fonction de densité pour apporter des éléments de réponse au problème de la détermination du nombre de clusters.

Les mêmes applications sont envisageables dans le cas de la fonction de régression. Par ailleurs, il est par exemple possible d'utiliser un estimateur des ensembles de niveau de la fonction de régression pour déterminer le trajet de l'écoulement de l'eau à partir de représentations numériques de la topographie d'une zone géographique. De la même manière, en imagerie médicale, il peut être utile d'estimer les zones où certaines fonctions de l'image dépassent un seuil fixé, par exemple pour déterminer automatiquement le lieu ou la nature d'une tumeur. On remarque que, dans ces deux exemples, l'utilisation de domaines compacts Λ et J se justifie pleinement. C'est en fait le cas dans la plupart des situations pratiques, et plus particulièrement en analyse d'images.

En dépit des nombreuses applications possibles, l'estimation des ensembles de niveau de la fonction de régression reste relativement peu étudiée dans la littérature. Müller [11] en parle brièvement dans son *survey*. On peut également citer les travaux plus récents de Cavalier [4], Scott et Davenport [16], et Willett et Nowak [13]. L'estimateur proposé par Cavalier est fondé

sur la maximisation de la masse en excès, et adapte celui proposé par Tsybakov [17] dans le cas de la fonction de densité. Sous certaines hypothèses, notamment sur la régularité de la fonction de régression et sur la forme de ses ensembles de niveau, il obtient une vitesse minimax. Scott et Davenport adoptent quant-à eux une approche de type *cost sensitive*. Pour résumer, ces auteurs mesurent la qualité de l'approximation par l'espérance du coût de mauvaise classification. La méthode se révèle efficace mais souffre cependant d'une trop grande dépendance au choix de la fonction de coût.

Les différents résultats de convergence seront donnés au sens de la différence symétrique (Figure 1.1), définie par

$$\mathcal{L}_n \Delta \mathcal{L} = (\mathcal{L}_n \cap \mathcal{L}^C) \cup (\mathcal{L}_n^C \cap \mathcal{L}),$$

où, comme à chaque fois qu'il n'y aura pas de confusion possible, $\mathcal{L}_n = \mathcal{L}_n(t)$ et $\mathcal{L} = \mathcal{L}(t)$.

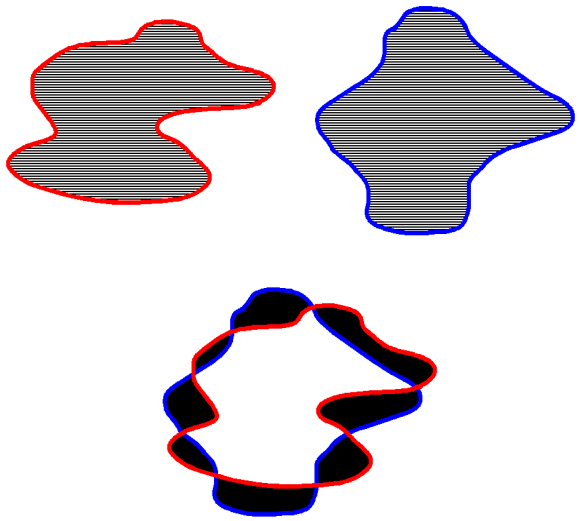

FIGURE 1.1 – Illustration de la différence symétrique (zone noire) entre deux ensembles A (en rouge) et B (en bleu).

Notre approche consiste à établir des résultats sous des hypothèses sur r et $\hat{r}_n$ aussi raisonnables que possible. En utilisant les résultats de Cuevas, González-Manteiga et Rodríguez-Casal [8], nous commençons par présenter un résultat de convergence valable pour tout estimateur consistant $\hat{r}_n$ de r. Ensuite, nous particularisons notre approche en considérant le cas de l'estimateur à noyau de la régression. Pour cet estimateur, nous obtenons une vitesse de convergence du même ordre que celle obtenu par Cadre [3] dans le cas de la densité.

Le chapitre est organisé de la façon suivante. Nous commençons par présenter les principaux résultats dans la Section 1.2. La Section 1.3 est ensuite dédiée à la confrontation de notre estimateur à des données simulées. Enfin, les preuves sont rassemblées dans la Section 1.4.

1.2 Résultats principaux

1.2.1 Convergence de l'estimateur

Dans toute la suite, $\|.\|$ représente la norme euclidienne sur un espace de dimension finie. Par ailleurs, pour toute fonction intégrable $g : \Lambda \to \mathbb{R}$, on note $\|g\|_p$ la norme définie par

$$\|g\|_p = \left(\int_\Lambda |g(x)|^p dx \right)^{1/p}.$$

Estimateur quelconque $\hat{r}_n$ de r

Dans ce paragraphe, nous supposons disposer d'un estimateur consistant (en un sens qui sera défini plus bas) $\hat{r}_n$ de r. On introduit l'hypothèse, notée **H**,

H $\lambda(\{r = t\}) = 0$.

Cette hypothèse signifie simplement que la mesure de Lebesgue ne charge pas l'ensemble où r vaut t.

Le Théorème 3 dans l'article de Cuevas, González-Manteiga et Rodríguez-Casal [8] établit la convergence presque sûre de $\lambda(\mathcal{L}_n \Delta \mathcal{L})$ pour tout estimateur consistant $\hat{r}_n$ de r. Le Théorème 1.2.1 ci-dessous complète ce résultat en traitant le cas de la convergence de $\mathbb{E}\lambda(\mathcal{L}_n \Delta \mathcal{L})$.

Théorème 1.2.1 *Supposons que* **H** *est vérifiée. Si*

$$\mathbb{E}\,\|\hat{r}_n - r\|_p \underset{n\to\infty}{\to} 0 \quad ou \quad \sup_{\Lambda} |\hat{r}_n - r| \underset{n\to\infty}{\to} 0 \quad p.s.\ ,$$

alors

$$\mathbb{E}\,\lambda\Big(\mathcal{L}_n(t)\Delta\mathcal{L}(t)\Big) \underset{n\to\infty}{\to} 0.$$

Ainsi, pourvu que l'estimateur $\hat{r}_n$ de r soit consistant au sens de la norme L_p ou de la norme sup, l'estimateur plug-in des ensembles de niveau de la fonction de régression est consistant au sens de la différence symétrique. Par exemple, l'estimateur BSE (Bosq et Lecoutre [2], Chapitre 7), celui des plus proches voisins ([2] Chapitre 8) ainsi que le régressogramme ([2], Chapitre 6) satisfont cette propriété. Dans le paragraphe suivant, nous étudions en particulier le cas de l'estimateur à noyau.

Estimateur à noyau r_n de r

Nous examinons à présent le cas de l'estimateur à noyau de la fonction de régression. On suppose qu'il est possible de réécrire r sous la forme

$$r(x) = \frac{\varphi(x)}{f(x)},$$

où f est la fonction de densité de X, et φ est définie par $\varphi(x) = r(x)f(x)$.

Soit K un noyau sur $\mathbb{R}^d$, c'est-à-dire une densité de probabilité sur $\mathbb{R}^d$. On note $K_h(x) = K(x/h)$. A partir d'un échantillon i.i.d. $\Big((X_1, Y_1), \ldots, (X_n, Y_n)\Big)$, on définit, pour tout $x \in \Lambda$,

$$\varphi_n(x) = \frac{1}{nh^d}\sum_{i=1}^{n} Y_i K_h(x - X_i) \text{ et } f_n(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K_h(x - X_i).$$

Pour tout $x \in \Lambda$, l'estimateur à noyau de r est alors défini par

$$r_n(x) = \begin{cases} \varphi_n(x)/f_n(x) & \text{si } f_n(x) \neq 0 \\ 0 \text{ sinon.} \end{cases}$$

Pour une étude détaillée de l'estimateur à noyau de la fonction de régression et de ses propriétés, nous renvoyons au livre de Prakasa Rao [15].

Les propriétés de convergence de l'estimateur à noyau de la régression (Bosq et Lecoutre [2]) nous permettent de déduire du Théorème 1.2.1 ci-dessus le résultat suivant :

Corollaire 1.2.1 *Supposons que* **H** *est vérifiée. Si K est borné, intégrable, à support compact et Lipschitzien, et si $h \to 0$ et $nh^d/\log n \to \infty$, alors*

$$\mathbb{E}\,\lambda\Big(\mathcal{L}_n(t)\Delta\mathcal{L}(t)\Big) \underset{n\to\infty}{\to} 0.$$

Avec ce résultat, nous particularisons celui de Cuevas, González-Manteiga et Rodríguez-Casal [8] au cas de l'estimateur à noyau. Cela nous permet d'établir des conditions concrètes sur la fenêtre pour définir notre estimateur. Dans le paragraphe suivant, nous nous attachons à établir une vitesse de convergence pour notre estimateur.

1.2.2 Vitesse de convergence

Dans ce paragraphe, nous ne considérons plus que le cas où r_n est l'estimateur à noyau de la fonction de régression. Dans la suite, $\Theta \subset (0, \sup_\Lambda r)$ représente un intervalle ouvert. Nous aurons besoin des hypothèses suivantes :

H1 Les fonctions r et f sont deux fois continûment différentiables, et $\inf_\Lambda f > 0$;

H2 Pour tout $t \in \Theta$,

$$\inf_{r^{-1}(\{t\})} \|\nabla r\| > 0,$$

où, ici et dans toute la suite, $\nabla\psi(x)$ représente l'opérateur gradient au point $x \in \mathbb{R}^d$ de la fonction différentiable $\psi : \mathbb{R}^d \to \mathbb{R}$.

Notons que les hypothèses **H1** et **H2** impliquent (Proposition $A.2$ dans [3])

$$\forall t \in \Theta : \quad \lambda(r^{-1}[t-\varepsilon, t+\varepsilon]) \to 0 \quad \text{quand } \varepsilon \to 0.$$

Cette propriété, utilisée par Polonik [14], est presque identique à l'hypothèse **H** du Paragraphe 1.2.1. Introduisons à présent les hypothèses sur le noyau K.

H3 K est un noyau tel que $\lim_{\|x\|\to\infty} \|x\|^s K(x) = 0$, continûment différentiable et à support compact. De plus, il existe une fonction décroissante $\mu : \mathbb{R}^+ \to \mathbb{R}$ telle que $K(x) = \mu(\|x\|)$ pour tout $x \in \mathbb{R}^d$.

L'hypothèse **H3** est par exemple satisfaite par le noyau gaussien ou encore celui d'Epanechnikov.

Dans toute la suite, on notera ∂A la frontière de tout ensemble $A \subset \Lambda$. Par ailleurs, on introduit $\mathcal{H}$ la mesure de Hausdorff de dimension $d-1$. Pour une présentation détaillée de la mesure de Hausdorff et de ses propriétés, nous renvoyons le lecteur au livre de Evans et Gariepy [9]. Nous nous contenterons ici de souligner le fait que cette mesure nous permet d'utiliser la formule de la coaire (Proposition A.1 dans [3]), cruciale dans la démonstration du Théorème 1.2.2. On note enfin $\widetilde{K} = \int K^2 d\lambda$.

Nous sommes désormais en mesure d'établir une vitesse de convergence pour $\mathbb{E}\,\lambda(\mathcal{L}_n(t)\Delta\mathcal{L}(t))$.

Théorème 1.2.2 *Sous les hypothèses* **H1** $-$ **H3**, *si* $nh^d/(\log n)^7 \to \infty$ *et* $nh^{d+4}\log n \to 0$, *alors il existe deux constantes strictement positives* C_1 *et* C_2 *telles que, pour presque tout* $t \in \Theta$,

$$\underline{\lim}_{n\to\infty} \left(\sqrt{nh^d}\mathbb{E}\,\lambda(\mathcal{L}_n(t)\Delta\mathcal{L}(t))\right) \geq C_1\sqrt{\frac{2t}{\pi}\widetilde{K}}\int_{\partial\mathcal{L}(t)} \frac{d\mathcal{H}}{\|\nabla r\|},$$

et

$$\overline{\lim_{n\to\infty}} \left(\sqrt{nh^d} \mathbb{E}\, \lambda(\mathcal{L}_n(t) \Delta \mathcal{L}(t)) \right) \leq C_2 \sqrt{\frac{2t}{\pi} \widetilde{K}} \int_{\partial \mathcal{L}(t)} \frac{d\mathcal{H}}{\|\nabla r\|}.$$

Notons que l'estimateur que nous proposons est facile à calculer et requiert des hypothèses raisonnables. Cavalier [4] obtient une meilleure vitesse, mais avec un estimateur plus difficile à calculer et des hypothèses plus restrictives. Il impose notamment aux ensembles de niveau d'être de forme étoilée, ce qui n'est pas notre cas. Cependant le choix de la fenêtre de notre estimateur est un problème délicat. Nous avons pris le parti de choisir en pratique la fenêtre optimale pour l'estimation de la fonction de régression, bien que rien n'assure en théorie qu'elle soit optimale pour l'estimation des ensembles de niveau.

1.3 Applications

Dans cette section, nous allons confronter notre estimateur à des données simulées. Nous reprenons avec ces simulations l'idée de l'analyse d'images. On considère la fonction $r : [-6.5, 4.5] \times [-6.5, 4.5] \to [-2, 2]$ définie par

$$r : (u, v) \mapsto \sin(u) + \sin(v).$$

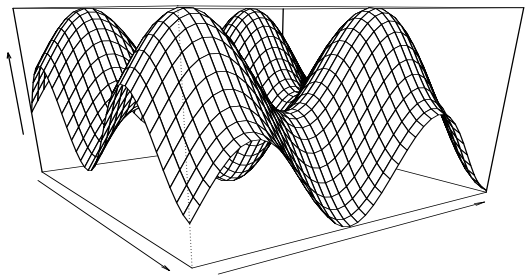

FIGURE 1.2 – Représentation de $r(u, v) = sin(u) + sin(v)$ sur $[-6.5, 4.5] \times [-6.5, 4.5]$.

Nous faisons le choix de cette fonction particulière pour vérifier si notre estimateur détecte correctement les différentes composantes connexes des ensembles de niveau. On voit en effet sur la Figure 1.2 que nous aurons, pour des valeurs de t adéquates, quatre composantes connexes pour $\mathcal{L}(t) = \{(u, v) : r(u, v) > t\}$. Nous nous intéressons à cet aspect particulier car, comme nous l'avons évoqué dans l'introduction, une possible application de l'estimation des ensembles de niveau est l'utilisation des composantes connexes pour faire de la classification.

Soit X une variable aléatoire suivant une loi binormale dans $\mathbb{R}^2$ (tronquée sur $[-6.5, 4.5]\times[-6.5, 4.5]$) centrée en $(-1, -1)$ et de matrice de covariance $6*Id$. On pose $Y_i = r(X_i)+\varepsilon_i$, où les ε_i sont indépendants et suivent une loi normale centrée et d'écart-type 0.1. Pour différentes valeurs de n et t, nous voulons estimer $\mathcal{L}(t)$ à partir d'un échantillon i.i.d. de taille n $\Big((X_1, Y_1), \ldots, (X_n, Y_n)\Big)$ distribué selon la loi du couple (X, Y). On considère les niveaux $t = 0.5$, $t = 1$ et $t = 1.5$. Cela nous permet notamment d'étudier le comportement de notre estimateur face à des composantes connexes plus ou moins éloignées les unes des autres. De plus, nous considérons différentes tailles d'échantillon : $n = 500$, $n = 2500$ et $n = 7500$. Dans tous les cas, nous calculons notre estimateur à partir des n données de l'échantillon, puis nous l'évaluons sur une grille régulière de 3364 points sur $[-6.5, 4.5] \times [-6.5, 4.5]$. Les résultats sont représentés dans les Figures 1.3 à 1.11. Chaque figure représente les ensembles de niveau vrais et estimés, ainsi que la superposition des deux ensembles, pour les différentes valeurs de n et t. Les tailles de fenêtre sont déterminées automatiquement par la fonction *npregbw* de R.

Dans l'ensemble des cas étudiés, on voit que l'ensemble de niveau estimé est proche du vrai ensemble de niveau, et que les quatres composantes connexes sont bien identifiées. On note par ailleurs sans surprise que la qualité de l'estimation augmente avec la taille de l'échantillon. Cependant, il apparaît que pour cette fonction f, un nombre relativement réduit d'observations ($n =$

2500) est suffisant pour obtenir une estimation satisfaisante. Pour un nombre encore plus réduit ($n = 500$), on remarque tout de même des écarts assez importants. Enfin, on note que la qualité de l'estimation est la même pour chaque valeur de t, ce qui est conforme à la théorie.

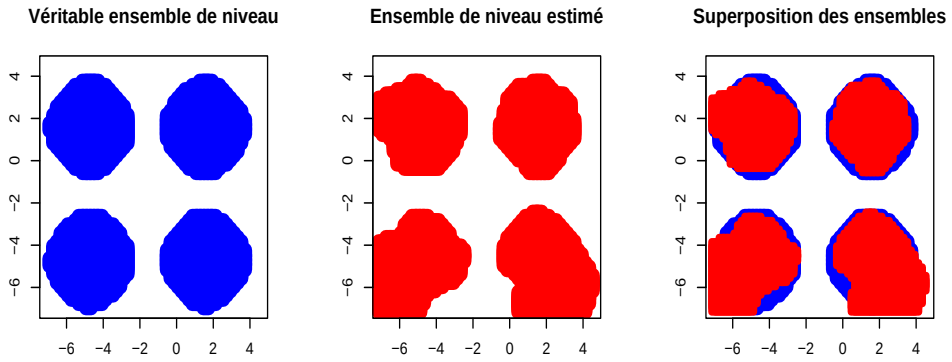

FIGURE 1.3 – Représentation graphique des ensembles de niveau vrais (bleu) et estimés (rouge) pour $n = 500$ et $t = 0.5$.

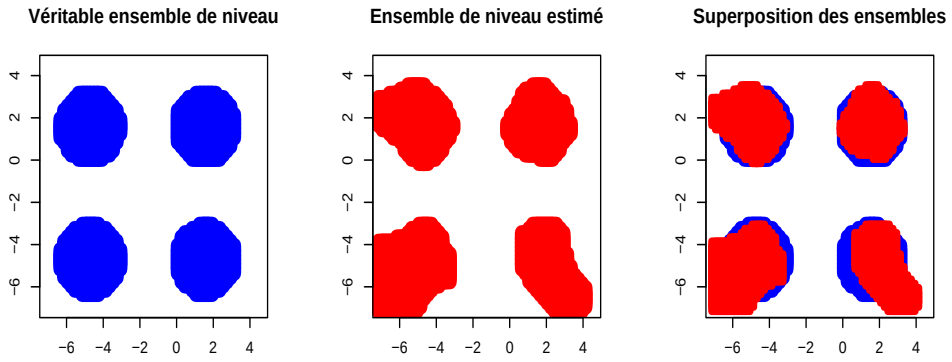

FIGURE 1.4 – Représentation graphique des ensembles de niveau vrais (bleu) et estimés (rouge) pour $n = 500$ et $t = 1$.

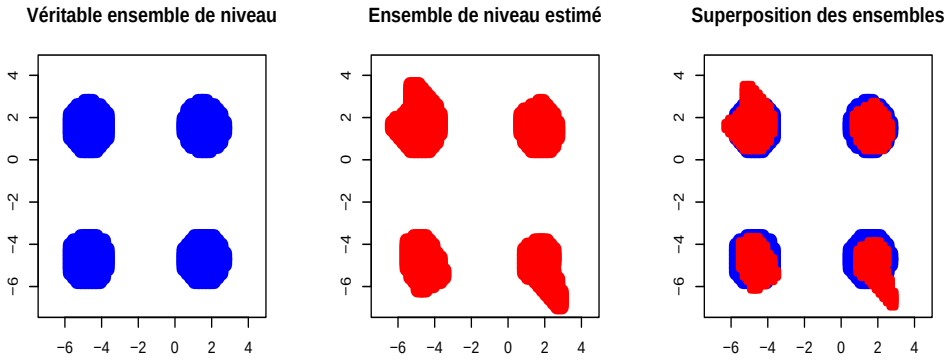

FIGURE 1.5 – Représentation graphique des ensembles de niveau vrais (bleu) et estimés (rouge) pour $n = 500$ et $t = 1.5$.

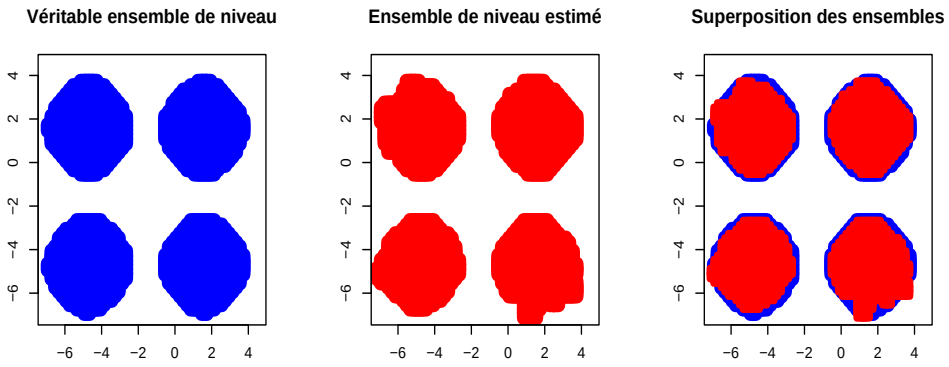

FIGURE 1.6 – Représentation graphique des ensembles de niveau vrais (bleu) et estimés (rouge) pour $n = 2500$ et $t = 0.5$.

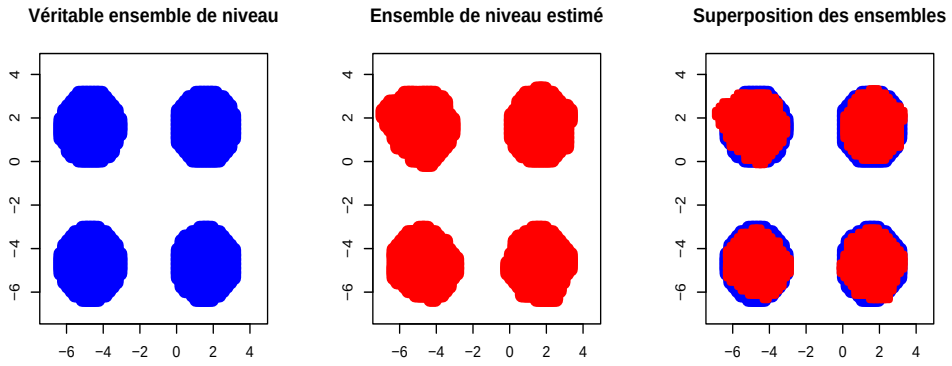

FIGURE 1.7 – Représentation graphique des ensembles de niveau vrais (bleu) et estimés (rouge) pour $n = 2500$ et $t = 1$.

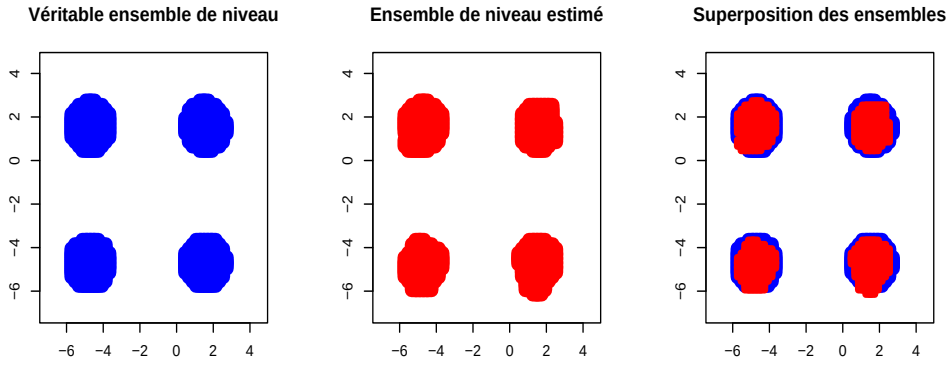

FIGURE 1.8 – Représentation graphique des ensembles de niveau vrais (bleu) et estimés (rouge) pour $n = 2500$ et $t = 1.5$.

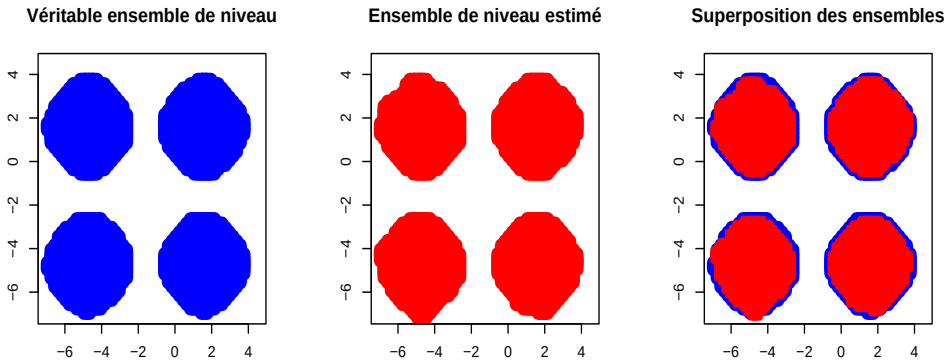

FIGURE 1.9 – Représentation graphique des ensembles de niveau vrais (bleu) et estimés (rouge) pour $n = 7500$ et $t = 0.5$.

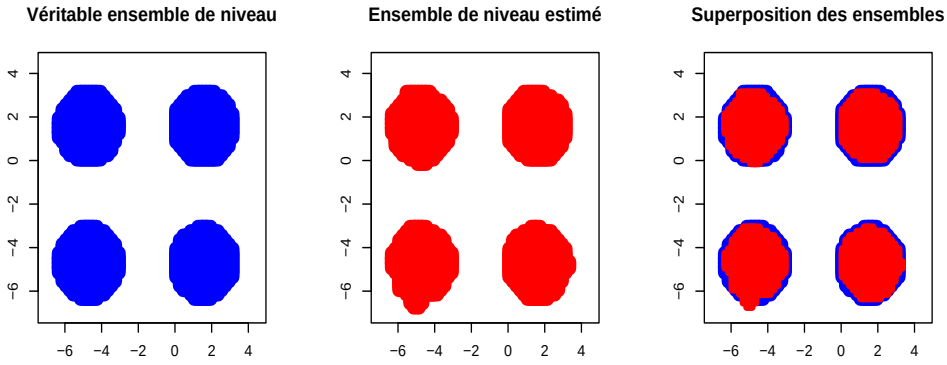

FIGURE 1.10 – Représentation graphique des ensembles de niveau vrais (bleu) et estimés (rouge) pour $n = 7500$ et $t = 1$.

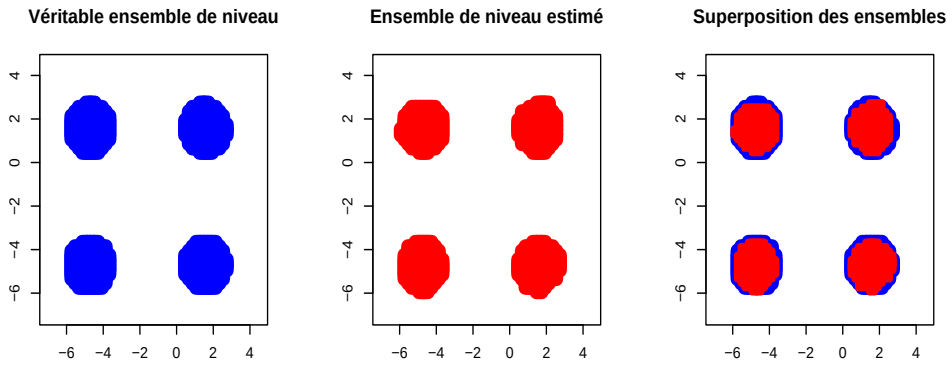

FIGURE 1.11 – Représentation graphique des ensembles de niveau vrais (bleu) et estimés (rouge) pour $n = 7500$ et $t = 1.5$.

1.4 Preuves

Cette section est dédiée à la preuve du Théorème 1.2.1 ainsi qu'à celle du Théorème 1.2.2. Dans toute la suite, c, C_1, et C_2 désigneront des constantes positives dont les valeurs pourront varier d'une ligne à l'autre.

1.4.1 Preuve du Théorème 1.2.1

Soit $\varepsilon > 0$. Pour tout $n \in \mathbb{N}$, il est clair que les ensembles (aléatoires) $E_{n,\varepsilon} = \{x \in \Lambda : |\hat{r}_n(x) - r(x)| \leq \varepsilon\}$ et $\mathcal{L}_n(t)\Delta\mathcal{L}(t)$ sont mesurables et que

$$\lambda(\mathcal{L}_n(t)\Delta\mathcal{L}(t)) = \lambda\left((\mathcal{L}_n(t)\Delta\mathcal{L}(t)) \cap E_{n,\varepsilon}\right) + \lambda\left((\mathcal{L}_n(t)\Delta\mathcal{L}(t)) \cap E_{n,\varepsilon}^c\right).$$

Comme $(\mathcal{L}_n(t)\Delta\mathcal{L}(t)) \cap E_{n,\varepsilon} \subset \{t - \varepsilon \leq r \leq t + \varepsilon\}$, on a

$$\lambda(\mathcal{L}_n(t)\Delta\mathcal{L}(t)) \leq \lambda(\{t - \varepsilon \leq r \leq t + \varepsilon\}) + \lambda(E_{n,\varepsilon}^c),$$

et donc

$$\mathbb{E}\lambda(\mathcal{L}_n(t)\Delta\mathcal{L}(t)) \leq \lambda(\{t - \varepsilon \leq r \leq t + \varepsilon\}) + \mathbb{E}\lambda(E_{n,\varepsilon}^c). \tag{1.1}$$

De plus, si $\sup_\Lambda |\hat{r}_n - r| \to 0$ p.s., alors $\lambda(E_{n,\varepsilon}^c) \underset{n\to\infty}{\to} 0$ et, comme Λ est borné, $\mathbb{E}\lambda(E_{n,\varepsilon}^c) \underset{n\to\infty}{\to} 0$ par le théorème de convergence dominée de Lebesgue. Si $\|\hat{r}_n - r\|_p \to 0$, alors

$$\begin{aligned}
\mathbb{E}\,\lambda(E_{n,\varepsilon}^c) &= \mathbb{E}\lambda(\{x \in \Lambda : |\hat{r}_n(x) - r(x)| > \varepsilon\}) \\
&= \mathbb{E}\int_\Lambda \mathbf{1}_{|\hat{r}_n(x) - r(x)| > \varepsilon} dx \\
&\leq \frac{1}{\varepsilon^p}\mathbb{E}\int_\Lambda |\hat{r}_n(x) - r(x)|^p dx \underset{n\to\infty}{\to} 0.
\end{aligned}$$

On déduit donc de (1.1) que, pour $\varepsilon > 0$,

$$\overline{\lim_{n\to\infty}}\, \mathbb{E}\lambda(\mathcal{L}_n(t)\Delta\mathcal{L}(t)) \leq \lambda(\{t - \varepsilon \leq r \leq t + \varepsilon\}).$$

Par ailleurs, puisque on a supposé $\lambda(\{r = t\}) = 0$, on a

$$\lim_{\varepsilon \searrow 0} \lambda(\{t - \varepsilon \leq r \leq t + \varepsilon\}) = 0$$

grâce au théorème de convergence monotone, ce qui conclut la preuve. □

1.4.2 Preuve du Théorème 1.2.2

Résultats préliminaires

On pose

$$\Omega_{n,c} = \Big\{ \sqrt{nh^d} \sup_{\Lambda} |r_n - r| \geq c\sqrt{\log n} \Big\}.$$

Lemme 1.4.1 *Sous les hypothèses* **H1** *et* **H3**, *si* $nh^{d+4}/\log n \to 0$, *alors il existe* $\Gamma > 0$ *tel que*

$$\sqrt{nh^d}\mathbb{P}(\Omega_{n,\Gamma}) \to 0.$$

Preuve du Lemme 1.4.1

Notons tout d'abord que comme r est continue, $\sup_\Lambda |r| < c$. Si on suppose que $\inf_\Lambda f > 0$, alors, puisque $\sup_\Lambda |f_n - f| \to 0$ p.s. sous le hypothèses du Lemme 1.4.1 (Bosq et Lecoutre [2]), il existe $\theta > 0$ tel que $\inf_\Lambda f_n > \theta$ p.s. pour n assez grand. On peut donc écrire

$$\begin{aligned} \sup_{\Lambda} |r_n - r| &= \sup_{\Lambda} \left| \frac{\varphi_n - \varphi}{f_n} + r\frac{f_n - f}{f_n} \right| \\ &\leq c\Big(\sup_{\Lambda} |\varphi_n - \varphi| + \sup_{\Lambda} |f_n - f| \Big). \end{aligned}$$

Dès lors, il suffit de montrer qu'il existe $c > 0$ tel que

$$\sqrt{nh^d}\mathbb{P}\left(\sup_{\Lambda} |\varphi_n - \varphi| \geq c\sqrt{\frac{\log n}{nh^d}} \right) \to 0,$$

pour obtenir le lemme (la démonstration étant identique pour $\sup_\Lambda |f_n - f|$). Comme

$$\sup_{\Lambda} |\varphi_n - \varphi| \leq \sup_{\Lambda} |\varphi_n - \mathbb{E}\,\varphi_n| + \sup_{\Lambda} |\mathbb{E}\,\varphi_n - \varphi|,$$

il suffit de montrer que

$$\sqrt{nh^d}\mathbb{P}\left(\sup_{\Lambda} |\varphi_n - \mathbb{E}\,\varphi_n| \geq \frac{c}{2}\sqrt{\frac{\log n}{nh^d}} \right) \to 0, \tag{1.2}$$

et

$$\sqrt{nh^d}\mathbb{P}\left(\sup_{\Lambda} |\mathbb{E}\,\varphi_n - \varphi| \geq \frac{c}{2}\sqrt{\frac{\log n}{nh^d}} \right) \to 0. \tag{1.3}$$

1.4 Preuves

La preuve de (1.3) est un exercice classique dès lors que le noyau K utilisé est pair et que $nh^{d+4}/\log n \to 0$, ce qui est le cas sous les hypothèses du Lemme 1.4.1. On va donc s'attacher à montrer (1.2).

On commence par recouvrir Λ par ℓ_n boules $B_k = B(x_k, \rho_n)$ $(k = 1, \ldots, \ell_n)$ de rayon ρ_n.

Pour $x \in \Lambda$ fixé, on considère B_k la boule contenant x. On pose alors, pour $x, x' \in \Lambda$,

$$A_n(x, x') = \frac{1}{n}\sum_{i=1}^{n} Y_i \left[K_h(x - X_i) - K_h(x' - X_i)\right] \\ -\mathbb{E}\,\frac{1}{n}\sum_{i=1}^{n} Y_i \left[K_h(x - X_i) - K_h(x' - X_i)\right],$$

et on obtient

$$\sup_{\Lambda} |\varphi_n - \mathbb{E}\varphi_n| \leq \sup_{1\leq k\leq \ell_n} |\varphi_n(x_k) - \mathbb{E}\varphi_n(x_k)| + \sup_{x\in\Lambda} |A_n(x, x_k)|. \tag{1.4}$$

Réglons tout d'abord le cas du second terme du membre de droite de (1.4). Comme K est supposé Lipschitzien, il existe $\gamma > 0$ tel que

$$\begin{aligned} |A_n(x, x_k)| &\leq ch^{-d-\gamma}\rho_n^{\gamma}\left(\frac{1}{n}\sum_{i=1}^{n}|Y_i| + \mathbb{E}|Y|\right) \\ &\leq ch^{-d-\gamma}\rho_n^{\gamma} \quad \text{car } Y \text{ est bornée.} \end{aligned}$$

Par conséquent

$$\mathbb{P}\left(\sup_{x\in\Lambda}|A_n(x, x_k)| > \frac{c}{4}\sqrt{\frac{\log n}{nh^d}}\right) \leq \mathbb{P}\left(ch^{-d-\gamma}\rho_n^{\gamma} > \frac{c}{4}\sqrt{\frac{\log n}{nh^d}}\right).$$

On peut toujours choisir

$$\rho_n = n^{-a}, a > 0 \quad \text{et} \quad \rho_n^{\gamma} = o\left(h^{d+\gamma}\sqrt{\frac{\log n}{nh^d}}\right),$$

de sorte que $\mathbb{P}\left(\sup_{x\in\Lambda}|A_n(x,x_k)|>\log n/\sqrt{nh^d}\right)=0$, ce qui règle le cas du deuxième terme du membre de droite de (1.4).

Pour finir, il faut traiter le premier terme : $|\varphi_n(x_k)-\mathbb{E}\varphi_n(x_k)|$. En utilisant les arguments de la preuve du Théorème 5.II.3 dans [2], on obtient

$$\forall\varepsilon>0,\ \mathbb{P}\left(\sup_{1\le k\le n}|\varphi_n(x_k)-\mathbb{E}\varphi_n(x_k)|>\varepsilon\right)<2\ell_n e^{-\frac{nh^d\varepsilon^2}{c}}.$$

Si on pose $\varepsilon=\varepsilon_0\sqrt{\log n/nh^d}$, on obtient alors

$$\begin{aligned}\mathbb{P}\left(\sup_{1\le k\le n}|\varphi_n(x_k)-\mathbb{E}\varphi_n(x_k)|>\varepsilon_0\sqrt{\frac{\log n}{nh^d}}\right) &\le c\ell_n n^{-2\varepsilon_0/c}\\ &\le cn^{-2\varepsilon_0/c}\rho_n^{-d}.\end{aligned}$$

En rappelant que $\rho_n=n^{-a}$, avec $a>0$, on obtient

$$\sqrt{nh^d}\mathbb{P}\left(\sup_{1\le k\le n}|\varphi_n(x_k)-\mathbb{E}\varphi_n(x_k)|>\varepsilon_0\sqrt{\frac{\log n}{nh^d}}\right)\le cn^{-2\varepsilon_0/c+1/2+ad}\sqrt{h^d}.$$

On peut toujours choisir $\varepsilon_0>\frac{(1/2+ad)c}{2}$, de sorte que le membre de droite tende vers 0. Il suffit alors de se rappeler (1.4) pour finir de démontrer (1.2). Par conséquent, on peut toujours choisir $\Gamma>0$ tel que

$$\sqrt{nh^d}\mathbb{P}(\Omega_{n,\Gamma})\to 0,$$

ce qui achève la preuve du lemme. □

Soit $t\in\Theta$. Pour tout $x\in\Lambda$, on définit

$$V_n(x,t)=\mathrm{Var}((Y-t)K_h(x-X))\ \text{ et }\ \widetilde{\mathbb{E}}\,r_n(x)=\mathbb{E}\,\varphi_n(x)/\mathbb{E}\,f_n(x).$$

Pour tout $x\in\Lambda$ tel que $V_n(x,t)\neq 0$, on pose

$$t_n(x)=\mathbb{E}\,f_n(x)\sqrt{\frac{nh^{2d}}{V_n(x,t)}}(t-\widetilde{\mathbb{E}}r_n(x)).$$

Par ailleurs, on considère les ensembles

$$\mathcal{V}_n^t=r^{-1}[t,t+\Gamma\sqrt{\log n/nh^d}]\cap\Lambda\ \text{ et }\ \bar{\mathcal{V}}_n^t=r^{-1}[t-\Gamma\sqrt{\log n/nh^d},t]\cap\Lambda.$$

Notons enfin Φ la fonction de répartition de la loi $\mathcal{N}(0,1)$, et $\overline{\Phi}(x)=1-\Phi(x)$.

Lemme 1.4.2 *Supposons que les hypothèses* **H1** *et* **H3** *sont satisfaites. Alors, il existe $c > 0$ telle que pour tout $n \geq 1$, $t \in \mathbb{R}$ et $x \in \Lambda$:*

$$|\mathbb{P}(r_n(x) \leq t) - \Phi(t_n(x))| \leq \frac{c}{\sqrt{nh^d}}.$$

Preuve du Lemme 1.4.2

Posons, pour $i = 1, \ldots, n$,

$$Z_i(x,t) = (Y_i - t)K_h(x - X_i)\ ,\ Z(x,t) = (Y - t)K_h(x - X).$$

On a alors $V_n(x,t) = \mathrm{Var}(Z(x,t))$. Nous allons à présent réécrire $\mathbb{P}(r_n(x) < t)$ en fonction de $V_n(x,t)$, $Z_i(x,t)$ et $\mathbb{E}\, Z(x,t)$ de manière à pouvoir utiliser l'inégalité de Berry-Esséen :

$$\begin{aligned}
&\mathbb{P}(r_n(x) < t) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i(x,t) < 0\right) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\Big(Z_i(x,t) - \mathbb{E}\, Z(x,t)\Big) < -\mathbb{E}\, Z(x,t)\right) \\
&= \mathbb{P}\left(\sqrt{\frac{n}{V_n(x,t)}}\frac{1}{n}\sum_{i=1}^{n}\Big(Z_i(x,t) - \mathbb{E}\, Z(x,t)\Big) < -\sqrt{\frac{n}{V_n(x,t)}}\mathbb{E}\, Z(x,t)\right) \\
&= \mathbb{P}\left(\sqrt{\frac{n}{V_n(x,t)}}\frac{1}{n}\sum_{i=1}^{n}\Big(Z_i(x,t) - \mathbb{E}\, Z(x,t)\Big) < t_n(x)\right).
\end{aligned}$$

L'inégalité de Berry-Esséen nous permet alors d'écrire

$$|\mathbb{P}(r_n(x) < t) - \Phi(t_n(x))| \leq \frac{c}{\sqrt{nh^d}}\mathbb{E}\left|(Y - t)K_h(x - X) - \mathbb{E}(Y - t)K_h(x - X)\right|^3. \tag{1.5}$$

Sous les hypothèses **H1** et **H3**, un calcul facile montre que

$$\sup_{x\in\Lambda}\left|(t - Y)K_h(x - X) - \mathbb{E}(Y - t)K_h(x - X)\right|^3 \leq ch^d$$

et

$$\inf_{x\in\Lambda} V_n(x,t) \geq ch^d.$$

Le lemme découle alors directement de (1.5). □

On définit à présent Θ_0 l'ensemble des $t \in \Theta$ tels que

$$\lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon} \lambda\Big(r^{-1}[t-\varepsilon, t]\Big) = \lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon} \lambda\Big(r^{-1}[t, t+\varepsilon]\Big) = \int_{\partial \mathcal{L}(t)} \|\nabla r\|^{-1} d\mathcal{H}.$$

Le résultat suivant est démontré dans l'article de Cadre [3] (Lemme 3.2).

Lemme 1.4.3 *Supposons que* **H1** *et* **H2** *sont satisfaites. Alors* $\Theta_0 = \Theta$ *presque partout.*

Notons que sous les hypothèses **H1** et **H2**, on obtient, grâce à la Proposition A.2 dans [3],

$$\lambda\Big(r^{-1}[t-\varepsilon, t+\varepsilon]\Big) = \lambda\Big(r^{-1}(t-\varepsilon, t+\varepsilon)\Big),$$

pour tout $t \in \Theta$ et $\varepsilon > 0$ suffisamment petit.

Posons finalement

$$v(x) = \mathrm{Var}(Y|X=x) + r^2(x),$$

et, pour $t \in \Theta$ et $x \in \Lambda$,

$$\bar{t}_n(x) = f(x) \sqrt{\frac{nh^d}{\widetilde{K} f(x)(v(x)+t^2)}} (t - r(x)).$$

Nous sommes alors en mesure d'énoncer et de démontrer le Lemme 1.4.4 ci dessous.

Lemme 1.4.4 *Supposons que* **H1** *et* **H2** *sont satisfaites. Si* $nh^d/(\log n)^7 \to \infty$ *et* $nh^{d+4} \log n \to 0$*, alors pour tout* $t \in \Theta_0$*,*

$$\lim_{n\to\infty} \sqrt{nh^d} \left[\int_{\mathcal{V}_n^t} \mathbb{P}(r_n(x) < t) dx - \int_{\mathcal{V}_n^t} \Phi(\bar{t}_n(x)) dx \right] = 0,$$

et

$$\lim_{n\to\infty} \sqrt{nh^d} \left[\int_{\overline{\mathcal{V}}_n^t} \mathbb{P}(r_n(x) > t) dx - \int_{\overline{\mathcal{V}}_n^t} \overline{\Phi}(\bar{t}_n(x)) dx \right] = 0.$$

Preuve du Lemme 1.4.4 Les démonstrations étant identiques, nous ne traiterons que la première équation.

Le Lemme 1.4.2 nous permet d'écrire

$$\sqrt{nh^d}\left[\int_{\mathcal{V}_n^t}\mathbb{P}(r_n(x)<t)dx-\int_{\mathcal{V}_n^t}\Phi(t_n(x))dx\right]\leq c\lambda(\mathcal{V}_n^t).$$

Comme $\lambda(r^{-1}[t-\varepsilon,t+\varepsilon])\to 0$, on a $\lambda(\mathcal{V}_n^t)\to 0$. Il suffit donc de montrer que

$$E_n=\sqrt{nh^d}\int_{\mathcal{V}_n^t}|\Phi(t_n(x))dx-\Phi(\bar{t}_n(x))dx|\to 0.$$

En utilisant le fait que Φ soit Lipschitzienne on obtient

$$E_n\leq c\sqrt{nh^d}\lambda(\mathcal{V}_n^t)\sup_{\mathcal{V}_n^t}|t_n-\bar{t}_n|. \tag{1.6}$$

Par définition de $t_n(x)$ et $\bar{t}_n(x)$, on a, pour tout $x\in\mathcal{V}_n^t$,

$$\begin{aligned}
&\frac{1}{\sqrt{nh^d}}|t_n(x)-\bar{t}_n(x)|\\
&\leq|t-r(x)|\left|\frac{f(x)}{\sqrt{\widetilde{K}f(x)(v(x)+t^2)}}-\frac{\mathbb{E}\,f_n(x)}{\sqrt{V_n(x,t)h^{-d}}}\right|\\
&\qquad+\sqrt{\frac{h^d}{V_n(x,t)}}\mathbb{E}\,f_n(x)|r(x)-\widetilde{\mathbb{E}}\,r_n(x)|\\
&\leq\sqrt{\frac{\log n}{nh^d}}\left|\sqrt{\frac{|f(x)V_n(x,t)h^{-d}-(\mathbb{E}\,f_n(x))^2\,\widetilde{K}(v(x)+t^2)|}{\widetilde{K}(v(x)+t^2)V_n(x,t)h^{-d}}}\right|\\
&\qquad+\sqrt{\frac{h^d}{V_n(x,t)}}\mathbb{E}\,f_n(x)|r(x)-\widetilde{\mathbb{E}}\,r_n(x)|
\end{aligned}\tag{1.7}$$

Comme $\mathcal{V}_n^t$ est contenu dans Λ, un calcul facile permet de déduire de **H1** et **H3** que

$$\sup_{\mathcal{V}_n^t}|\widetilde{\mathbb{E}}\,r_n-r|\leq ch^2. \tag{1.8}$$

De plus, si on pose

$$V_n^1(x)=\mathrm{Var}\;K_h(x-X),\quad V_n^2=\mathrm{Var}\;Y\,K_h(x-X),$$

on peut alors écrire

$$\begin{aligned}
&|f(x)V_n(x,t)h^{-d} - (\mathbb{E}\, f_n(x))^2\, \widetilde{K}(v(x)+t^2)| \\
&\quad \le |f(x)|\left|V_n(x,t)h^{-d} - \widetilde{K}\mathbb{E}\, f_n(x)(v(x)+t^2)\right| + c|f(x) - \mathbb{E}\, f_n(x)| \\
&\quad \le |f(x)|\left|V_n(x,t)h^{-d} - \widetilde{K}f(x)(v(x)+t^2)\right| + c|f(x) - \mathbb{E}\, f_n(x)| \\
&\quad \le |f(x)|\Big(t^2|V_n^1(x)h^{-d} - \widetilde{K}f(x)| + |V_n^2(x)h^{-d} - \widetilde{K}f(x)v(x)| \\
&\qquad\quad +2t\,|\mathrm{Cov}\,(YK_h(x-X), K_h(x-X))|\Big) + c|f(x) - \mathbb{E}\, f_n(x)| \\
&\quad \le c\Big(|V_n^1(x)h^{-d} - \widetilde{K}f(x)| + |V_n^2(x)h^{-d} - \widetilde{K}f(x)v(x)| \\
&\qquad\quad + |\mathrm{Cov}\,(YK_h(x-X), K_h(x-X))| + |f(x) - \mathbb{E}\, f_n(x)|\Big).
\end{aligned}$$

A nouveau, comme $\mathcal{V}_n^t$ est contenu dans Λ, un calcul facile permet de déduire de **H1** et **H3** que

$$\sup_{x\in\mathcal{V}_n^t} |f(x)V_n(x,t)h^{-d} - (\mathbb{E}\, f_n(x))^2\, \widetilde{K}(v(x)+t^2)| \le ch. \tag{1.9}$$

On déduit de (1.7), (1.8) et (1.9) que

$$\sup_{x\in\mathcal{V}_n^t} |t_n(x) - \bar{t}_n(x)| \le c\Big(\sqrt{h\log n} + \sqrt{nh^{k+4}}\Big).$$

Ainsi, par (1.6) et puisque $t \in \Theta_0$, on a que pour n assez grand

$$E_n \le c\sqrt{\log n}\Big(\sqrt{h\log n} + \sqrt{nh^{k+4}}\Big),$$

et ce dernier terme tend vers 0 grâce aux hypothèses sur h, d'où le lemme. □

Preuve du Théorème 1.2.2

Notons tout d'abord que

$$\mathbb{E}\,\lambda\Big(\mathcal{L}_n(t)\Delta\mathcal{L}(t)\Big) = \int_{\Lambda\cap\{r\ge t\}} \mathbb{P}(r_n(x) < t)dx + \int_{\Lambda\cap\{r<t\}} \mathbb{P}(r_n(x) \ge t)dx.$$

Posons à présent

$$\mathbb{P}_{n,t}(x) = \mathbb{P}(r_n(x) < t),$$

et rappelons que

$$\mathcal{V}_n^t = r^{-1}[t, t + \Gamma\sqrt{\log n/nh^d}] \cap \Lambda \text{ et } \bar{\mathcal{V}}_n^t = r^{-1}[t - \Gamma\sqrt{\log n/nh^d}, t] \cap \Lambda.$$

Comme $\sqrt{nh^d}\mathbb{P}\Big(\sqrt{nh^d}\sup_{\Lambda}|r_n - r| \geq \Gamma\sqrt{\log n}\Big) \to 0$ d'après le Lemme 1.4.1, il suffit de montrer qu'il existe deux constantes strictement positives C_1, C_2 telles que

$$\min\left(\varliminf_{n\to\infty}\sqrt{nh^d}\int_{\mathcal{V}_n^t}\mathbb{P}_{n,t}(x)dx, \varliminf_{n\to\infty}\sqrt{nh^d}\int_{\bar{\mathcal{V}}_n^t}\mathbb{P}_{n,t}(x)dx\right) \geq C_1\sqrt{\frac{2t}{\pi}\widetilde{K}}\int_{\partial\mathcal{L}(t)}\frac{d\mathcal{H}}{\|\nabla r\|}$$

et

$$\max\left(\varlimsup_{n\to\infty}\sqrt{nh^d}\int_{\mathcal{V}_n^t}\mathbb{P}_{n,t}(x)dx, \varlimsup_{n\to\infty}\sqrt{nh^d}\int_{\bar{\mathcal{V}}_n^t}\mathbb{P}_{n,t}(x)dx\right) \leq C_2\sqrt{\frac{2t}{\pi}\widetilde{K}}\int_{\partial\mathcal{L}(t)}\frac{d\mathcal{H}}{\|\nabla r\|}$$

De plus, le Lemme 1.4.3 nous apprend qu'il suffit de montrer le Théorème 1.2.2 pour tout $t \in \Theta_0$. Soient $t \in \Theta_0$ et

$$I_n = \int_{\mathcal{V}_n^t}\Phi(\bar{t}_n(x))dx, \qquad \overline{I}_n = \int_{\bar{\mathcal{V}}_n^t}\overline{\Phi}(\bar{t}_n(x))dx$$

Grâce au Lemme 1.4.4, il suffit de montrer qu'il existe deux constantes strictement positives C_1 et C_2 telles que

$$C_1\sqrt{\frac{t}{2\pi}\widetilde{K}}\int_{\partial\mathcal{L}(t)}\frac{d\mathcal{H}}{\|\nabla r\|} \leq \varliminf_{n\to\infty}\sqrt{nh^d}I_n \leq \varlimsup_{n\to\infty}\sqrt{nh^d}I_n \leq C_2\sqrt{\frac{t}{2\pi}\widetilde{K}}\int_{\partial\mathcal{L}(t)}\frac{d\mathcal{H}}{\|\nabla r\|},$$

et

$$C_1\sqrt{\frac{t}{2\pi}\widetilde{K}}\int_{\partial\mathcal{L}(t)}\frac{d\mathcal{H}}{\|\nabla r\|} \leq \varlimsup_{n\to\infty}\sqrt{nh^d}\overline{I}_n \leq \varliminf_{n\to\infty}\sqrt{nh^d}\overline{I}_n \leq C_2\sqrt{\frac{t}{2\pi}\widetilde{K}}\int_{\partial\mathcal{L}(t)}\frac{d\mathcal{H}}{\|\nabla r\|}.$$

On ne montrera que le premier encadrement. Pour cela on écrit

$$I_n = \frac{1}{\sqrt{2\pi\widetilde{K}}}\int_{\mathcal{V}_n^t}\int_{-\infty}^{b_n(x)}\exp\left(\frac{-u^2}{2\widetilde{K}}\right)du\,dx$$

où $b_n(x) = \sqrt{f(x)nh^d}(t - r(x))/\sqrt{v(x) + t^2}$.
Par ailleurs,

$$b_n(x) = \sqrt{\frac{|\varphi(x)|}{v(x) + t^2}}b_n'(x),$$

avec $b'_n(x) = \sqrt{nh^d}(t - r(x))/\sqrt{|r(x)|}$. On peut alors trouver deux constantes strictement positives C_1 et C_2 telles que

$$C_1 b'_n(x) \leq b_n(x) \leq C_2 b'_n(x),$$

ce qui nous permet d'obtenir

$$I_n \geq \frac{C_1}{\sqrt{2\pi\widetilde{K}}} \int_{\mathcal{V}_n^t} \int_{-\infty}^{b'_n(x)} \exp\left(\frac{-C_1^2 u^2}{2\widetilde{K}}\right) du\, dx,$$

et

$$I_n \leq \frac{C_2}{\sqrt{2\pi\widetilde{K}}} \int_{\mathcal{V}_n^t} \int_{-\infty}^{b'_n(x)} \exp\left(\frac{-C_2^2 u^2}{2\widetilde{K}}\right) du\, dx.$$

Il suffit alors de reprendre les arguments de la preuve de la Proposition 3.1 dans [3] pour conclure que

$$\frac{\sqrt{t\widetilde{K}}}{C_1\sqrt{2\pi}} \int_{\partial\mathcal{L}(t)} \frac{1}{\|\nabla r\|} \partial\mathcal{H} \leq \underline{\lim}_{n\to\infty} \sqrt{nh^d} I_n \leq \overline{\lim}_{n\to\infty} \sqrt{nh^d} I_n \leq \frac{\sqrt{t\widetilde{K}}}{C_2\sqrt{2\pi}} \int_{\partial\mathcal{L}(t)} \frac{1}{\|\nabla r\|} \partial\mathcal{H},$$

d'où le théorème. □

Bibliographie

[1] G. Biau, B. Cadre, and B. Pelletier. A graph-based estimator of the number of clusters. *ESAIM. Probability and Statistics*, 11 :272–280, 2007.

[2] D. Bosq and J. P. Lecoutre. *Théorie de l'Estimation Fonctionnelle*. Ecole Nationale de la Statistique et de l'Administration Economique et Centre d'Etudes des Programmes Economiques. Economica, 1987.

[3] B. Cadre. Kernel estimation of density level sets. *Journal of Multivariate Analysis*, 97(4) :999–1023, 2006.

[4] L. Cavalier. Nonparametric estimation of regression level sets. *Statistics*, 29(2) :131–160, 1997.

[5] A. Cuevas, M. Febrero, and R. Fraiman. Estimating the number of clusters. *The Canadian Journal of Statistics*, 28(1) :367–382, 2000.

[6] A. Cuevas, M. Febrero, and R. Fraiman. Cluster analysis : a further approach based on density estimation. *Computational Statistics and Data Analysis*, 36(4) :441–459, 2001.

[7] A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, 25(6) :2300–2312, 1997.

[8] A. Cuevas, W. González-Manteiga, and A. Rodríguez-Casal. Plug-in estimation of general level sets. *Australian and New Zealand Journal of Statistics*, 48(1) :7–19, 2006.

[9] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. Studies in Advanced Mathematics. CRC Press, 2000.

[10] J. A. Hartigan. Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association*, 82(397) :267–270, 1987.

[11] D. Müller. The excess mass approach in statistics. *Beiträge zur Statistik, Universität Heidelberg*, 3, 1993.

[12] D. W. Müller and G. Sawitzki. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86(415) :738–746, 1991.

[13] R. D. Nowak and R. M. Willett. Minimax optimal level-set estimation. *IEEE Transactions on Image Processing*, 16(12) :2965–2979, 2007.

[14] W. Polonik. Measuring mass concentrations and estimating density contour clusters–an excess mass approach. *The Annals of Statistics*, 23(3) :855–881, 1995.

[15] B. L. S. Prakasa Rao. *Nonparametric Functional Estimation*. Academic Press, Orlando, 1983.

[16] C. Scott and M. Davenport. Regression level set estimation via costsensitive classification. *IEEE Transaction on Signal Processing*, 55 :2752–2757, 2007.

[17] A. B. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3) :948–969, 1997.

[18] G. Walther. Granulometric smoothing. *The Annals of Statistics*, 25(6) :2273–2299, 1997.

Conclusion et perspectives

Dans ce travail de thèse, nous avons apporté une contribution supplémentaire aux thèmes de l'apprentissage statistique et de la statistique fonctionnelle. Dans la mesure du possible, nous nous sommes attachés à étudier à la fois les aspects théoriques des méthodes présentées et leurs utilisations pratiques.

Dans la première partie de ce travail, nous avons adapté au problème de la régression fonctionnelle la méthode des k plus proches voisins, utilisée par Biau, Bunea et Wegkamp (2005) dans le contexte de la classification. Les résultats obtenus dans ce chapitre ont montré la pertinence d'une réduction préalable de la dimension des observations par le biais d'une projection. S'agissant de la recherche future, il nous semble important de réaliser une étude comparative des performances de la méthode en fonction des bases choisies pour la projection des observations.

Dans le premier chapitre de la deuxième partie, nous avons adapté aux données fonctionnelles le principe du clustering par quantification utilisé par Linder (2002) dans un contexte fini-dimensionnel. Pour des raisons de robustesse nous avons choisi d'adopter un critère d'erreur de type L_1. Nous nous sommes attachés à développer un algorithme qui soit à la fois fiable et utilisable en pratique. Une prochaine étape importante sera de développer une procédure automatique pour choisir le nombre de clusters. Dans un second chapitre, nous avons présenté le résultat d'une collaboration avec M. Gerlotto, chercheur de l'Institut pour la Recherche et le Développement (IRD). L'utilisation des méthodes présentées dans le premier chapitre pour discrimi-

ner des bancs d'anchois a ainsi permis de construire un nouvel indicateur facilement utilisable pour repérer les bancs victimes de l'attaque d'un prédateur.

La dernière partie de cette thèse a apporté des éléments de réponse au problème de l'estimation des ensembles de niveau de la fonction de régression, grâce à l'utilisation de la méthode plug-in présentée par Cadre (2006) pour la fonction de densité. Cette méthode nous a permis de construire un estimateur facilement calculable, sous des hypothèses peu restrictives. Une étude sur des échantillons de données simulées a confirmé la validité pratique de notre méthode. Ce travail offre un certain nombre de perspectives intéressantes. En particulier, il serait intéressant d'utiliser une méthode plug-in avec l'estimateur de la fonction de régression présenté dans la première partie de cette thèse, afin de considérer le cas de données fonctionnelles.

Pour finir, nous pensons que la prise en compte de la nature fonctionnelle des données apporte un nouvel éclairage pertinent sur un grand nombre de problématiques actuelles, dont l'étude du comportement de populations où d'individus est un exemple probant. C'est pourquoi nous envisageons de poursuivre nos travaux dans cette voie, en nous attachant à toujours associer les méthodes théoriques à leurs utilisations pratiques.

Annexe A

L_1-quantization and clustering in Banach spaces

Nous donnons dans cette annexe l'article en anglais correspondant au premier chapitre de la deuxième partie.

A.1 Introduction

Clustering consists in partitioning a data set into subsets (or clusters), so that the data in each subset share some common trait. Proximity is determined according to some distance measure. For a thorough introduction of the subject, we refer to the book by Kaufman and Rousseeuw [14]. The origin of clustering goes back to 45 years ago, when some biologists and sociologists began to search for automatics methods to build different groups with their data. Today, clustering is used in many fields. For example, in medical imaging, it can be used to differentiate between different types of tissue and blood in a three dimensional image. Market researchers use it to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers. There are also many different applications in artificial intelligence, sociology, medical research, or political sciences.

In the present paper, the clustering method we investigate lays on the tech-

nique of quantization, commonly used in signal compression (Graf and Luschgy [12], Linder [17]). Given a normed space $(\mathcal{H}, \|.\|)$, a codebook (of size k) is defined by a subset $\mathcal{C} \subset \mathcal{H}$ with cardinality k. Then, each $x \in \mathcal{H}$ is represented by a unique $\hat{x} \in \mathcal{C}$ via the function q,

$$\begin{aligned} q: \ \mathcal{H} &\to \mathcal{C} \\ x &\to \hat{x}, \end{aligned}$$

which is called a quantizer. Here we come back to the clustering, as we create clusters in the data by regrouping the observations which have the same image by q.

Denote by d the distance induced by the norm on $\mathcal{H}$:

$$\begin{aligned} d : \mathcal{H} \times \mathcal{H} &\to \mathbb{R}^+ \\ (x, y) &\to \|x - y\|. \end{aligned}$$

In this paper, observations are modeled by a random variable X on $\mathcal{H}$, with distribution μ. The quality of the approximation of X by $q(X)$ is then given by the distortion $\mathbb{E}\, d\big(X, q(X)\big)$. Thus the aim is to minimize $\mathbb{E}\, d\big(X, q(X)\big)$ among all possible quantizers. However, in practice, the distribution μ of the observations is unknown, and we only have at hand n independent observations $X_1, \ldots, X_n$ with the same distribution than X. The goal is then to minimize the empirical distortion :

$$\frac{1}{n} \sum_{i=1}^{n} d(X_i, q(X_i)).$$

Since the early work of Hartigan [13] and Pollard [19, 20, 21], the performances of clustering have been considered by many authors. Convergence properties of the minimizer q_n^* of the empirical distortion have been mostly studied in the case when $\mathcal{H} = \mathbb{R}^d$. Consistency of q_n^* was shown by Pollard [19, 21] and Abaya and Wise [1]. Rates of convergence have been considered by Pollard [20], Linder, Lugosi, and Zeger [18], Linder [17].

A.1 Introduction

As a matter of facts, in many practical problems, input data items are in the form of random functions (speech recordings, spectra, images) rather than standard vectors, and this casts the clustering problem into the general class of functional data analysis. Even though in practice such observations are observed at discrete sampling points, the challenge in this context is to infer the data structure by exploiting the infinite-dimensional nature of the observations. The last few years have witnessed important developments in both the theory and practice of functional data analysis, and many traditional data analysis tools have been adapted to handle functional inputs. The book by Ramsay and Silverman [22] provides a comprehensive introduction to the area. Recently, Biau, Devroye, and Lugosi [2] gave some consistency results in Hilbert spaces and with a L_2-based distortion.

Thus, the first novelty in this paper is to consider data taking place in a separable and reflexive Banach space, with no restriction on their dimension. The second novelty is that we consider a L_1-based distortion, which leads to more robust estimators. For a discussion of the advantage of the L_1-distance we refer the reader to the paper by Kemperman [15].

This setup calls for substantially different arguments to prove results which are known to be true when considering finite dimensional spaces and a L_2-based distortion. In particular, specific notions will be required, such as weak topology (Dunford and Schwartz [10]), lower semi-continuity (Ekeland and Temam [10]) and entropy (Van der Vaart and Wellner [23]).

The document is organized as follows. We first provide the formal context of quantization in Banach space in the first part of Section 2. Then, we focus on the problem of the existence of an optimal quantizer. In Sections 3 and 4 we study two consistent estimators of this optimal quantizer, and we confront them to real-life data in Section 5. Proofs are collected in Appendix A.

A.2 Quantization in a Banach space

A.2.1 General framework

The fact that the closed bounded balls are not compact is a major problem when considering infinite dimensional spaces. To overcome this, the classical solution is to consider reflexive spaces, i.e., spaces in which the closed bounded balls are compact for the weak topology (Dunford and Schwartz [9]). Thus, throughout the document, $(\mathcal{H}, \|.\|)$ will denote a reflexive and separable Banach space. We let X be a $\mathcal{H}$-valued random variable with distribution μ such as $\mathbb{E}\|X\| < \infty$.

Given a set $\mathcal{C} = \{y_i\}_{i=1}^k$ of points in $\mathcal{H}^k$, any Borel application $q : \mathcal{H} \to \mathcal{C}$ is called a quantizer. The set $\mathcal{C}$ is called a codebook, and the y_i, $i = 1, \ldots, k$ are the centers of $\mathcal{C}$. The error made by replacing X by $q(X)$ is measured by the distortion :

$$D(\mu, q) = \mathbb{E}\, d(X, q(X)) = \int_{\mathcal{H}} \|x - q(x)\| \mu(dx).$$

Note that $D(\mu, q) < \infty$ since $\mathbb{E}\|X\| < \infty$. For a given k, the aim is to minimize $D(\mu, .)$ among the set $\mathcal{Q}_k$ of all possible k-quantizers. The optimal distortion is then defined by

$$D_k^*(\mu) = \inf_{q \in \mathcal{Q}_k} D(\mu, q).$$

When it exists, a quantizer q^* satisfying $D(\mu, q^*) = D_k^*(\mu)$ is said to be an optimal quantizer.

Any quantizer is characterized by its codebook $\mathcal{C} = \{y_i\}_{i=1}^k$ and a partition of $\mathcal{H}$ in cells $S_i = \{x \in \mathcal{H} : q(x) = y_i\},\ i = 1, \ldots, k$ via the rule

$$q(x) = y_i \iff x \in S_i.$$

Thus, from now on, we will define a quantizer by its codebook and its cells.

A.2 Quantization in a Banach space

Let us consider the particular family of Voronoi partitions, constructed by the nearest neighbors rule. That is, for each center of the codebook, a cell is constituted by the elements $x \in \mathcal{H}$ which are the closest to him (Gersho and Gray [11]). A quantizer with such a partition is named a nearest neighbor quantizer, and we denote by $\mathcal{Q}_{knn}$ the set of all k-nearest neighbor quantizers. It can be easily proven (see Lemma 1 in Linder [17]) that

$$\inf_{q \in \mathcal{Q}_k} D(\mu, q) = \inf_{q \in \mathcal{Q}_{knn}} D(\mu, q).$$

More precisely, given two quantizers $q \in \mathcal{Q}_k$ and $q' \in \mathcal{Q}_{knn}$ with the same codebook, we have

$$D(\mu, q') \leq D(\mu, q).$$

Therefore, in the following, we will restrict ourselves to nearest neighbor quantizers.

A complementary result (see Lemma 2 in Linder [17]) is that for a quantizer q with codebook $\mathcal{C}$ and partition S, a quantizer q' with the same partition but with a codebook defined by

$$y'_i \in \arg\min_{y \in \mathcal{H}} \mathbb{E}\left[\|X - y\| \mid X \in S_i\right], \ i = 1, \ldots, k,$$

satisfies

$$D(\mu, q') \leq D(\mu, q).$$

From the two previous optimality results, on the codebook and associated partition, we can derive a simple algorithm in order to find a good quantizer. This algorithm is called the Lloyd algorithm and based on the so-called Lloyd iteration (Gersho and Gray [11], Chapter 6). The outline is as follows :

1. Choose randomly an initial codebook ;
2. Given a codebook C_m, build the associated Voronoi partition ;
3. Build C_{m+1}, the optimal codebook for the previous partition ;
4. Stop when the distortion no longer decreases.

Unfortunately, this algorithm has two drawbacks : it depends on the initial codebook chosen, and it does not necessarily converge to the optimal distortion. In Section 4 we will discuss an alternative to this algorithm, leading to an optimal quantizer.

A.2.2 Existence of an optimal quantizer

The aim of this section is to show that the minimization problem of $D(\mu, q)$ has at least one solution. Recall that we consider only nearest neighbor quantizers, which can be entirely characterized by their codebook $(y_1, \ldots, y_k)$, and set $\mathbf{y}_k = (y_1, \ldots, y_k)$.

We denote by

$$D(\mu, q) = D(\mu, \mathbf{y}_k)$$

the associated distortion. Therefore our first task is to prove that the function $D(\mu, .)$ has at least one minimum, or, in other words, that there exists at least one optimal codebook.

Theorem A.2.1 *Assume that $\mathcal{H}$ is a reflexive and separable Banach space. Then, the function $D(\mu, .)$ admits at least one minimum.*

Theoretically speaking, it is of interest to search for an optimal quantizer. To make the link with clustering, Theorem A.2.1 states that there exists at least one optimal repartition of the space $\mathcal{H}$ in different clusters. The next step is to consider the statistical case, in which the distribution of X is unknown.

A.3 A consistent estimator

A.3.1 Construction and consistency

In a statistical context, the distribution μ of X is unknown and we only have at hand n random variables, $X_1, \ldots, X_n$, independent and distributed as X.

A.3 A consistent estimator

Let the empirical measure μ_n be defined as

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[X_i \in A]},$$

for any measurable set $A \subset \mathcal{H}$. For any quantizer q, the associated empirical distortion is then given by

$$D(\mu_n, q) = \frac{1}{n}\sum_{i=1}^{n} \|X_i - q(X_i)\|.$$

An (empirical) quantizer $q_n^* = q_n^*(., X_1, \ldots, X_n)$ satisfying

$$q_n^* \in \arg\min_{q \in \mathcal{Q}_k} \sum_{i=1}^{n} \|X_i - q(X_i)\|$$

is said to be empirically optimal. In particular, if we set (with a slight abuse of notation)

$$D(\mu, q_n^*) = \mathbb{E}\left[\|X - q_n^*(X)\| \mid X_1 \ldots, X_n\right],$$

we have

$$D(\mu_n, q_n^*) = D_k^*(\mu_n).$$

From Theorem A.2.1, we know that for every n, an empirically optimal quantizer always exists.

The following theorem, which is an adaptation of Theorem 2 in Linder [17], establishes the asymptotic optimality of the quantizer q_n^* with respect to the distortion.

Theorem A.3.1 *Assume that $\mathcal{H}$ is a reflexive and separable Banach space, and set $k \geq 1$. Then, any sequence of empirically optimal k-quantizers $(q_n^*)_{n\geq 1}$ satisfies*

$$\lim_{n\to\infty} D(\mu, q_n^*) = D_k^*(\mu) \qquad a.s.$$

and

$$\lim_{n\to\infty} D(\mu_n, q_n^*) = D_k^*(\mu) \qquad a.s.$$

A.3.2 Rate of convergence

Most results in the literature concern the situation when $\mathcal{H} = \mathbb{R}^d$ and the distortion is a L_2-based one (Pollard [20], Linder [17], Linder, Lugosi, and Zeger [18]). For example, it is shown in [17] that if there exists $T > 0$ such that $\mathbb{P}[\|X\| \leq T] = 1$, then

$$\mathbb{E}\, D(\mu, q_n^*) - D^*(\mu) \leq CT^2 \sqrt{\frac{k(d+1)\ln(k(d+1))}{n}},$$

where $C > 0$ is a universal constant.

Recently, Biau, Devroye, and Lugosi [2] proved that when $\mathcal{H}$ is an Hilbert space, and the distortion is a L_2-based one, then

$$\mathbb{E}\, D(\mu, q_n^*) - D^*(\mu) \leq C\frac{k}{\sqrt{n}},$$

where $C > 0$ is a universal constant.

In the sequel, our goal is to establish a rate of convergence in a Banach space and with a L_1-criterion. This will require some new notions.

Let $\mathcal{P}(\mathcal{H})$ be the set of all probability measures on $\mathcal{H}$.

Définition A.3.1 *Let $p \in [1, \infty[$.*

1. *The L_p-Wasserstein distance between $\phi, \xi \in \mathcal{P}(\mathcal{H})$ is defined by :*

$$\rho_p(\phi, \xi) = \inf_{X \sim \phi, Y \sim \xi} \Big(\mathbb{E}\, d\big(X, Y\big)^p \Big)^{\frac{1}{p}}.$$

2. *A probability $\phi \in \mathcal{P}(\mathcal{H})$ satisfies a transportation inequality $T_p(\lambda)$ if there exists $\lambda > 0$ such that, for any probability $\xi \in \mathcal{P}(\mathcal{H})$,*

$$\rho_p^p(\phi, \xi) \leq \sqrt{\frac{2}{\lambda} H(\xi|\phi)},$$

where $H(\xi|\phi) = \displaystyle\int_{\mathcal{H}} \frac{d\xi}{d\phi} \log\left(\frac{d\xi}{d\phi}\right) d\phi$ is the Kullback information between ϕ and ξ.

Remarks :

- The L_p-Wasserstein distance, also called L_p-Kantorovich distance, is known to be appropriate for the quantization problem (Graf and Luschgy, Section 3 [12]) ;
- For this choice of distance, in view of getting rates of convergence, the so-called transportation inequalities, or Talagrand inequalities, are well designed (Ledoux [16]).

Generally speaking, it is a difficult task to determine whether a probability $\mu \in \mathcal{P}(\mathcal{H})$ satisfies a transportation inequality $T_p(\lambda)$. However, the problem is simpler when $p = 1$, as expressed in the theorem below proven in Djellout, Guillin, and Wu [7] (Theorem 2.3 and Section 1).

Theorem A.3.2 *A probability $\phi \in \mathcal{P}(\mathcal{H})$ satisfies a transportation inequality $T_1(\lambda)$ if and only if, for all $\alpha < \lambda/2$,*

$$\int_{\mathcal{H}} e^{\alpha\|x-y\|^2} d\mu(x) < \infty$$

for one (and therefore for all) y in $\mathcal{H}$.

In the sequel, we will only consider the case $p = 1$, and we set $\rho = \rho_1$. For any set $\Lambda \subset \mathcal{H}$, let $\mathcal{P}(\Lambda)$ be the set of all probability measures on Λ. Let also $\mathcal{N}(r, \Lambda)$ be the smallest number of balls of radius r (for the metric ρ) required to cover $\mathcal{P}(\Lambda)$, that is

$$\begin{aligned}&\mathcal{N}(r, \Lambda)\\ &\quad = \inf\left\{ n \in \mathbb{N} \text{ s.t. } \exists x_1, \ldots, x_n \in \mathcal{P}(\Lambda) : \textstyle\bigcup_{i=1}^n B_{\mathcal{P}(\Lambda)}(x_i, r) \supset \mathcal{P}(\Lambda) \right\},\end{aligned}$$

where $B_{\mathcal{P}(\Lambda)}(x_i, r)$ is the ball in $\mathcal{P}(\Lambda)$ centered at x_i and with radius r (for the metric ρ). The quantity $\ln(\mathcal{N}(r, \Lambda))$ is the entropy of $\mathcal{P}(\Lambda)$ (Van der Vaart and Wellner [23]).

In the same way, let $N(r, \Lambda)$ be the smallest number of balls of radius r required to cover Λ, with respect to the metric of $\mathcal{H}$.

In order to state a rate of convergence for $D(\mu, q_n^*)$, we introduce the following assumptions :

H1 : There exists $\lambda > 0$ such that μ satisfies a transportation inequality $T_1(\lambda)$;

H2 : Any closed bounded ball $B \subset \mathcal{H}$ is totally bounded. That is, for all $r > 0$, $N(r, B)$ is finite.

Note that **H1** is satisfied for paths of stochastic differential equations

$$dX_t = b(X_t)dt + s(X_t)dW_t,$$

where $t \in [0, T]$, $T < \infty$, and $b(.)$, $s(.)$ satisfy suitable properties (Djellout, Guillin and Wu [7], Corollary 4.1). **H2** is satisfied, for example, if $\mathcal{H}$ is a Sobolev space on a compact domain of $\mathbb{R}^d$ (Cucker and Smale [6], example 3).

From now on, B_R stands for the ball of center 0 and radius R in $\mathcal{H}$. According to assumption **H2** and Theorem A.1 in Bolley, Guillin, and Villani [4], there exists a positive constant C such that for all $r, R > 0$,

$$\mathcal{N}(r, B_R) \leq \left(\frac{CR}{r}\right)^{N(r/2, B_R)}. \tag{A.1}$$

Theorem A.3.3 *Assume that $\mathcal{H}$ is a reflexive and separable Banach space, and* **H1**, **H2** *are satisfied. Then, for all $\lambda' < \lambda$ and $\varepsilon > 0$, there exist three positive constants K, γ, and R_1 such that if $R = R_1 \max\left(1, \varepsilon^2, \ln\left(1/\varepsilon^2\right)\right)^{1/2}$ and $n \geq K \ln\left(\mathcal{N}(\gamma\varepsilon, B_R)\right)/\varepsilon^2$, we have :*

$$\mathbb{P}\left[\rho(\mu, \mu_n) \geq \varepsilon\right] \leq e^{-(\lambda'/2)n\varepsilon^2}.$$

Using the inequality

$$D(\mu, q_n^*) - D^*(\mu) \leq 2\rho(\mu, \mu_n),$$

we deduce the following corollary.

Corollary A.3.1 *Assume that $\mathcal{H}$ is a reflexive and separable Banach space, and* **H1**, **H2** *are satisfied. Then, for all $\lambda' < \lambda$ and $\varepsilon > 0$, there exist three positive constants K, γ, and R_1 such that if $R = R_1 \max\left(1, \varepsilon^2, \ln\left(1/\varepsilon^2\right)\right)^{1/2}$ and $n \geq K \ln\left(\mathcal{N}(\gamma\varepsilon, B_R)\right)/\varepsilon^2$, we have :*

$$\mathbb{P}\left[D(\mu, q_n^*) - D(\mu, q^*) \geq \varepsilon\right] \leq e^{-(\lambda'/8)n\varepsilon^2}.$$

Let $\mathcal{R}$ be the function from $\mathbb{R}_+^*$ to $\mathbb{R}_+^*$ defined by

$$\mathcal{R}(x) = R_1 \max\left(1, x^2, \ln\left(1/x^2\right)\right)^{1/2},$$

and denote $\mathcal{M}$ the function from $\mathbb{R}_+^*$ to $\mathbb{R}_+^*$ defined by

$$\mathcal{M}(x) = K \ln\left(\mathcal{N}(\gamma x, B_{\mathcal{R}(x))}\right)/x^2. \tag{A.2}$$

Theorem A.3.4 below gives us the desired rate of convergence.

Theorem A.3.4 *Assume that $\mathcal{H}$ is a reflexive and separable Banach space,* **H1**, **H2** *are satisfied, and $\mathcal{M}$ is invertible on some interval $]0, a]$. Then, there exists $C_0 > 0$ such that*

$$\mathbb{E}\, D\left(\mu, q_n^*\right) - D\left(\mu, q^*\right) \leq C_0 \max(\mathcal{M}^{-1}(n), n^{-1/2}).$$

Note there is no restriction on the support of μ. In particular, we do not require that the support of μ is bounded. This is an important point, since such an assumption is not verified, for example, by the distributions of classical diffusion processes, yet widely used in stochastic modeling.

Example : Suppose that assumptions **H1** and **H2** are satisfied. Consider the example 3 in Cucker and Smale [6], in which $\mathcal{H}$ is a Sobolev space on a

compact domain set of $\mathbb{R}^d$. Using the entropy of the balls $B_R \subset \mathcal{H}$ (Cucker and Smale [6]) and Theorem A.3.4, we have

$$\mathbb{E}\, D\left(\mu, q_n^*\right) - D\left(\mu, q^*\right) \leq \frac{C}{(\ln n)^{s/d}},$$

where C is a positive constant.

A.3.3 Algorithm

Calculating q_n^* appears to be a NP-complete problem. In order to approximate q_n^* one can adapt the Lloyd Lloyd algorithm, which has been presented in Section 2, to the statistical context in which we use μ_n instead of μ. Moreover, rather to calculate empirical medians in each cell, a possible solution is to consider medoids, i.e., centers taken within the sample $\{X_1, \ldots, X_n\}$. For more details about the Lloyd algorithm and medoids, we refer the reader to the book by Kaufman and Rousseeuw [14].

However, this Lloyd algorithm with medoids has the same drawbacks as the Lloyd algorithm presented in section 2 : non optimality and dependence on initial codebook. Thus, in the next section, we will present a new estimator, in order to overcome these drawbacks.

A.4 Minimization on data

A.4.1 Construction and Consistency

The basic idea of the estimator presented in this section consists in searching the minimum of the empirical distortion $D(\mu_n, .)$ within the sample $\{X_1, \ldots, X_n\}$. It is a generalization of a method of Cadre [5], who considered the case $k = 1$ only. Formally, our estimator $\mathbf{y}_{k,n}^* = (y_{1,n}^*, \ldots, y_{k,n}^*)$ is defined by

$$\mathbf{y}_{k,n}^* \in \underset{\mathbf{z} \in \{X_1, \ldots, X_n\}^k}{\arg\min}\ D(\mu_n, \mathbf{z}).$$

Note $\|.\|_k$ a norm on $\mathcal{H}^k$ (as an example, for $\mathbf{z} = (z_1, \ldots, z_k) \in \mathcal{H}^k$, $\|\mathbf{z}\|_k = \max_{i=1,\ldots,k} \|z_i\|$), and $B_{\mathcal{H}^k}(\mathbf{z}, r)$ the associated closed ball in $\mathcal{H}^k$ centered at $\mathbf{z}$ and with radius r.

Theorem A.4.1 *Assume that $\mathcal{H}$ is a reflexive and separable Banach space, and there exists $\mathbf{y}_k^*$ an optimal codebook for μ, which satisfies*

$$\forall \varepsilon > 0,\ \mathbb{P}\left[(X_1, \ldots, X_k) \in B_{\mathcal{H}^k}(\mathbf{y}_k^*, \varepsilon)\right] > 0. \tag{A.3}$$

Then,

$$\lim_{n\to\infty} D(\mu, \mathbf{y}_{k,n}^*) = D_k^*(\mu) \ a.s.$$

Remark : The condition (A.3) in Theorem A.4.1 simply requires that the probability that k observations fall in the neighborhood of $\mathbf{y}_k^*$ is not zero. The necessity of this condition is easy to understand. Indeed, suppose there exists $\varepsilon > 0$ such that for all optimal codebook $\mathbf{y}_k^*$ for μ, $(X_1, \ldots, X_k) \notin B_{\mathcal{H}^k}(\mathbf{y}_k^*, \varepsilon)$ with probability 1. Then, by construction, $D(\mu, \mathbf{y}_{k,n}^*)$ can not converge to $D_k^*(\mu)$.

Theorem A.4.2 *Assume that $\mathcal{H}$ is a reflexive and separable Banach space, and* (A.3)*,* **H1** *and* **H2** *hold. Then, we have*

$$\lim_{n\to\infty} \mathbb{E}D(\mu, \mathbf{y}_{k,n}^*) = D_k^*(\mu).$$

A.4.2 Rate of convergence

The next theorem states that $D(\mu, \mathbf{y}_{n,k}^*)$ converges to $D_k^*(\mu)$ at the same rate as $D(\mu, q_n^*)$. Remember that the function $\mathcal{M}$ is defined in (A.2), and let $\mathbf{y}_k^*$ be an optimal codebook for μ. For $\varepsilon > 0$ we set

$$f(\mathbf{y}_k^*, \varepsilon) = \mathbb{P}\big[(X_1, \ldots, X_k) \in B_{\mathcal{H}^k}(\mathbf{y}_k^*, \varepsilon)\big].$$

We also introduce the assumption :

H3 : There exist a decreasing function $V : \mathbb{N}^* \to \mathbb{R}_+^*$ and positive constants u, v, C such that

$$\max\left(\int_0^u (1 - f(\mathbf{y}_k^*, \varepsilon))^{\lfloor n/k \rfloor}\, d\varepsilon, \int_v^{+\infty} (1 - f(\mathbf{y}_k^*, \varepsilon))^{\lfloor n/k \rfloor}\, d\varepsilon\right) \leq V(n).$$

Theorem A.4.3 *Assume that $\mathcal{H}$ is a reflexive and separable Banach space, and* **H1** *and* **H2** *are satisfied. Let $\mathbf{y}_k^*$ be an optimal codebook for μ satisfying* **H3**. *Then, if $\mathcal{M}$ is invertible on some interval $]0, b]$ there exists a positive constant C_0 such that, for n large enough,*

$$\mathbb{E}D\left(\mu, \mathbf{y}_{k,n}^*\right) - D_k^*(\mu) \leq C_0 \max(\mathcal{M}^{-1}(n), V(n), \lfloor n/k \rfloor^{-1/2}).$$

Remarks :

- Assumption **H3** requires that the probability that data are present in a neighborhood of an optimal quantizer rises fast enough with n. It is an essential assumption in the proof of Theorem A.4.3.
- Assumption **H3** is satisfied if the following assumptions hold :

  H4 : There exists $c_1 > 0$ such that $f(\mathbf{y}_k^*, \varepsilon) \geq 1 - \exp(-\varepsilon^2)$ for $\varepsilon \in]0, c_1]$;

  H5 : There exists $c_2 > 0$ such that $f(\mathbf{y}_k^*, \varepsilon) \geq 1 - \exp(-\varepsilon^2)$ for $\varepsilon \in [c_2, +\infty[$.
- Assume that **H4** and **H5** are satisfied. Then we have

  $$\mathbb{E}D\left(\mu, \mathbf{y}_{k,n}^*\right) - D_k^*(\mu) \leq C_0 \max(\mathcal{M}^{-1}(n), \lfloor n/k \rfloor^{-1/2}).$$

  That is, $D(\mu, \mathbf{y}_{n,k}^*)$ converges to $D_k^*(\mu)$ at the same rate as $D(\mu, q_n^*)$.
- Assumption **H5** is satisfied if μ has a bounded support.

A.4.3 Algorithms

In order to calculate $\mathbf{y}_{k,n}^*$, we provide an algorithm which we will call Alter algorithm. The outline is the following :

1. List all possible codebooks, i.e., all possible k-tuple of data ;

2. Calculate the empirical distortion associated to the first codebook ;
3. For each successive codebook, calculate the associated empirical distortion. Each time a codebook has an associated empirical distortion smaller than the previous, store the codebook ;
4. Return the codebook which has the smallest distortion.

This algorithm overcomes the two drawbacks of the Lloyd algorithm : it does not depend on initial conditions and it converges to the optimal distortion. Unfortunately its complexity is $o(n^k)$ and it is impossible to use it for high values of n or k.

In order to overcome this complexity problem, we define the Alter-fast iteration, working as follows :

1. Select randomly $n_1 < n$ data in the whole data set (n_1 should be small) ;
2. Run the Alter algorithm on these n_1 data (empirical distortions should be calculated using the whole data set) ;
3. Store the obtained codebook.

Then we derive an accelerated version of the Alter algorithm, which we call Alter-fast algorithm. The outline is the following :

1. Run n_2 times the Alter-fast iteration (n_2 should be high) ;
2. Select, among all the obtained codebooks, the one which minimizes the associated empirical distortion (calculated using the whole data set).

The Alter-fast algorithm provide a usable alternative for the Alter algorithm, in the same way as the Lloyd algorithm using medoids was an alternative to the Lloyd algorithm. Its complexity is $o(n_2 \times n_1^k)$. We will see in the next section that the Alter-fast algorithm seems to perform almost as well as the Alter algorithm on real-life data.

A.5 Application : speech recognition

Here we use a part of the TIMIT database (http://www-stat.stanford.edu/~tibs/ElemStatLearn/). The data are log-periodograms corresponding to recording phonemes of 32 ms duration. We are interested in 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 "ao" as the first vowel in "water" (1022 items). The database is a multi speaker database. Each speaker is recorded at a 6 kHz sampling rate and we retain only the first 256 frequencies (see Figure A.1).

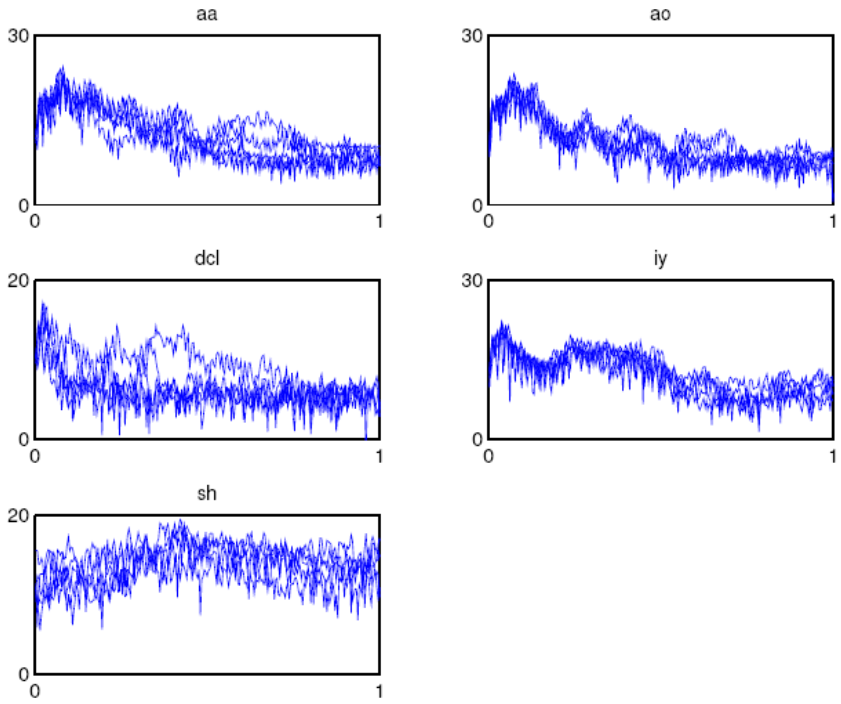

FIGURE A.1 – A sample of log-periodograms for fives phonemes.

Thus the data consist of 4509 series of length 256. We compare here the Lloyd and Alter-fast algorithms. We split the data into a learning and a

testing set. The quantizer is constructed using only the first set and its performance (i.e., the rate of good classification) is evaluated from the second one. We give the rates of good classification associated to the codebooks selected by the Lloyd and and Alter-fast algorithms in Table A.1. Recall that, for each center, a cluster includes the data which are closer to this center than to any other. Moreover we give the variance induced by the dependence on initial conditions : the initial codebook for the Lloyd algorithm, and the successive reduced data set for the Alter-fast algorithm. We note that the results of the Alter-fast algorithm are better than those of the Lloyd algorithm.

Algorithm	Rate of good classification
Lloyd	0.80 (var=0.0047)
Alter-fast	0.84 (var=0.00014)

TABLE A.1 – Rate of good classification with the five phonemes.

The phonemes “ao” and “aa” appear to be particularly difficult to classify. To illustrate this phenomenon, we confront the Lloyd, Alter, and Alter-fast algorithms on these two phonemes only. The rates of good classification are given in Table A.2 (note that we gave no variance for the Alter algorithm, since it does not depends on any initial condition). As expected, the results are not satisfactory. We note however that the Alter algorithm results are more reliable than the Lloyd algorithm ones, and that the rates of good classification obtained from the Alter and Alter-fast algorithms are almost equivalent. We also note that we improve over the results of Bleakley [3] (Chapter 2), who is using different SVM algorithms in a supervised learning context.

Finally, we provide a similar study by removing the phonemes “ao”from the database (see Table A.3). The results are significantly better than those obtained with the whole database.

Algorithm	Rate of good classification
Lloyd	0.64 (var=0.0031)
Alter	0.71
Alter-fast	0.68 (var=0.00015)
Max. bin. kernel [3]	0.61
Min. bin. kernel [3]	0.63

TABLE A.2 – Rate of good classification of phonemes “aa” and “ao”.

Algorithm	Rate of good classification
Lloyd	0.87 (var=0.0032)
Alter-fast	0.90 (var=0.0001)

TABLE A.3 – Rate of good classification without the phoneme “ao”.

A.6 Conclusion

This paper thus provided an answer to the problem of functional L_1-clustering : we first proved that for any measure $\mu \in \mathcal{P}(\mathcal{H})$ with finite moment, an optimal quantization always exists (Theorem A.2.1). Then we proposed a consistent estimator of q^* (Theorem A.3.1), and we state its rate of convergence (Theorem A.3.4). In order to offset the main drawbacks of the Lloyd algorithm, we then proposed the Alter algorithm and its accelerated version, the Alter-fast algorithm. Finally, a confrontation of our algorithms on real-life data states the practical suitability of our theoretical results.

One of the most interesting points in our results is that the assumptions we make are as light as possible. For example, we made no restriction on the support of μ, and the assumptions **H1**, **H2** are satisfied in classical stochastic modeling.

Appendix : Proofs

Proof of Theorem A.2.1

Before we prove Theorem A.2.1, we will need to introduce the following definition.

Définition A.6.1 *A function $\phi : \mathcal{H} \to \bar{\mathbb{R}}$ is called lower semi-continuous for the weak topology (abbreviated weakly l.s.c.) if it satisfies one of the following equivalent conditions :*
(i) $\forall t \in \mathbb{R}, \{u \in \mathcal{H} : \phi(u) \leq t\}$ *is closed for the weak topology.*
(ii) $\forall \bar{u} \in \mathcal{H}, \liminf_{u \xrightarrow{w} \bar{u}} \phi(u) \geq \phi(\bar{u})$ *(where* $\xrightarrow{w}$ *note the weak convergence in* $\mathcal{H}$*).*

For a proof of this equivalence and of the following proposition, we refer the reader to the book by Ekeland and Temam [10].

Proposition A.6.1 *With the notation of Definition A.6.1, the two following properties hold :*
(i) If ϕ is continuous and convex, then it is weakly l.s.c.
(ii) If ϕ is weakly l.s.c. on a set Λ which is compact for the weak topology, then ϕ has a minimum on Λ.

Lemma A.6.1 is a straightforward adaptation of the results proven in the first part of the proof of Theorem 1 in Linder [17].

Lemma A.6.1 *There exists $A > 0$ and $\ell \leq k$ such that*

$$\inf_{\mathbf{y}_k \in \mathcal{H}^k} D(\mu, \mathbf{y}_k) = \inf_{\mathbf{y}_\ell \in B_A^\ell} D(\mu, \mathbf{y}_\ell).$$

For all x in $\mathcal{H}$, we define the functions $g_{i,x} : \mathcal{H}^k \to \mathbb{R}$ and $g_x : \mathcal{H}^k \to \mathbb{R}$ by :

$$g_{i,x}(\mathbf{y}_k) = \|x - y_i\|,$$

and

$$g_x(\mathbf{y}_k) = \min_{i=1,\ldots,k} g_{i,x}(\mathbf{y}_k).$$

Lemma A.6.2 *For any x in $\mathcal{H}$, the function g_x is weakly l.s.c. on $\mathcal{H}^k$.*

Proof of Lemma A.6.2 For each x in $\mathcal{H}$, the functions $g_{i,x}$ are continuous and convex, thus they are weakly l.s.c. according to Proposition A.6.1. For all t in $\mathbb{R}$, the sets

$$\left\{\mathbf{y}_k \in \mathcal{H}^k : g_{i,x}(\mathbf{y}_k) \leq t\right\}$$

are then weakly closed. We deduce that

$$\left\{\mathbf{y}_k \in \mathcal{H}^k : g_x(\mathbf{y}_k) \leq t\right\} = \bigcup_{i=1}^{k} \left\{\mathbf{y}_k \in \mathcal{H}^k : g_{i,x}(\mathbf{y}_k) \leq t\right\}$$

is weakly closed. Lemma A.6.2 follows by using statement (i) in Definition A.6.1. □

Lemma A.6.3 *The function $D(\mu, .)$ is weakly l.s.c. on $\mathcal{H}^k$.*

Proof of Lemma A.6.3 For each $\mathbf{y}_k^* \in \mathcal{H}^k$, we can write :

$$\begin{aligned}
\liminf_{\mathbf{y}_k \overset{w}{\to} \mathbf{y}_k^*} D(\mu, \mathbf{y}_k) &= \liminf_{\mathbf{y}_k \overset{w}{\to} \mathbf{y}_k^*} \int_{\mathcal{H}} g_x(\mathbf{y}_k)\mu(dx) \\
&\geq \int_{\mathcal{H}} \liminf_{\mathbf{y}_k \overset{w}{\to} \mathbf{y}_k^*} g_x(\mathbf{y}_k)\mu(dx) \\
&\text{(by Fatou's Lemma)} \\
&\geq \int_{\mathcal{H}} g_x(\mathbf{y}_k^*)\mu(dx) \\
&\text{(by Lemma } A.6.2 \text{ and statement } (ii) \text{ in Definition } A.6.1) \\
&= D(\mu, \mathbf{y}_k^*),
\end{aligned}$$

which proves that $D(\mu, .)$ satisfies the condition (ii) of Definition A.6.1. □

We are now in a position to prove Theorem A.2.1.

Proof of Theorem A.2.1 According to Lemma A.6.1, there exists $R > 0$ such that the infimum of $D(\mu, .)$ on $\mathcal{H}^k$ is also the infimum of $D(\mu, .)$ on B_R^k. Moreover, on the one hand B_R^k is compact for the weak topology, and on the other hand $D(\mu, .)$ is weakly l.s.c. according to Lemma A.6.3. Thus, according to Proposition A.6.1, the function $D(\mu, .)$ reaches its infimum on B_R^k. □

Proof of Theorem A.3.3

The proof is adapted from the proof of Theorem 1 by Bolley, Guillin, and Villani [4]. It can be decomposed in three steps :

1. First, we show we can consider truncated version of the probability measures μ and μ_n on the ball B_R ;
2. Then we cover the space $\mathcal{P}(B_R)$ by small balls of radius r ;
3. Finally, we optimize the various parameters introduced in the proof.

Each of the next three lemmas matches a step.

Let $R > 0$. We consider μ^R defined, for all Borel set $A \subset \mathcal{H}$, by :

$$\mu^R[A] = \frac{\mu[A \cap B_R]}{\mu[B_R]} = \mu[A|B_R].$$

Consider now the independent random variables $\{X_i\}_{i=1}^n$ with distribution μ and $\{Y_i\}_{i=1}^n$ with distribution μ^R. We define, for $i \leq n$,

$$X_i^R = \begin{cases} X_i \text{ if } \|X_i\| \leq R \\ Y_i \text{ if } \|X_i\| > R. \end{cases}$$

Let δ_x be the Dirac measure at point x. The empirical measures μ_n and μ_n^R are defined by

$$\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i} \text{ and } \mu_n^R = \frac{1}{n}\sum_{i=1}^n \delta_{X_i^R}.$$

Note $E_\alpha = \int_{\mathcal{H}} \exp(\alpha\|x\|^2)\mu(dx)$. Since we suppose that μ satisfies a $T_1(\lambda)$-inequality, we have, for $\alpha < \lambda/2$, $E_\alpha < \infty$.

Lemma A.6.4 *Let $\eta \in]0,1[$, $\varepsilon, \theta > 0$, $\alpha_1 \in]0, \lambda/2[$, and $\alpha \in]\alpha_1, \lambda/2[$. Then, for all $R > \max\left(\sqrt{1/2\alpha}, 2\theta/\alpha_1\right)$, we have*

$$\begin{aligned} \mathbb{P}\left[\rho\left(\mu, \mu_n\right) > \varepsilon\right] \;\leq\; & \mathbb{P}\left[\rho\left(\mu^R, \mu_n^R\right) > \eta\varepsilon - 2E_\alpha R e^{-\alpha R^2}\right] \\ & + \exp\left(-n\left[\theta\left(1-\eta\right)\varepsilon - E_\alpha e^{(\alpha_1-\alpha)R^2}\right]\right). \end{aligned}$$

Proof of Lemma A.6.4 For a fixed $\varepsilon > 0$, we bound $\mathbb{P}[\rho(\mu, \mu_n) > \varepsilon]$ in function of μ^R and μ_n^R. First, following the arguments of the proof of Theorem 1.1 by Bolley, Guillin, and Villani (step 1) [4], it can be proven that for all $\alpha < \lambda/2$ and $R \geq \sqrt{1/2\alpha}$,

$$\rho(\mu, \mu^R) \leq 2E_\alpha R e^{-\alpha R^2}. \tag{A.4}$$

Second, the probability measures μ_n and μ_n^R satisfy

$$\rho(\mu_n, \mu_n^R) \leq \frac{1}{n}\sum_{i=1}^n \|X_i^R - X_i\| \leq \frac{1}{n}\sum_{i=1}^n Z_i,$$

where $Z_i = 2\|X_i\|\mathbf{1}_{\|X_i\|>R}$ $(i = 1, \ldots, n)$. Using a similar argument as in the proof of Theorem 1.1 by Bolley, Guillin, and Villani (step 1) [4], we deduce that if ε, θ are positive and $\alpha < \lambda/2$,

$$\mathbb{P}\left[\rho(\mu_n, \mu_n^R) > \varepsilon\right] \leq \exp\left(-n\left[\theta\varepsilon - E_\alpha e^{(\alpha_1-\alpha)R^2}\right]\right). \tag{A.5}$$

The conclusion follows from (A.4), (A.5), and the triangular inequality for ρ. $\square$

Lemma A.6.5 *Given $\theta, \alpha, \alpha_1, \lambda_1 > 0$ such that $\lambda_1 < \lambda$, $\alpha \in]\alpha_1, \lambda/2[$, and $\zeta > 1$, there exist positive constants $\delta_1, \lambda_2 < \lambda_1$, K_1 and K_2 such that, for all $R > \zeta \max\left(\sqrt{1/2\alpha}, 2\theta/\alpha_1\right)$ and $\varepsilon > 0$,*

$$\begin{aligned}\mathbb{P}\left[\rho(\mu, \mu_n) > \varepsilon\right] \;\leq\; & \mathcal{N}(\delta_1\varepsilon/2, B_R)\exp\left(-n\left[\frac{\lambda_2}{2}\varepsilon^2 - K_1R^2e^{-\alpha R^2}\right]\right) \\ & + \exp\left(-n\left[K_2\zeta\varepsilon - K_3e^{(\alpha_1-\alpha)R^2}\right]\right),\end{aligned}$$

where K_3 is a positive constant depending only on θ and α_1.

Proof of Lemma A.6.5 We start by proving that μ^R satisfies a modified $T_1(\lambda)$-inequality. Let Λ be a Borel set of $\mathcal{P}(B_R)$. Following the arguments of the proof of Theorem 1.1 of Bolley, Guillin, and Villani (step 2) [4], one may write

$$\mathbb{P}[\mu_n^R \in \Lambda] \leq \exp\left(-n\inf_{\nu\in\Lambda} H(\nu|\mu^R)\right). \tag{A.6}$$

From now on, we consider that $\mathcal{P}(B_R)$ is equipped with the distance ρ. Consider $\delta > 0$ and A a measurable subset of $\mathcal{P}(B_R)$. We set $\mathcal{N}^A = \mathcal{N}(\delta/2, A)$. Then there exist $\mathcal{N}^A$ balls $B_i, i = 1, \ldots, \mathcal{N}^A$, covering A. Each of this balls is convex and included in the δ-neighborhood A_δ of A. Moreover, by assumption **H2**, the balls B_i are totally bounded.

It is easily inferred from equation (A.6) that

$$\mathbb{P}[\mu_n^R \in A] \leq \mathcal{N}^A \exp\left(-n \inf_{\nu \in A_\delta} H(\nu|\mu^R)\right). \tag{A.7}$$

Define now

$$A = \left\{\nu \in \mathcal{P}(B_R) : \rho(\nu, \mu^R) \geq \eta\varepsilon - 2E_\alpha R e^{-\alpha R^2}\right\}.$$

According to the basic inequality

$$\forall a \in]0,1[, \exists\, C > 0 \text{ such that } \forall x, y \in \mathbb{R}, (x-y)^2 \geq (1-a)x^2 - Cy^2, \tag{A.8}$$

we have, for any $\nu \in \mathcal{H}$,

$$\forall \lambda_1 < \lambda, \exists K > 0 \text{ such that } H(\nu|\mu^R) \geq \frac{\lambda_1}{2}\rho^2(\mu^R, \nu) - KR^2 e^{-\alpha R^2}.$$

Thus, we can write

$$\forall \nu \in A_\delta, \quad H(\nu|\mu^R) \geq \frac{\lambda_1}{2}\rho^2(\mu^R, \nu) - KR^2 e^{-\alpha R^2} \geq \frac{\lambda_1}{2}m^2 - KR^2 e^{-\alpha R^2},$$

where

$$m = \max\left(\eta\varepsilon - 2E_\alpha R e^{-\alpha R^2} - \delta, 0\right).$$

From this and equation (A.7) we conclude that

$$\mathbb{P}\left[\rho(\mu^R, \mu_n^R) \geq \eta\varepsilon - 2E_\alpha R e^{-\alpha R^2}\right] \leq \mathcal{N}^{\mathcal{A}} \exp\left(-n\left[\frac{\lambda_1}{2}m^2 - KR^2 e^{-\alpha R^2}\right]\right). \tag{A.9}$$

Now, given $\lambda_2 < \lambda_1$, it follows from (A.8) that there exist three positive constants δ_1, η_1 and K_1 depending only on α, λ_1, and λ_2 such that

$$\frac{\lambda_1}{2}m^2 - KR^2 e^{-\alpha R^2} \geq \frac{\lambda_2}{2}\varepsilon^2 - K_1 R^2 e^{-\alpha R^2},$$

where $\delta = \delta_1 \varepsilon$. This leads, together with (A.9), to

$$\mathbb{P}\left[\rho(\mu^R, \mu_n^R) \geq \eta\varepsilon - 2E_\alpha R e^{-\alpha R^2}\right] \leq \mathcal{N}^{\mathcal{A}} \exp\left(-n\left[\frac{\lambda_2}{2}\varepsilon^2 - K_1 R^2 e^{-\alpha R^2}\right]\right). \tag{A.10}$$

To bound $\mathcal{N}^A$, we observe that since $A \subset \mathcal{P}(B_R)$,

$$\mathcal{N}^A \leq \mathcal{N}(\delta/2, B_R) = \mathcal{N}(\delta_1\varepsilon/2, B_R).$$

The conclusion follows by Lemma A.6.4 and inequality (A.10). □

The following lemma simplifies the results of the previous.

Lemma A.6.6 *Let $\lambda' < \lambda$, $\alpha < \lambda/2$, and $\alpha' < \alpha$. There exists $\delta_1 > 0$ such that, for all $\varepsilon > 0$,*

$$\mathbb{P}\left[\rho(\mu, \mu_n) > \varepsilon\right] \leq \exp\left(-\frac{\lambda'}{2} n\varepsilon^2\right) + \exp\left(-\alpha' n\varepsilon^2\right),$$

as soon as

$$R^2 \geq R_2 \max\left(1, \varepsilon^2, \ln\left(\frac{1}{\varepsilon^2}\right)\right) \text{ and } n \geq K_4 \frac{\ln\left(\mathcal{N}(\delta_1\varepsilon/2, B_R)\right)}{\varepsilon^2},$$

where R_2 and K_4 are some positive constants depending on μ through λ and α.

Proof of Lemma A.6.6 On the one hand, under the assumptions and notation of Lemma A.6.5, we have, for all $\lambda' < \lambda_2$,

$$\ln\left(\mathcal{N}(\delta_1\varepsilon/2, B_R) \exp\left(-n\left[\frac{\lambda_2}{2}\varepsilon^2 - K_1 R^2 e^{-\alpha R^2}\right]\right)\right) \leq \frac{-n\lambda'\varepsilon^2}{2} \tag{A.11}$$

as soon as R, $R/\ln(1/\varepsilon^2)$ and $n\varepsilon^2/\ln(\mathcal{N}(\delta_1\varepsilon/2, B_R))$ are large enough (see the third step of the proof of Theorem 1.1 by Bolley, Guillin, and Villani [4]).

On the other hand, let $\alpha' < \alpha_2 < \alpha_1$. We can choose ζ such that $K_2\zeta = \alpha_2\varepsilon$. With this choice we obtain

$$\exp\left(-n\left[K_2\zeta\varepsilon - K_3 e^{(\alpha_1-\alpha)R^2}\right]\right) = \exp\left(-n\left[\alpha_2\varepsilon^2 - K_3 e^{(\alpha_1-\alpha)R^2}\right]\right),$$

which can be bounded by $\exp(-\alpha' n\varepsilon^2)$, for R and $R^2/\ln(1/\varepsilon^2)$ large enough. This, together with (A.11), leads to the conclusion. $\square$

Theorem A.3.3 is then a straightforward consequence of Lemma A.6.6, noticing that, for any $K < \min((\lambda'/2), \alpha')$ and n large enough, we have

$$\exp\left(-\frac{\lambda'}{2}n\varepsilon^2\right) + \exp\left(-\alpha' n\varepsilon^2\right) \leq \exp\left(-Kn\varepsilon^2\right).$$

Proof of Theorem A.3.4

Let $\varepsilon > 0$ be small enough. According to Corollary A.3.1 we have

$$\mathbb{P}\left[D(\mu, q_n^*) - D(\mu, q^*) > \varepsilon\right] \leq e^{-(\lambda'/8)n\varepsilon^2},$$

as soon as $n \geq \mathcal{M}(\varepsilon)$. Therefore we can write :

$$\begin{aligned}
\mathbb{E}D(\mu, q_n^*) - D(\mu, q^*) &= \int_0^{+\infty} \mathbb{P}\left[D(\mu, q_n^*) - D(\mu, q^*) > \varepsilon\right] d\varepsilon \\
&= \int_0^{\mathcal{M}^{-1}(n)} \mathbb{P}\left[D(\mu, q_n^*) - D(\mu, q^*) > \varepsilon\right] d\varepsilon \\
&\quad + \int_{\mathcal{M}^{-1}(n)}^{+\infty} \mathbb{P}\left[D(\mu, q_n^*) - D(\mu, q^*) > \varepsilon\right] d\varepsilon \\
&\leq \mathcal{M}^{-1}(n) + \int_0^{+\infty} e^{-(\lambda'/8)n\varepsilon^2} d\varepsilon \\
&\leq C_0 \max(\mathcal{M}^{-1}(n), n^{-1/2}),
\end{aligned}$$

as desired. $\square$

Proof of Theorem A.4.1

One can easily show that

$$D(\mu_n, \mathbf{y}_{k,n}^*) - D(\mu, \mathbf{y}_{k,n}^*) \leq \rho(\mu, \mu_n). \tag{A.12}$$

Thus, by Lemma 4 in Linder [17] and Varadarajan's Theorem [8], we deduce that :

$$D(\mu_n, \mathbf{y}_{k,n}^*) - D(\mu, \mathbf{y}_{k,n}^*) \to 0 \text{ a.s. as } n \to \infty. \tag{A.13}$$

Let $p \leq n$ and $\mathbf{z} \in \{X_1, \ldots, X_p\}^k$. Since $D(\mu_n, \mathbf{y}^*_{k,n}) \leq D(\mu_n, \mathbf{z})$ and, by the law of large number, $D(\mu_n, \mathbf{z}) \to D(\mu, \mathbf{z})$ a.s., we have

$$\limsup_n D(\mu_n, \mathbf{y}^*_{k,n}) \leq D(\mu, \mathbf{z}) \text{ a.s.}$$

From (A.13), we deduce that, for all $p \geq 1$,

$$\limsup_n D(\mu, \mathbf{y}^*_{k,n}) \leq \min_{\mathbf{z} \in \{X_1, \ldots, X_p\}^k} D(\mu, \mathbf{z}). \tag{A.14}$$

Let us now evaluate the limit of the right-hand term in the equation (A.14) as $p \to \infty$. Note, for $\varepsilon > 0$ and $p \geq 1$,

$$\begin{aligned} N(p, \varepsilon) \quad = \quad & \Bigg[\exists\, \mathbf{z}^* \in \operatorname*{arg\,min}_{\mathbf{z} \in \{X_1, \ldots, X_p\}^k} D(\mu, \mathbf{z}) \cap B_{\mathcal{H}^k}(\mathbf{y}^*_k, \varepsilon), \\ & D(\mu, \mathbf{z}^*) \geq D(\mu, \mathbf{y}^*_k) + 2\varepsilon \Bigg]. \end{aligned}$$

Since, $\forall\, \mathbf{y}_k, \mathbf{y}'_k \in \mathcal{H}^k,\ |D(\mu, \mathbf{y}_k) - D(\mu, \mathbf{y}'_k)| \leq \|\mathbf{y}_k - \mathbf{y}'_k\|_k$, we obtain

$$N(p, \varepsilon) \quad \subset [D(\mu, \mathbf{y}^*_k) \geq D(\mu, \mathbf{y}^*_k) + \varepsilon] = \emptyset.$$

Therefore as soon as $p \geq k$,

$$\begin{aligned} & \mathbb{P}\Big[\min_{\mathbf{z} \in \{X_1, \ldots, X_p\}^k} D(\mu, \mathbf{z}) - D(\mu, \mathbf{y}^*_k) > 2\varepsilon \Big] \\ & \leq \mathbb{P}\Big[N\left(p, \varepsilon\right) \Big] + \mathbb{P}\Big[\forall\, \mathbf{z} \in \{X_1, \ldots, X_p\}^k,\ \mathbf{z} \notin B_{\mathcal{H}^k}(\mathbf{y}^*_k, \varepsilon) \Big] \\ & \leq \mathbb{P}\Big[\left(X_1, \ldots, X_k\right) \notin B_{\mathcal{H}^k}\left(\mathbf{y}^*_k, \varepsilon\right) \Big]^{\lfloor p/k \rfloor} \\ & = \Big(1 - \mathbb{P}\big[(X_1, \ldots, X_k) \in B_{\mathcal{H}^k}(\mathbf{y}^*_k, \varepsilon) \big] \Big)^{\lfloor p/k \rfloor}, \end{aligned} \tag{A.15}$$

where $\lfloor . \rfloor$ stands for the integer part function. Then, by the Borel-Cantelli lemma,

$$\lim_{p \to \infty} \min_{\mathbf{z} \in \{X_1, \ldots, X_p\}^k} D(\mu, \mathbf{z}) = D(\mu, \mathbf{y}^*_k) \text{ a.s.}$$

This result, together with (A.14), leads to the conclusion. □

Proof of Theorem A.4.2

On the one hand we can write :

$$\begin{aligned}
&D(\mu, \mathbf{y}^*_{k,n}) - D^*_k(\mu) \\
&= D(\mu, \mathbf{y}^*_{k,n}) - D(\mu_n, \mathbf{y}^*_{k,n}) + D(\mu_n, \mathbf{y}^*_{k,n}) - D^*_k(\mu) \\
&\leq |D(\mu, \mathbf{y}^*_{k,n}) - D(\mu_n, \mathbf{y}^*_{k,n})| + |D(\mu_n, \mathbf{y}^*_{k,n}) - D^*_k(\mu)| \\
&\leq \rho(\mu, \mu_n) + |D(\mu_n, \mathbf{y}^*_{k,n}) - D^*_k(\mu)|,
\end{aligned}$$

according to (A.12).

On the other hand,

$$\lim_{n\to\infty} D(\mu_n, \mathbf{y}^*_{k,n}) = D^*_k(\mu) \text{ a.s.}$$

Moreover,

$$\begin{aligned}
D(\mu_n, \mathbf{y}^*_{k,n}) &= \min_{\mathbf{z}\in\{X_1,\dots,X_n\}^k} \frac{1}{n}\sum_{i=1}^n \min_{j=1,\dots,k} \|X_i - \mathbf{z}_j\| \\
&\leq \frac{1}{n}\sum_{i=1}^n \|X_i - X_1\| \\
&\leq \frac{1}{n}\sum_{i=1}^n \|X_i\| + \|X_1\|.
\end{aligned}$$

Hence, $D(\mu_n, \mathbf{y}^*_{k,n})$ is equi-integrable, which proves that it converges in L_1.

Finally, $\mathbb{E}\rho(\mu, \mu_n) \to 0$ by Theorem A.3.3, and we deduce the proof of Theorem A.4.2. $\square$

Proof of Theorem A.4.3

First we can write

$$\begin{aligned}
D(\mu, \mathbf{y}^*_{k,n}) - D^*_k(\mu) &= D(\mu, \mathbf{y}^*_{k,n}) - D(\mu_n, \mathbf{y}^*_{k,n}) \\
&+ D(\mu_n, \mathbf{y}^*_{k,n}) - \min_{\mathbf{z}\in\{X_1,\dots,X_n\}^k} D(\mu, \mathbf{z}) \\
&+ \min_{\mathbf{z}\in\{X_1,\dots,X_n\}^k} D(\mu, \mathbf{z}) - D^*_k(\mu).
\end{aligned}$$

Then, according to Lemma 3 in Linder [17], we have

$$D(\mu, \mathbf{y}_{k,n}^*) - D(\mu_n, \mathbf{y}_{k,n}^*) \leq \rho(\mu, \mu_n)$$

and

$$D(\mu_n, \mathbf{y}_{k,n}^*) - \min_{\mathbf{z} \in \{X_1, \ldots, X_n\}^k} D(\mu, \mathbf{z}) \leq \rho(\mu, \mu_n).$$

Thus,

$$D(\mu, \mathbf{y}_{k,n}^*) - D_k^*(\mu) \leq 2\rho(\mu, \mu_n) + \min_{\mathbf{z} \in \{X_1, \ldots, X_n\}^k} D(\mu, \mathbf{z}) - D_k^*(\mu). \qquad \text{(A.16)}$$

Moreover, according to the inequality (A.15), we have for $n \geq k$:

$$\mathbb{P}\left[\min_{\mathbf{z} \in \{X_1, \ldots, X_n\}^k} D(\mu, \mathbf{z}) - D_k^*(\mu) \geq 2\varepsilon\right] \leq \left[1 - f(\mathbf{y}_k^*, \varepsilon)\right]^{\lfloor n/k \rfloor}.$$

We deduce

$$\begin{aligned}
&\mathbb{E} \min_{\mathbf{z} \in \{X_1, \ldots, X_n\}^k} D(\mu, \mathbf{z}) - D_k^*(\mu) \\
&= \int_0^{+\infty} \mathbb{P}\Big[\min_{\mathbf{z} \in \{X_1, \ldots, X_n\}^k} D(\mu, \mathbf{z}) - D_k^*(\mu) \geq \varepsilon\Big] d\varepsilon \\
&\leq 2 \int_0^{+\infty} \Big(1 - f(\mathbf{y}_k^*, \varepsilon)\Big)^{\lfloor n/k \rfloor} d\varepsilon \\
&\leq 2\left(\int_{[0,u] \cup [v, \infty[} \Big(1 - f(\mathbf{y}_k^*, \varepsilon)\Big)^{\lfloor n/k \rfloor} d\varepsilon + \int_u^v \Big(1 - f(\mathbf{y}_k^*, \varepsilon)\Big)^{\lfloor n/k \rfloor} d\varepsilon \right) \\
&\leq 2\left(2V(n) + \int_u^v \Big[1 - f(\mathbf{y}_k^*, \varepsilon)\Big]^{\lfloor n/k \rfloor} d\varepsilon \right) \\
&\quad \text{(according to assumption } \mathbf{H3}) \\
&\leq 2\left(2V(n) + (v-u)\Gamma^{\lfloor n/k \rfloor}\right) \\
&\leq C \max\left(\lfloor n/k \rfloor^{-1/2}, V(n)\right) \text{ for } n \text{ large enough,}
\end{aligned}$$

where $\Gamma < 1$ and C are some positive constants. Theorem A.4.3 follows from (A.16), Theorem A.3.3 and Theorem A.3.4. $\square$

Bibliography

[1] E. Abaya and G. Wise. Convergence of vector quantizers with application to optimal quantization. *SIAM Journal on Applied Mathematics*, 44 :183–189, 1984.

[2] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54 :781–790, 2007.

[3] K. Bleakley. *Quelques Contributions à l'Analyse Statistique et à la Classification des Graphes et des Courbes. Applications à l'Immunobiologie et à la Reconstruction des Réseaux Biologiques.* PhD thesis, Université Montpellier II, 2007.

[4] F. Bolley, A. Guillin, and C. Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3-4) :541–593, 2007.

[5] B. Cadre. Convergent estimators for the L_1-median of a Banach valued random variable. *Statistics*, 35(4) :509–521, 2001.

[6] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1) :1–49, 2002.

[7] H. Djellout, A. Guillin, and L. Wu. Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *Annals of Probability*, 32(3B) :2702–2732, 2004.

[8] R. M. Dudley. *Real Analysis and Probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

[9] N. Dunford and J. T. Schwartz. *Linear Operators. Part I.* Wiley Classics Library. John Wiley & Sons Inc., New York, 1988. General theory, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1958 original, A Wiley-Interscience Publication.

[10] I. Ekeland and R. Temam. *Analyse Convexe et Problèmes Variationnels.* Dunod, 1974. Collection Études Mathématiques.

[11] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Norwell, MA, USA, 1991.

[12] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000.

[13] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, New York-London-Sydney, 1975. Wiley Series in Probability and Mathematical Statistics.

[14] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1990.

[15] J. H. B. Kemperman. The median of a finite measure on a Banach space. In *Statistical data analysis based on the L_1-norm and related methods (Neuchâtel, 1987)*, pages 217–230. North-Holland, Amsterdam, 1987.

[16] M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001.

[17] T. Linder. Learning-theoretic methods in vector quantization. In *Principles of Nonparametric Learning (Udine, 2001)*, volume 434 of *CISM Courses and Lectures*, pages 163–210. Springer, Vienna, 2002.

[18] T. Linder, G. Lugosi, and K. Zeger. Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Transactions on Information Theory*, 40 :1728–1740, 1994.

[19] D. Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, 9 :135–140, 1981.

[20] D. Pollard. A central limit theorem for k-means clustering. *The Annals of Probability*, 10 :919–926, 1982.

[21] D. Pollard. Quantization and the method of k-means. *IEEE Transactions on Information Theory*, 28 :199–205, 1982.

[22] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.

[23] A. W. Van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

Résumé

L'objectif de cette Thèse est d'apporter une contribution au problème de l'apprentissage statistique, notamment en développant des méthodes pour prendre en compte des données fonctionnelles. Dans la première partie, nous développons une approche de type plus proches voisins pour la régression fonctionnelle. Dans la deuxième, nous étudions les propriétés de la méthode de quantification dans des espaces de dimension infinie. Nous appliquons ensuite cette méthode pour réaliser une étude comportementale de bancs d'anchois. Enfin, la dernière partie est dédiée au problème de l'estimation des ensembles de niveaux de la fonction de régression dans un cadre multivarié.

Mots-clefs : Apprentissage statistique, Apprentissage supervisé, Apprentissage non supervisé, Données fonctionnelles, Classification, Régression, Quantification, Plus proches voisins, Estimateurs à noyaux, Ensembles de niveaux.

Abstract

The goal of this thesis is to contribute to the domain of statistical learning, and includes the development of methods that can deal with functional data. In the first section, we develop a Nearest Neighbor approach for functional regression. In the second, we study the properties of a quantization method in infinitely-dimensional spaces. We then apply this approach to a behavioral study of schools of anchovies. The last section is dedicated to the problem of estimating level sets of the regression function in a multivariate context.

Keywords : Statistical learning, Supervised Learning, Unsupervised Learning, Functional data, Classification, Regression, Quantization, Nearest Neighbor, Kernel estimators, Level sets.