Université Pierre et Marie Curie – Paris VI

T H È S E

Présentée pour obtenir

Le titre de docteur en sciences de
l'université Pierre et Marie Curie – Paris VI

Spécialité : Mathématiques

Ecole Doctorale : Sciences Mathématiques de Paris Centre

Apprentissage statistique non supervisé : grande dimension et courbes principales.

par

Aurélie FISCHER

Soutenue le 9 juin 2011 devant le jury composé de :

M. Gérard BIAU	Professeur, Université Paris VI	Directeur de thèse
M. Jérôme DEDECKER	Professeur, Université Paris V	Examinateur
M. Paul DEHEUVELS	Professeur, Université Paris VI	Président
M. Fabrice GAMBOA	Professeur, Université Toulouse III	Rapporteur
M. Balázs KEGL	Chargé de recherches, Université Paris XI	Examinateur
M. Pascal MASSART	Professeur, Université Paris XI	Examinateur

Rapporteurs :

M. Fabrice GAMBOA	Professeur, Université Toulouse III
M. Gábor LUGOSI	Professeur, Université Pompeu Fabra de Barcelone

Apprentissage statistique non supervisé : grande dimension et courbes principales.

Remerciements

« Qu'allez-vous choisir ? Si vous n'êtes pas absolument sûre de ce que vous voulez faire, choisissez l'Ecole Centrale car elle vous laisse le plus de latitude pour corriger votre trajectoire. »

Devant les différentes possibilités qui s'offraient à moi à l'issue des classes préparatoires, voilà en substance le conseil de mon professeur de Math Spé, Jean-Denis Eiden, qui ne m'en voudra pas, je l'espère, d'avoir ainsi un peu rapidement résumé sa pensée. Il serait exagéré de dire que je savais exactement, à ce moment-là, quelle profession je souhaitais exercer. L'enseignement me plaisait déjà, mais j'avais alors une trop vague idée de ce qu'était la recherche pour décider d'en faire mon métier. Pourtant, je n'ai pas suivi l'avis de mon professeur... Plutôt que l'Ecole Centrale, représentant une carrière d'ingénieur qui ne me correspondait pas, c'est le magistère de l'ENS que j'ai choisi, et ce choix, que je ne regrette pas, m'a menée jusqu'ici.

* § * §§ * §§§ * §§§ * §§§ * §§§§§ * §§§ * §§§ * §§§ * §§ * § *

En premier lieu, je remercie de tout cœur Gérard Biau d'avoir bien voulu encadrer ma thèse et de m'avoir proposé des axes de recherche partant de jolis objets mathématiques. Il s'est toujours montré intéressé et disponible, malgré un emploi du temps chargé dû à ses diverses responsabilités. Pour la clarté de ses explications et sa rigueur scientifique, pour la confiance qu'il m'a accordée, son enthousiasme et sa bonne humeur, et parce qu'il est le vivant exemple de ce qu'il m'a enseigné, je tiens à lui exprimer toute ma gratitude. J'espère vivement que nous aurons l'occasion de travailler ensemble par la suite.

J'adresse également mes remerciements à Paul Deheuvels, directeur du LSTA, pour m'avoir accueillie dans ce laboratoire au sein duquel j'ai disposé de conditions idéales pour écrire cette thèse. Il me fait un grand honneur en présidant le jury.

Je suis reconnaissante à Fabrice Gamboa et Gábor Lugosi de l'intérêt qu'ils ont témoigné pour ce travail en acceptant la difficile tâche de rapporteur. Je les remercie bien sincèrement pour leur lecture attentive de mon manuscrit et leurs remarques constructives.

Je suis heureuse de retrouver dans mon jury de thèse Pascal Massart, qui m'a proposé au cours du Master 2 d'intéressantes lectures grâce auxquelles j'ai découvert des mathématiques plaisantes et beaucoup appris. Je lui sais gré d'être tombé remarquablement juste en m'orientant vers Gérard pour ma thèse.

Je remercie Jérôme Dedecker, mon voisin à Chevaleret, qui avait gentiment accepté d'être mon tuteur pédagogique et m'a toujours encouragée, et Bálazs Kégl, qui a aimablement répondu à mes questions concernant ses travaux, d'être eux aussi présents dans ce jury.

Merci à Benjamin Auder. Notre collaboration m'a donné la possibilité d'étudier une application bien concrète de nos thèmes de recherche en ingénierie nucléaire.

Merci à tous les membres du laboratoire, avec qui j'ai eu beaucoup de plaisir à discuter de mathématiques, comme d'innombrables autres sujets. Merci à Michel Broniatowski pour son appui au moment où j'ai déposé ma thèse.

Agathe * Bertrand * Etienne * Fanny * Frédéric * Jean-Patrick
Nathalie * Olivier * Philippe * Stéphane * Tabea

Parce que la recherche en mathématiques ne se conçoit pas sans bibliothèque, merci à Pascal Epron. Je tiens également à remercier Anne Durrande, Louise Lamart et Corinne van Vlierberghe pour leur gentillesse et leur efficacité.

A Jussieu, comme à Chevaleret, à l'occasion d'un groupe de travail, ou d'une conférence, j'ai rencontré nombre de doctorants bien sympathiques. Merci à tous !

J'ai la chance d'avoir des amis formidables, grâce à qui ces quelques années à Paris ont été lumineuses. Merci pour ce que nous avons partagé, de la musique aux longueurs à la piscine, en passant par les expériences culinaires et ces nombreuses conversations, qu'elles soient profondes ou ponctuées de fous rires, merci pour tout ce qui m'a permis de m'échapper, quand il le fallait, des problèmes de convergence et autres inégalités.

Anne * Claire * Emeline * Madeleine * Marie-Amélie * Olivier
Pascal * Philippe * Pierre * Sabine * Sophie * Sylvain

Je remercie mes parents Anne Lise et Jean Paul de donner ce bel exemple de l'enseignement, dans la simplicité, avec les plus petits comme avec les plus grands, et Jean Christophe de m'avoir sauvée de quelques catastrophes informatiques. Mais ils savent bien que c'est pour tellement plus que cela qu'avec joie je leur dis ce MERCI.

Table des matières

Introduction

1.1. Présentation de la thèse

L'apprentissage statistique désigne un ensemble de méthodes et d'algorithmes permettant d'extraire de l'information pertinente de données ou d'apprendre un comportement à partir de l'observation d'un phénomène. En général, ce processus est associé à la possibilité de mesurer en un certain sens la qualité et la précision des résultats. L'apprentissage comprend deux grandes branches : l'apprentissage supervisé d'une part et l'apprentissage non supervisé d'autre part. L'apprentissage supervisé se fixe pour objectif de trouver une fonction reliant des entrées $X_1, \ldots, X_n$ à des sorties $Y_1, \ldots, Y_n$, alors que l'apprentissage non supervisé vise à construire un modèle permettant de décrire un ensemble d'entrées $X_1, \ldots, X_n$.

Plus précisément, dans le cas supervisé, la finalité est de déterminer une nouvelle sortie Y à partir d'une nouvelle entrée X, connaissant un ensemble d'observations $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. Lorsque les Y_i prennent des valeurs discrètes, il s'agit d'un problème de classification — en classification binaire, par exemple, on cherche à attribuer à X une étiquette 0 ou 1 —, tandis que des Y_i à valeurs réelles nous placent dans le cadre de la régression. En apprentissage non supervisé, en revanche, il n'y a pas de sortie, et il s'agit alors de représenter au mieux les observations $X_1, \ldots, X_n$, de manière à la fois précise et compacte.

Le contexte général de cette thèse est celui de l'apprentissage non supervisé. Nous examinons plus particulièrement les problématiques du clustering et des courbes principales. Nous avons choisi de diviser le présent document en deux parties, chacune consacrée à l'une de ces deux notions, suivies d'une annexe qui rassemble tout d'abord quelques outils techniques utilisés tout au long de la thèse, puis reprend les articles correspondants rédigés en anglais.

La **première partie** concerne la **quantification** et le **clustering**. Le clustering se donne pour objet de repérer dans la structure des données des groupes homogènes et bien séparés, sans qu'il existe pour autant d'étiquettes ou de groupes connus à l'avance (Duda, Hart et Stork [76]). En termes plus probabilistes, un autre aspect du même problème est la quantification (ou compression de données avec perte), qui consiste à remplacer une variable aléatoire X par une représentation

compacte $q(X)$ (voir par exemple le livre de Gersho et Gray [94]). Plus formellement, si X est à valeurs dans un espace $\mathcal{X}$, le quantificateur q est une fonction qui envoie tout élément de $\mathcal{X}$ dans un sous-ensemble fini de $\mathcal{X}$. L'erreur commise en remplaçant X par sa version quantifiée $q(X)$ est alors évaluée grâce à la distorsion $\mathbb{E}[\mathcal{d}(X, q(X))]$, où $\mathcal{d}$ désigne une fonction de perte positive, appelée mesure de distorsion. Dans le premier chapitre, nous étudions les propriétés théoriques de la quantification en utilisant une classe de mesures de distorsion appelées divergences de Bregman (Bregman [40]). Le deuxième chapitre, rédigé en anglais, traite d'une méthode de clustering de courbes dans le cadre de l'industrie nucléaire. Cet article écrit en collaboration avec Benjamin Auder (doctorant au CEA de Cadarache) est à paraître dans la revue *Journal of Statistical Computation and Simulation*. Dans le troisième chapitre, nous nous intéressons au choix du nombre de groupes.

Les **courbes principales** constituent le thème de la **seconde partie**. Cette notion, introduite dans les années 1980 par Hastie et Stuetzle [104], peut être vue comme une généralisation non linéaire de l'analyse en composantes principales. Une courbe principale est une courbe paramétrée de $\mathbb{R}^d$ « passant au milieu » d'un nuage de points ou d'une distribution. Il existe plusieurs points de vue donnant un sens mathématique à cette notion intuitive et le premier chapitre présente une synthèse bibliographique sur ce sujet. Le second chapitre s'attache à déterminer une classe de courbes appropriée afin d'obtenir une courbe principale convenable.

L'**annexe** qui complète ce document présente tout d'abord un résumé sur les moyennes de Rademacher, puis quelques rappels de sélection de modèle et ensuite les définitions essentielles relatives aux courbes paramétrées. Elle s'achève par trois articles rédigés en anglais : l'article correspondant au Chapitre 1 de la première partie sous sa version publiée dans la revue *Journal of Multivariate Analysis*, un article reprenant le Chapitre 3 de la première partie et enfin, un article écrit en collaboration avec Gérard Biau, qui reprend la Section 2.2 du Chapitre 2 de la seconde partie.

Dans tout le document, $\mathbf{X}$ désigne un vecteur aléatoire de $\mathbb{R}^d$ et $\mathbf{f}$ une courbe paramétrée.

1.2. Première partie : Quantification et clustering

Le **premier chapitre** de cette partie est consacré à l'étude de quelques propriétés de la quantification et du clustering, dans le cas où la mesure de distorsion

est une divergence de Bregman. Il s'insère dans la continuité des travaux de Banerjee, Merugu, Dhillon et Ghosh [17], qui ont montré que l'algorithme de clustering des k-means (Lloyd [132], Linde, Buzo et Gray [129]) peut être généralisé à cette classe de mesures de distorsion. De nombreuses mesures de dissimilarité fréquemment utilisées en statistique et en théorie de l'information sont des cas particuliers de divergences de Bregman, d'où l'intérêt de considérer cette classe de mesures de distorsion indexée par des fonctions strictement convexes, introduite par Bregman [40] en 1967. En outre, certaines de ces divergences s'appliquent à des fonctions ou encore à des mesures de probabilité, ce qui en fait des outils appropriés pour classer des observations de grande dimension ou de nature complexe. Cette caractéristique est très appréciable compte tenu de l'afflux croissant de telles données dans de nombreux domaines.

En dimension finie, la divergence de Bregman associée à une fonction ϕ strictement convexe et différentiable est donnée par

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle,$$

où $\langle\cdot,\cdot\rangle$ désigne le produit scalaire canonique de l'espace euclidien $(\mathbb{R}^d, \|\cdot\|)$, et $\nabla\phi(y)$ le gradient de ϕ au point y. La distance euclidienne standard au carré est par exemple obtenue pour $\phi(x) = \|x\|^2$. Cette définition se généralise à un espace de Banach en écrivant

$$d_\phi(x, y) = \phi(x) - \phi(y) - D_y\phi(x - y),$$

où $D_y\phi(h)$ est la différentielle de ϕ au point y appliquée à h.

Pour éviter d'alourdir la présentation, nous omettons volontairement certains détails qui seront explicités dans le chapitre correspondant. Soit X une variable aléatoire de loi μ à valeurs dans $\mathcal{X}$, $X_1, \ldots, X_n$ un échantillon de X et d_ϕ une divergence de Bregman. Un quantificateur q envoyant tout $x \in \mathcal{X}$ sur l'un des k éléments $c_1, \ldots, c_k$ de $\mathcal{X}$ est caractérisé par cette famille de représentants $c_1, \ldots, c_k$, appelée table de codage ou ensemble de centres, et par la partition de $\mathcal{X}$ en k cellules $S_1, \ldots, S_k$ induite par la relation $x \in S_j$ si, et seulement si, $q(x) = c_j$. L'erreur résultant de la substitution de $q(X)$ à X est mesurée par la distorsion

$$W(\mu, q) = \mathbb{E}[d_\phi(X, q(X))],$$

dont l'équivalent pour la mesure empirique est donné par

$$W(\mu_n, q) = \frac{1}{n}\sum_{i=1}^{n} d_\phi(X_i, q(X_i)).$$

Notons

$$W^\star(\mu) = \inf_q W(\mu, q)$$

la distorsion optimale, qui sert de référence pour juger de la précision d'un quantificateur. Pour un ensemble de centres donné, la meilleure partition au sens de la distorsion est la partition dite de Voronoi, définie en affectant un élément x à S_j si, et seulement si, x est plus proche de c_j que de tout autre c_ℓ. Cette propriété a pour conséquence qu'il suffit de considérer les quantificateurs associés à la partition de Voronoi, appelés quantificateurs des plus proches voisins. Un tel quantificateur est donc décrit par sa table de codage. En fonction de $\mathbf{c} = (c_1, \ldots, c_k)$, la distorsion se récrit

$$W(\mu, \mathbf{c}) = \mathbb{E}\Big[\min_{j=1,\ldots,k} d_\phi(X, c_j)\Big].$$

De même, la distorsion empirique devient

$$W(\mu_n, \mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n}\Big[\min_{j=1,\ldots,k} d_\phi(X_i, c_j)\Big].$$

Dans un premier temps, utilisant un argument de continuité lié à une propriété de compacité, nous montrons que, sous des conditions appropriées, il existe $\mathbf{c}^\star$ tel que

$$W(\mu, \mathbf{c}^\star) = W^\star(\mu).$$

En d'autres termes, il existe une table de codage qui est optimale au sens de la distorsion $W(\mu, \mathbf{c})$. Ceci est vrai en particulier pour la mesure empirique μ_n. Une question naturelle consiste alors à se demander si une table de codage empirique optimale $\mathbf{c}_n^\star$ basée sur un ensemble d'observations $X_1, \ldots, X_n$ constitue, pour n suffisamment grand, une bonne approximation d'un ensemble de centres optimaux $\mathbf{c}^\star$. C'est pourquoi nous nous intéressons dans un deuxième temps à la convergence de $W(\mu, \mathbf{c}_n^\star)$ vers le minimum de distorsion $W^\star(\mu)$. En supposant que X reste presque sûrement dans une boule de rayon R et sous certaines conditions, le résultat principal prend ainsi la forme

$$\mathbb{E}\left[W(\mu, \mathbf{c}_n^\star)\right] - W^\star(\mu) \leq \frac{4kC(\phi, R)}{\sqrt{n}}, \tag{1.1}$$

où $C(\phi, R) > 0$ est une constante qui ne dépend que du rayon de la boule et de la divergence de Bregman d_ϕ utilisée. Puisque l'espace ambiant peut être de dimension élevée voire infinie, l'intérêt de ce type de borne non-asymptotique est de ne pas faire intervenir la dimension. L'inégalité (1.1), démontrée en utilisant des moyennes de Rademacher comme mesure de complexité pour une classe fonctionnelle, est explicitée dans le chapitre pour quelques divergences de Bregman

usuelles. Par exemple, pour la perte exponentielle, correspondant à $\phi(x) = e^x$, la borne devient

$$\mathbb{E}[W(\mu, \mathbf{c}_n^\star)] - W^\star(\mu) \leq \frac{4k(2R-1)e^R}{\sqrt{n}}.$$

Nous retrouvons également le cas particulier d'une norme hilbertienne (Biau, Devroye et Lugosi [32]) :

$$\mathbb{E}[W(\mu, \mathbf{c}_n^\star)] - W^\star(\mu) \leq \frac{12kR^2}{\sqrt{n}}.$$

Quelques simulations illustrant l'utilisation de différentes divergences de Bregman en clustering complètent ce chapitre. La Figure 1.1 montre ainsi le résultat du clustering en deux groupes d'un ensemble de 100 observations distribuées suivant la loi uniforme sur un cercle et une bande, pour la norme euclidienne au carré, la distance de Kullback-Leibler et la distance de Itakura-Saito.

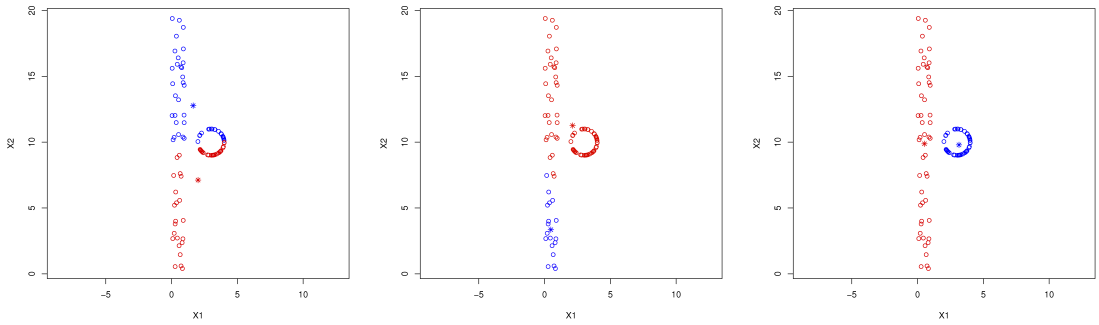


FIGURE 1.1.: Clustering de données distribuées uniformément sur une bande ou un cercle (k=2,n=100). [A] Norme euclidienne au carré. [B] Distance de Kullback-Leibler généralisée. [C] Distance de Itakura-Saito.

Le **deuxième chapitre**, qui résulte d'une collaboration avec Benjamin Auder (doctorant au CEA de Cadarache), s'inscrit dans le contexte de l'industrie nucléaire. Il concerne plus précisément le problème important posé par la prévention des risques d'accidents au niveau de la cuve d'un réacteur. Pour étudier ces phénomènes, des simulations sont réalisées à l'aide d'un code de calcul appelé CATHARE (Code Avancé de THermohydraulique pour les Accidents des Réacteurs à Eau). Etant donné des paramètres physiques d'entrée, le code CATHARE fournit certaines courbes de pression, de température et de coefficient d'échanges thermiques, qui seront utilisées ensuite pour déterminer s'il peut se produire une rupture de la cuve. La Figure 1.2 montre un exemple de telles courbes.

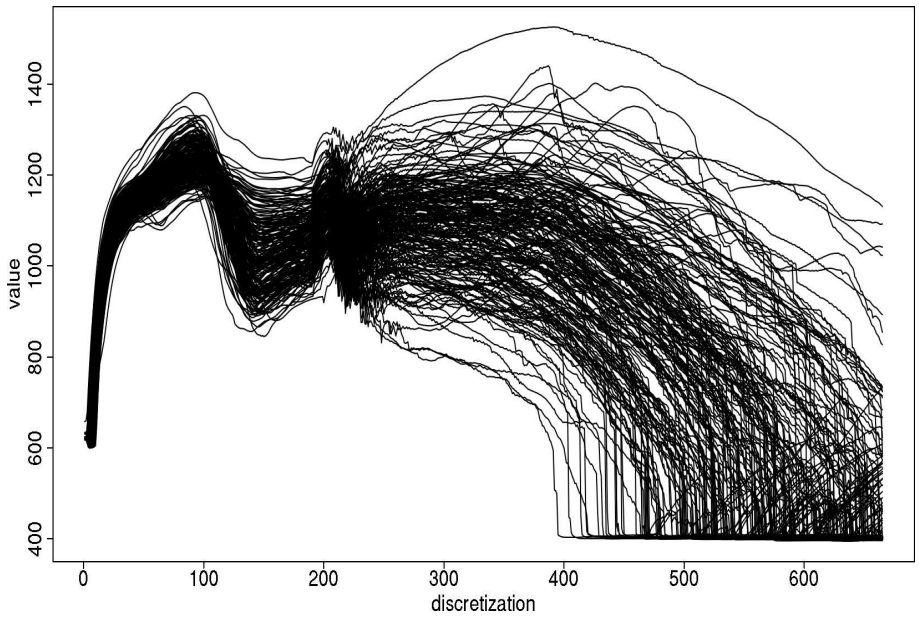


FIGURE 1.2.: Courbes de coefficient d'échanges thermiques.

Il s'avère cependant que le code CATHARE est très coûteux en temps de calcul et peut, pour cette raison, difficilement être utilisé directement pour des calculs de fiabilité. Il est donc nécessaire de l'approcher par un "métamodèle", c'est-à-dire un code simplifié et plus rapide. Au CEA, ce métamodèle est construit en pratiquant une régression à partir d'une série de résultats obtenus antérieurement avec CATHARE. Afin de gagner en précision, il est néanmoins indispensable de classer les courbes au préalable, et d'effectuer ainsi la régression pour chacun des groupes pris séparément. Le clustering d'objets de dimension potentiellement infinie posant problème du point de vue des calculs numériques, l'objet de ce deuxième chapitre est précisément d'étudier les propriétés d'une technique de réduction de la dimension pour cette étape de clustering de courbes.

Sur le plan théorique, nous supposons que les courbes à classer sont des éléments de $L^2([0,1])$ qui se décomposent sur une base hilbertienne avec des coefficients appartenant au sous-ensemble $\mathcal{S}$ de l'espace ℓ^2 des suites de carré sommable donné par

$$\mathcal{S} = \left\{ \mathbf{x} = (x_j)_{j\geq 1} \in \ell^2 : \sum_{j=1}^{+\infty} \varphi_j x_j^2 \leq R^2 \right\},$$

où $R > 0$ et $(\varphi_j)_{j\geq 1}$ est une suite positive strictement croissante tendant vers l'infini. Il est important de noter que l'ensemble $\mathcal{S}$ est étroitement lié au choix de la base hilbertienne, même si cela n'apparaît pas explicitement dans la définition.

Adoptant des notations similaires à celles employées dans la présentation du premier chapitre, la distorsion est définie par

$$W_\infty = \mathbb{E}\left[\min_{\ell=1,\dots,k} \|X - c_\ell\|^2\right].$$

Posons

$$W_d = \mathbb{E}\left[\min_{\ell=1,\dots,k} \|\Pi_d(X) - \Pi_d(c_\ell)\|^2\right],$$

où Π_d désigne la projection sur $\mathbb{R}^d$. La version empirique de cette distorsion « d-dimensionnelle » est donnée par

$$W_{d,n}(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n} \min_{\ell=1,\dots,k} \|\Pi_d(X_i) - \Pi_d(c_\ell)\|^2.$$

La stratégie de réduction de la dimension mise en œuvre dans ce chapitre consiste à effectuer le clustering dans l'espace de projection de dimension d. Si $\hat{\mathbf{c}}_{d,n}$ désigne un minimiseur de $W_{d,n}(\mathbf{c})$, un contrôle de l'écart entre $W_\infty(\hat{\mathbf{c}}_{d,n})$ et la distorsion minimale $W_\infty^\star$ permet d'évaluer la qualité de la procédure. Le principal résultat exprime ainsi le fait que la perte en espérance pour le clustering dans l'espace de dimension infinie est bornée par la perte « fini-dimensionnelle » correspondante à laquelle s'ajoute un terme représentant le coût de la projection sur $\mathbb{R}^d$:

$$\mathbb{E}[W_\infty(\hat{\mathbf{c}}_{d,n})] - W_\infty^\star \leq \mathbb{E}[W_d(\hat{\mathbf{c}}_{d,n})] - W_d^\star + \frac{8R^2}{\varphi_d}.$$

Cette inégalité permet de discuter le choix de la dimension de projection d en fonction de n, puisque le terme $\mathbb{E}[W_d(\hat{\mathbf{c}}_{d,n})] - W_d^\star$ peut être borné par un terme de l'ordre de $1/\sqrt{n}$. Par exemple, si $\mathcal{S}$ est un ellipsoïde de Sobolev, avec

$$\varphi_j = \begin{cases} j^{2\beta} & \text{si } j \text{ pair} \\ (j-1)^{2\beta} & \text{si } j \text{ impair,} \end{cases}$$

la dimension d de projection doit être de l'ordre de $n^{1/4\beta}$ pour conserver une vitesse en $1/\sqrt{n}$.

Comme nous l'avons mentionné plus haut, la forme de l'ensemble $\mathcal{S}$ dépend de la base hilbertienne considérée. Il est donc essentiel de choisir une base appropriée. D'un point de vue pratique, notre approche est basée sur un algorithme permettant de construire une base à l'aide de paquets d'ondelettes suivant la méthode de Coifman et Wickerhauser [54], en incluant une comparaison avec les bases de Haar et de Fourier ainsi que la base de l'analyse en composantes principales fonctionnelle. Outre l'application au problème industriel, une illustration sur données simulées est également présentée.

Dans ces deux premiers chapitres, le nombre de groupes k est supposé connu. Cependant, le choix du nombre de classes à spécifier constitue une question majeure en clustering. Pour obtenir un résultat pertinent, en évitant autant que possible de couper artificiellement une classe ou de fusionner plusieurs groupes, il est en effet indispensable de déterminer correctement k. Dans le **troisième chapitre**, nous discutons ainsi une méthode permettant de le sélectionner automatiquement. Il existe dans la littérature diverses heuristiques pour le choix du nombre de groupes. L'approche que nous proposons se base sur l'idée naturelle consistant à l'évaluer à partir de la structure des données, par l'intermédiaire de la distorsion empirique $W(\mu_n, \mathbf{c})$, fonction décroissante de k. Pour tout k, la minimisation en $\mathbf{c} = (c_1, \ldots, c_k)$ de $W(\mu_n, \mathbf{c})$ fournit une table de codage $\hat{\mathbf{c}}_k$. L'objectif est alors de sélectionner la meilleure $\hat{\mathbf{c}}_k$ sur toutes les valeurs possibles de k. Ce problème est envisagé dans le chapitre sous l'angle de la sélection de modèle par pénalisation développée par Birgé et Massart [35] et Barron, Birgé et Massart [21]. Il s'agit de trouver une fonction de pénalité adéquate, telle que la minimisation du critère pénalisé

$$\mathrm{crit}(k) = W(\mu_n, \hat{\mathbf{c}}_k) + \mathrm{pen}(k)$$

conduise au bon nombre de groupes.

Dans ce contexte, nous prouvons que pour une fonction de pénalité $\mathrm{pen}(k)$ de l'ordre de $\sqrt{k/n}$, l'ensemble de centres $\tilde{\mathbf{c}}$ obtenu par minimisation de la distorsion empirique pénalisée vérifie

$$\mathbb{E}[W(\mu, \tilde{\mathbf{c}})] \leq \inf_{1 \leq k \leq n} \left(W(\mu, S_k) + \mathrm{pen}(k) \right) + r_n,$$

où $W(\mu, S_k) = \inf_{\mathbf{c} \in S_k} W(\mu, \mathbf{c})$ et r_n tend vers 0 lorsque n tend vers l'infini. Ce troisième chapitre se termine par la mise en pratique de la méthode de choix de k dans quelques simulations et expériences sur données réelles.

1.3. Deuxième partie : Courbes principales

Les courbes principales peuvent être vues comme une généralisation non linéaire de la notion de première composante principale. Ce sont des courbes paramétrées de $\mathbb{R}^d$ passant « au milieu » d'une distribution ou d'un nuage de points (voir l'Annexe C pour quelques rappels sur les courbes paramétrées). Un exemple simple de courbe principale est donné dans la Figure 1.3. Comme dans la section précédente, certains détails dans les notations sont omis afin de ne pas noyer le lecteur. Les précisions sont données dans les chapitres correspondants.

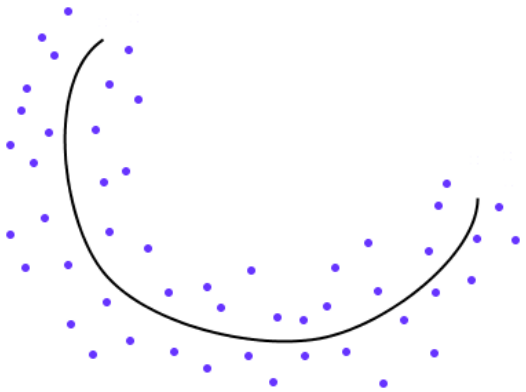


FIGURE 1.3.: Un exemple de courbe principale.

Selon la définition originelle de Hastie et Stuetzle [104], une courbe principale pour un vecteur aléatoire $\mathbf{X}$ de $\mathbb{R}^d$ est une courbe $\mathbf{f}$ vérifiant la propriété d'auto-consistance, c'est-à-dire

$$\mathbf{f}(t) = \mathbb{E}[\mathbf{X}|t_{\mathbf{f}}(\mathbf{X}) = t],$$

où l'indice de projection $t_{\mathbf{f}}$, illustré par la Figure 1.4, est défini par

$$t_{\mathbf{f}}(\mathbf{x}) = \sup\{t, \|\mathbf{x} - \mathbf{f}(t)\| = \inf_{t'} \|\mathbf{x} - \mathbf{f}(t')\|\}.$$

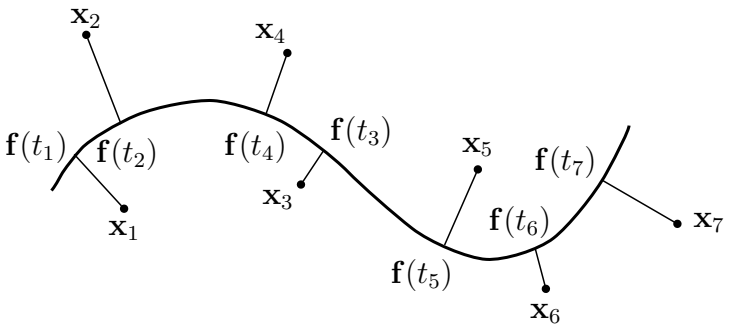


FIGURE 1.4.: Indice de projection. Pour tout i, t_i désigne $t_{\mathbf{f}}(\mathbf{x}_i)$.

La propriété d'auto-consistance s'interprète en disant que chaque point de la courbe est la moyenne des observations qui se projettent sur la courbe au voisinage

de ce point. Différentes définitions ont été proposées à la suite de celle de Hastie et Stuetzle [104]. Elles sont généralement associées à un algorithme fournissant effectivement une approximation de courbe principale. Le plus souvent, la définition est énoncée dans un cadre probabiliste, pour un vecteur aléatoire $\mathbf{X}$ dont on suppose la loi connue, et ensuite adaptée à la situation pratique, où l'on dispose seulement d'un échantillon $\mathbf{X}_1, \ldots, \mathbf{X}_n$ de $\mathbf{X}$. Ainsi, Banfield et Raftery [18] étendent la procédure aux courbes fermées et développent une méthode permettant de réduire le biais d'estimation, tandis que Tibshirani [180] adopte un point de vue semi-paramétrique et définit les courbes principales à partir d'un modèle de mélange. Ce dernier est associé à un algorithme EM (*Expectation-Maximization*, Dempster, Laird et Rubin [66], Xu et Jordan [193]). Dans la définition de Kégl, Krzyżak, Linder et Zeger [117], une courbe principale de longueur L pour $\mathbf{X}$ est une courbe paramétrée minimisant le critère de type moindres carrés

$$\Delta(\mathbf{f}) = \mathbb{E}\left[\inf_t \|\mathbf{X} - \mathbf{f}(t)\|^2\right]$$

parmi toutes les courbes de longueur au plus L. Sandilya et Kulkarni [166] proposent une définition similaire, mais la contrainte de longueur est remplacée par une contrainte sur la courbure. La définition de Delicado [64] est basée sur la généralisation d'une propriété de la première composante principale d'une loi normale multivariée, exprimant le fait que la variance totale de la loi conditionnelle de la variable, sachant qu'elle appartient à un hyperplan, est minimale lorsque celui-ci est orthogonal à la première composante principale. Il est aussi possible de donner une définition des courbes principales reposant sur l'analyse en composantes principales effectuée localement (Einbeck, Tutz et Evers [82]).

Etant donné la diversité de points de vue sur la notion de courbe principale, il nous a semblé important de synthétiser ces différentes idées. Ainsi, le **premier chapitre** propose une mise au point bibliographique sur les courbes principales. Nous présentons en détail plusieurs définitions en tâchant d'expliciter au mieux les liens existant entre elles. Ensuite sont rassemblés les éléments de la littérature ayant trait à la question des courbes principales d'ordre supérieur. En effet, tout comme on s'intéresse en analyse en composantes principales aux composantes principales successives, une loi de probabilité ou un ensemble d'observations peut donner lieu à plusieurs courbes principales. Pour clore le chapitre, nous donnons finalement un aperçu rapide des différentes applications des courbes principales. Ces applications se révèlent extrêmement variées, allant de la reconnaissance de caractères au domaine médical en passant par la géographie ou encore l'écologie.

Dans le **second chapitre**, nous adoptons la définition des courbes principales

reposant sur la minimisation du critère de type moindres carrés

$$\Delta(\mathbf{f}) = \mathbb{E}\left[\inf_t \|\mathbf{X} - \mathbf{f}(t)\|^2\right].$$

Afin de déterminer une courbe principale à partir d'un échantillon $\mathbf{X}_1, \ldots, \mathbf{X}_n$ de $\mathbf{X}$, on cherche à minimiser

$$\Delta_n(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n} \inf_t \|\mathbf{X}_i - \mathbf{f}(t)\|^2,$$

le critère empirique associé à $\Delta(\mathbf{f})$. Observons que la minimisation de $\Delta_n(\mathbf{f})$ sur une classe de courbes trop pauvre ne peut permettre de retrouver correctement la forme des données, tandis qu'autoriser à l'inverse trop de courbes risque d'entraîner un phénomène d'interpolation, comme l'illustre la Figure 1.5. Par exemple, si l'on optimise le critère empirique $\Delta_n(\mathbf{f})$ sur toutes les courbes paramétrées, la courbe principale résultante passera par tous les points des données.

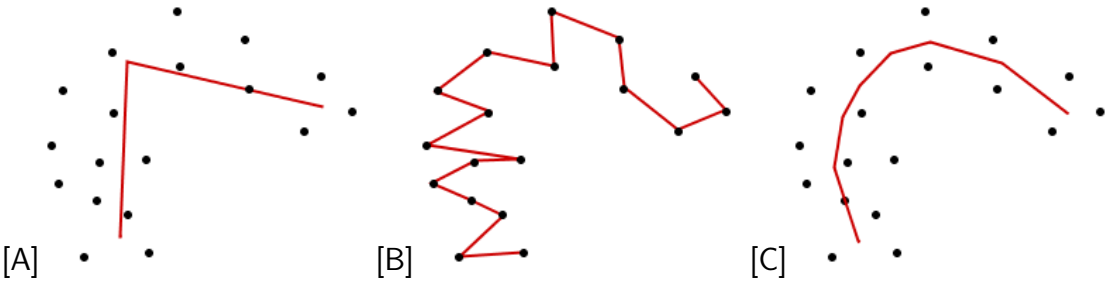


FIGURE 1.5.: Courbes principales obtenues avec [A] une classe contenant peu de courbes, [B] une classe trop riche et [C] une classe convenable.

Par conséquent, pour construire une courbe principale convenable, il est nécessaire de sélectionner une classe de courbes réalisant un bon compromis entre ajustement aux données et un certain degré de régularité. Pour ce faire, nous considérons des collections indexées par la longueur ou la courbure des courbes qu'elles contiennent, ainsi que par le nombre de segments lorsqu'il s'agit de lignes polygonales. L'approche retenue pour choisir ces paramètres est fondée sur la minimisation de critère pénalisé, qui repose sur la théorie de sélection de modèle introduite par Birgé et Massart [35] et Barron, Birgé et Massart [21].

En premier lieu, nous supposons que nous observons des points $\mathbf{X}_1, \ldots, \mathbf{X}_n$ de $\mathbb{R}^d$ tels que

$$\mathbf{X}_i = \mathbf{x}_i^\star + \sigma\boldsymbol{\xi}_i, \quad i = 1, \ldots, n,$$

où les $\mathbf{x}_i^\star$ sont inconnus, les $\boldsymbol{\xi}_i$ sont des vecteurs gaussiens standards indépendants de $\mathbb{R}^d$ et $\sigma > 0$, et nous cherchons une courbe principale correspondant à ces observations. Notons $\overrightarrow{\mathbf{X}}$ le vecteur de $\mathbb{R}^{nd}$ constitué de tous les $\mathbf{X}_i$. Il s'agit d'un vecteur gaussien, et en adoptant la même écriture pour les $\mathbf{x}_i^\star$ et les $\boldsymbol{\xi}_i$, le modèle s'écrit

$$\overrightarrow{\mathbf{X}} = \overrightarrow{\mathbf{x}^\star} + \sigma\, \overrightarrow{\boldsymbol{\xi}}.$$

Considérant des courbes ayant des extrémités fixées F et G, nous introduisons une collection dénombrable de modèles $\{\mathcal{F}_\ell\}_{\ell\in\mathcal{L}}$ de courbes de longueur ℓ. Notre but est de choisir la longueur adéquate. Pour chaque longueur ℓ, la minimisation du risque empirique $\Delta_n(\mathbf{f})$ mène à une certaine courbe $\hat{\mathbf{f}}_\ell$. L'idée consiste alors à sélectionner la longueur $\hat{\ell}$ en minimisant en ℓ le critère $\Delta_n(\hat{\mathbf{f}}_\ell)$ auquel a été ajoutée une pénalité convenable, de nature à empêcher le choix d'une trop grande longueur. En notant $\tilde{\mathbf{x}}_i = \hat{\mathbf{x}}_{i\hat{\ell}}$ les projections des observations sur la courbe résultante $\hat{\mathbf{f}}_{\hat{\ell}}$ et $\|\cdot\|$ la norme euclidienne normalisée de $\mathbb{R}^d$, la qualité de l'estimation peut être mesurée au moyen de la perte

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\|\tilde{\mathbf{x}}_i - \mathbf{x}_i^\star\|^2.$$

Le résultat principal dans ce contexte gaussien s'énonce ainsi : étant donné une famille de poids $\{w_\ell\}_{\ell\in\mathcal{L}}$ vérifiant $\sum_{\ell\in\mathcal{L}} e^{-w_\ell} = \Sigma < +\infty$, si le niveau de bruit σ n'est pas trop grand, il existe des constantes c_1, c_2 telles que pour tout $\eta > 1$, si

$$\mathrm{pen}(\ell) \geq \eta\sigma^2 \left[c_1 \left(\ln\left(\frac{\ell^{1/d}\lambda^{1-1/d}}{\sigma} \right) + c_2 \right) + \frac{4w_\ell}{nd} \right],$$

où $\lambda = \sqrt{\ell^2 - FG^2}$, alors, presque sûrement, il existe un minimiseur $\hat{\ell}$ du critère pénalisé

$$\mathrm{crit}(\ell) = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - \hat{\mathbf{x}}_{i\ell}\|^2 + \mathrm{pen}(\ell).$$

En outre,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\|\tilde{\mathbf{x}}_i - \mathbf{x}_i^\star\|^2 \leq c(\eta)\Big[\inf_{\ell\in\mathcal{L}} \{ d^2(\overrightarrow{\mathbf{x}^\star}, \mathcal{C}_\ell) + \mathrm{pen}(\ell) \} + \frac{\sigma^2}{nd}(\Sigma + 1) \Big],$$

où $\mathcal{C}_\ell \subset \mathbb{R}^{nd}$ dépend de la classe $\mathcal{F}_\ell$ et $d^2(\overrightarrow{\mathbf{x}^\star}, \mathcal{C}_\ell) = \inf_{\overrightarrow{\mathbf{y}}\in\mathcal{C}_\ell} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{x}_i^\star\|^2$.

Dans la seconde section de ce chapitre, nous cherchons à établir un résultat similaire, mais avec une inégalité portant cette fois-ci non plus sur les estimateurs des points échantillonnés sur la courbe, mais sur l'estimateur de la courbe principale elle-même. Dans cette perspective, il faut changer de cadre de travail

et supposer $\mathbf{X}$ presque sûrement bornée. Considérant alors des modèles de lignes polygonales, nous examinons d'abord le cas des courbes principales de longueur bornée de Kégl, Krzyżak, Linder et Zeger [117], puis celui des courbes principales de courbure intégrale bornée de Sandilya et Kulkarni [166].

Dans le premier cas, d'après Kégl, Krzyżak, Linder et Zeger [117], il existe une courbe paramétrée minimisant le critère $\mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})]$ sur toutes les courbes de longueur au plus L, où $\Delta(\mathbf{f}, \mathbf{x}) = \inf_t \|\mathbf{x} - \mathbf{f}(t)\|^2$. Soit

$$\mathbf{f}^\star \in \arg \min_{\mathbf{f}, \mathscr{L}(\mathbf{f}) \leq L} \mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})],$$

où $\mathscr{L}(\mathbf{f})$ désigne la longueur de la courbe $\mathbf{f}$. Pour $k \geq 1$ et $\ell \in \mathcal{L} \subset]0, L]$, le modèle $\mathcal{F}_{k,\ell}$ est défini comme la classe des lignes polygonales à k segments et de longueur au plus ℓ. A chaque modèle $\mathcal{F}_{k,\ell}$ correspond une courbe $\hat{\mathbf{f}}_{k,\ell}$ minimisant le critère empirique $\Delta_n(\mathbf{f})$ sur tous les éléments de ce modèle et notre objectif est de choisir la meilleure courbe principale parmi les estimateurs de la collection $\{\hat{\mathbf{f}}_{k,\ell}\}_{k\geq 1, \ell \in \mathcal{L}}$. Pour ce faire, nous construisons une fonction de pénalité $\text{pen}(k, \ell)$, et les paramètres $(\hat{k}, \hat{\ell})$ retenus sont ceux qui minimisent le critère $\Delta_n(\mathbf{f}_{k,\ell})$ pénalisé par $\text{pen}(k, \ell)$. La qualité de l'estimateur $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k},\hat{\ell}}$ sélectionné est évaluée en contrôlant la perte

$$\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) = \mathbb{E}[\Delta(\tilde{\mathbf{f}}, \mathbf{X}) - \Delta(\mathbf{f}^\star, \mathbf{X})].$$

Plus précisément, l'utilisation de moyennes de Rademacher (Bartlett, Boucheron et Lugosi [22], Koltchinskii [122]) et la majoration d'une intégrale de Dudley [77] constituent les ingrédients principaux menant au résultat suivant : si $\{x_{k,\ell}\}_{k\geq 1, \ell \in \mathcal{L}}$ désigne une famille de poids positifs telle que $\sum_{k\geq 1, \ell \in \mathcal{L}} e^{-x_{k,\ell}} = \Sigma$, pour une pénalité de la forme

$$\text{pen}(k, \ell) \geq \frac{1}{\sqrt{n}} \left[c_1 \sqrt{k} + c_2 \frac{\ell}{\sqrt{k}} + c_0 \right] + \delta^2 \sqrt{\frac{x_{k,\ell}}{2n}},$$

où les constantes c_i ne dépendent que de L, d et δ, la courbe $\tilde{\mathbf{f}}$ obtenue en minimisant le critère $\Delta_n(\mathbf{f})$ pénalisé vérifie l'inégalité

$$\mathbb{E}[\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}})] \leq \inf_{k\geq 1, \ell \in \mathcal{L}} \Big(\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\ell}) + \text{pen}(k, \ell) \Big) + \frac{\delta^2 \Sigma}{2^{3/2}} \sqrt{\frac{\pi}{n}},$$

où $\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\ell}) = \inf_{\mathbf{f} \in \mathcal{F}_{k,\ell}} \mathcal{D}(\mathbf{f}^\star, \mathbf{f})$.

Grâce à un résultat géométrique reliant la courbure intégrale d'une courbe paramétrée à sa longueur (Alexandrov et Reshetnyak [7]), il est possible de mettre en

œuvre une approche similaire dans le contexte des courbes principales de Sandilya et Kulkarni [166], où la contrainte est mise sur la courbure. Les modèles $\mathcal{F}_{k,\kappa}$ sont ici indexés par le nombre de segments $k \geq 1$ et la courbure intégrale $\kappa \in \mathcal{K}$. De même que la longueur d'une ligne polygonale est la somme des segments qui la composent, la courbure intégrale est la somme des angles aux sommets. Les courbes $\mathbf{f}^\star$ et $\tilde{\mathbf{f}}$ peuvent être définies dans ce cadre par analogie avec le cas précédent. Alors, en notant $\{x_{k,\kappa}\}_{k\geq 1,\kappa\in\mathcal{K}}$ une famille de poids positifs, si

$$\mathrm{pen}(k,\kappa) \geq \frac{\delta^2}{\sqrt{n}}\left[c_1\sqrt{k} + +c_2 \max\left(\frac{\zeta(\kappa)}{\sqrt{k}}, \sqrt{k \ln \frac{\zeta(\kappa)}{k}}\right) + c_0 + \sqrt{\frac{x_{k,\kappa}}{2}}\right],$$

où ζ est une fonction croissante de κ et les c_i ne dépendent que de d, nous obtenons

$$\mathbb{E}[\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}})] \leq \inf_{k\geq 1,\kappa\in\mathcal{K}} \Big(\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\kappa}) + \mathrm{pen}(k,\kappa)\Big) + \frac{\delta^2\Sigma}{2^{3/2}}\sqrt{\frac{\pi}{n}},$$

où $\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\kappa}) = \inf_{\mathbf{f}\in\mathcal{F}_{k,\kappa}} \mathcal{D}(\mathbf{f}^\star, \mathbf{f})$.

Finalement, deux algorithmes permettant de calculer une courbe principale d'un point de vue pratique ont été implémentés et appliqués à des données simulées ainsi qu'à des données réelles. Pour évaluer les constantes intervenant dans les fonctions de pénalité, nous avons choisi d'employer l'heuristique de pente, méthode de calibration de pénalités introduite par Birgé et Massart [37] et étendue par Arlot et Massart [10]. Le premier algorithme, qui estime à la fois $\hat{k}$ et $\hat{\ell}$, est basé sur une version bivariée de cette heuristique, tandis que dans le second, adaptation du Polygonal Line Algorithm de Kégl, Krzyżak, Linder et Zeger [117], $\hat{k}$ est sélectionné grâce à l'interface CAPUSHE de Baudry, Maugis et Michel [25], la courbure étant contrôlée localement.

La Figure 1.6 présente les sorties des deux algorithmes pour un exemple d'image binaire de chiffre faisant partie de la base de données NIST Special Database 19 (`http://www.nist.gov/srd/nistsd19.cfm`) et la Figure 1.8 montre les courbes principales résultantes pour une zone sismique extraite du jeu de données de tremblements de terre provenant de l'institut United States Geological Survey (`http://earthquake.usgs.gov/research/data/centennial.php`) reproduit dans la Figure 1.7 (Engdahl et Villaseñor [83]).

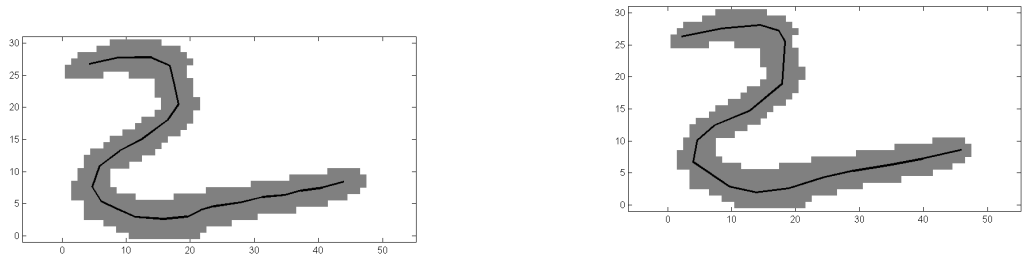


FIGURE 1.6.: Courbes principales sélectionnées pour le chiffre 2. [A] Méthode **MS1** : $\hat{k} = 23$, $\hat{\ell} = 80$. [B] Méthode **MS2** : $\hat{k} = 17$.

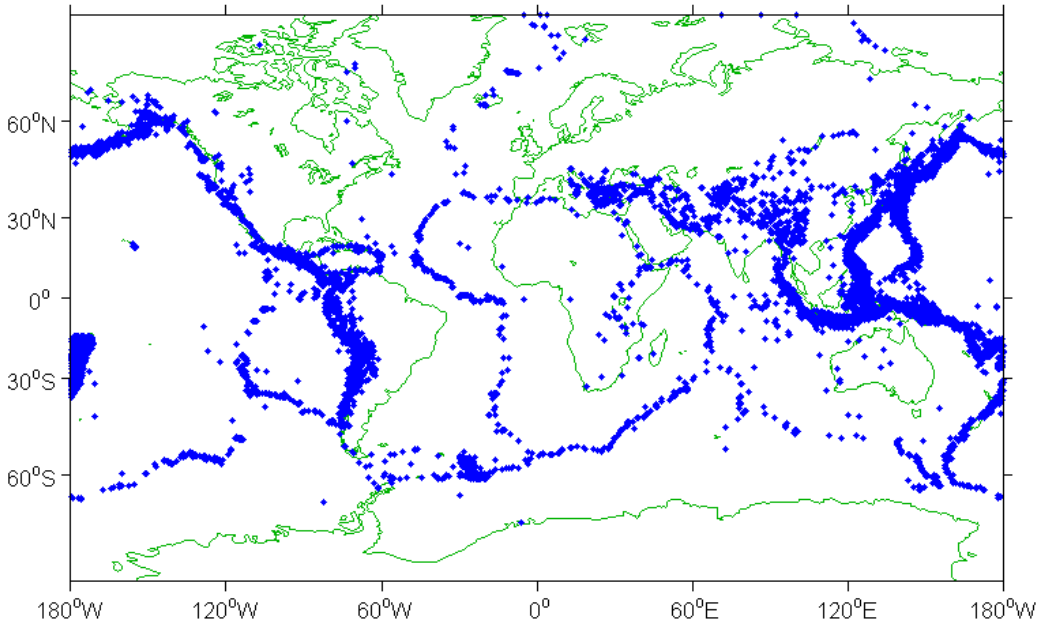
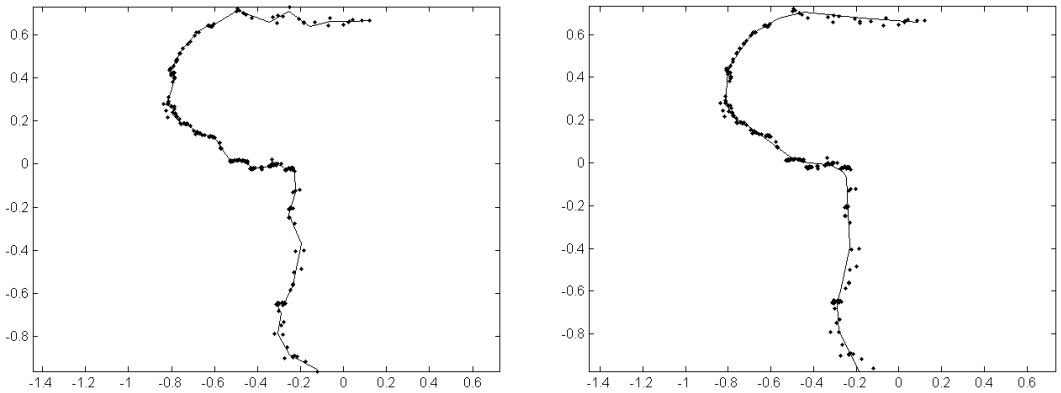


FIGURE 1.7.: Impacts sismiques à la surface du globe.



[A] [B]

FIGURE 1.8.: Courbes principales sélectionnées pour une zone sismique de l'océan Atlantique. [A] Méthode **MS1** : $\hat{k} = 55$, $\hat{\ell} = 31$. [B] Méthode **MS2** : $\hat{k} = 30$.

Première partie .

Quantification et clustering

1. Quantification et clustering avec des divergences de Bregman*

Sommaire

*Ce chapitre a donné lieu à un article repris dans l'Annexe D, publié dans la revue *Journal of Multivariate Analysis*.

1.1. Introduction

Quantification et clustering sont deux aspects d'une même question, dont les applications concernent des domaines aussi variés que la biologie, l'informatique ou les sciences sociales. Le premier mot correspond à la formulation probabiliste du problème, et le second au point de vue statistique. En compression de données et théorie de l'information, la quantification équivaut à une compression « avec perte » : il s'agit de remplacer des données par une représentation efficace et compacte, à partir de laquelle il est ensuite possible de les reconstruire, avec une certaine précision, évaluée par un critère d'erreur. Plus formellement, une variable aléatoire X à valeurs dans un espace $\mathcal{X}$ est représentée par $q(X)$, où l'application q envoie $\mathcal{X}$ dans un sous-ensemble fini d'éléments de $\mathcal{X}$. Une fonction appelée distorsion permet de contrôler l'erreur due à ce remplacement. Cette théorie est exposée de manière détaillée par Gersho et Gray [94], Graf et Luschgy [98] et Linder [131]. A la quantification correspond dans le contexte statistique l'une des formes les plus répandues de clustering, la méthode des k-means initiée par Lloyd [132]. Le clustering consiste, à partir d'un amas de données ($X_1, \ldots, X_n$ supposées indépendantes et distribuées suivant la loi de X), à former des groupes de telle manière que les données soient très semblables entre elles et que les différents groupes soient aussi séparés que possible (Duda, Hart et Stork [76]). Ces groupes jouent le rôle des éléments choisis comme représentants en quantification. En effectuant une étape de clustering, on espère retirer des données certaines informations, en dégager des caractéristiques, y repérer des phénomènes sous-jacents. C'est une méthode d'apprentissage non supervisé, puisque contrairement à la classification supervisée qui consiste à classer des observations dans des groupes connus à l'avance, ce qui revient à attribuer à chaque donnée une étiquette, il s'agit ici de regrouper les données qui sont proches les unes des autres en un certain sens, sans disposer de classes préexistantes dans lesquelles les répartir.

Une notion de distance intervient en quantification et en clustering, à travers une fonction de deux variables appelée mesure de distorsion. La plupart du temps, c'est le carré de la distance euclidienne de $\mathbb{R}^d$ qui est utilisé. Cependant, dans de nombreux domaines, les données à analyser sont des courbes. En particulier, il est souvent utile d'étudier l'évolution d'une quantité au cours du temps. En biologie ou dans le domaine médical, on peut par exemple penser à une courbe de croissance, ou à l'évolution d'un taux sanguin. On peut également songer à un enregistrement vocal ou encore, en économie, au cours d'un actif ou à la variation d'un taux d'intérêt. D'autres courbes reflètent une variation dans l'espace ou donnent la forme d'un objet. Bien que les données fonctionnelles se présentent en réalité comme un ensemble de valeurs discrètes qui ont été enregistrées, mesurées, comme on pourrait en théorie disposer d'une infinité de points qui soient aussi proches les uns

des autres que l'on veut, il est intéressant de tenir compte de leur nature fonctionnelle (Ramsay et Silverman [159]). Il est donc souhaitable d'envisager d'autres notions de distance, adaptées aux données de grande dimension. Dans le cadre du clustering, Biau, Devroye et Lugosi [32] et Cadre et Paris [44] considèrent le cas d'une norme hilbertienne au carré, et Laloë [125] celui des normes L^1. Luschgy et Pagès [136, 137] étudient la quantification de processus gaussiens et de diffusions avec une norme L^p. Par ailleurs, même en dimension finie, la distance euclidienne n'est pas toujours adéquate. Dereich et Vormoor [68] traitent ainsi le problème de la quantification avec une norme d'Orlicz, tandis que Teboulle, Berkhin, Dhillon, Guan et Kogan [120] remarquent qu'une mesure de distorsion non symétrique peut être plus appropriée pour certains types de données.

En 1967, Bregman [40] a introduit une classe de mesures de distorsion indexées par des fonctions ϕ strictement convexes. Dans $\mathbb{R}^d$, une divergence de Bregman d_ϕ est de la forme

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle,$$

avec $\langle\cdot,\cdot\rangle$ le produit scalaire de $\mathbb{R}^d$ et $\nabla\phi(y)$ le gradient de ϕ en y. Avec $\phi(x) = \|x\|^2$, on retrouve le carré de la distance euclidienne. Cette définition se généralise au cadre des espaces de Hilbert, ainsi qu'à celui des espaces de Banach en écrivant

$$d_\phi(x, y) = \phi(x) - \phi(y) - D_y\phi(x - y),$$

avec $D_y\phi$ la différentielle de ϕ en y (Alber et Butnariu [4], Frigyik, Srivastava et Gupta [89]). Une divergence de Bregman ne correspond pas toujours à une vraie distance. En effet, elle peut ne pas être symétrique et l'on ne dispose pas nécessairement de l'inégalité triangulaire. En choisissant correctement ϕ, outre la distance euclidienne au carré, on obtient entre autres comme cas particuliers de divergences de Bregman la distance de Mahalanobis, la distance de Kullback-Leibler ou le carré d'une norme L^2.

Banerjee, Merugu, Dhillon et Ghosh [17] établissent une relation entre divergences de Bregman et lois de la famille exponentielle. De même que la distance euclidienne convient dans le cas de la loi normale, une autre divergence de Bregman est en lien avec une autre loi de la famille exponentielle. Ces auteurs montrent que l'algorithme de clustering des k-means (Lloyd [132]) se généralise à ces divergences et proposent donc de les utiliser pour faire du clustering. En fait, Wu, Xiong, Chen et Zhou [192] caractérisent les mesures de distorsion en dimension finie pour lesquelles cet algorithme fonctionne et distinguent les « distances k-means » de type I, qui sont les divergences de Bregman, et celles de type II, obtenues pour une fonction ϕ convexe mais non strictement convexe. Nielsen, Boissonnat et Nock [149] s'intéressent aux nombreuses propriétés des divergences de Bregman en dimension

finie, notamment à leurs caractéristiques géométriques, et observent que ces objets possèdent des applications à la géométrie algorithmique et à l'apprentissage. Les divergences de Bregman sont également liées au contexte bayésien de la méthode du maximum d'entropie sur la moyenne (Csiszár, Gamboa et Gassiat [57], Gamboa, Loubes et Rochet [92]). Les propriétés des divergences de Bregman fonctionnelles sont étudiées par Frigyik, Srivastava et Gupta [89]. La définition est déjà donnée dans le cas fonctionnel par Alber et Butnariu [4], qui développent une importante notion de projection basée sur ces divergences. Notons que Basseville propose dans [24] une synthèse bibliographique sur les divergences de Bregman et d'autres types de mesure de distance (f-divergences et α-divergences).

Puisque les divergences de Bregman englobent un large éventail de mesures de distorsion, en dimension finie ou infinie, et que l'algorithme des k-means s'étend à ces divergences, nous allons nous intéresser d'un point de vue théorique au problème de la quantification et du clustering avec des divergences de Bregman dans le cadre d'un espace de Banach (voir l'Annexe 1.8.1 pour un rappel sur les variables aléatoires à valeurs dans un espace de Banach). Nous considérons une variable aléatoire X de loi μ à valeurs dans un espace de Banach réflexif et séparable afin de disposer de certaines caractéristiques topologiques propres à ces espaces. Notons

$$W(\mu, q) = \mathbb{E}[d_\phi(X, q(X))]$$

la distorsion et, si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes de même loi que X et μ_n désigne la mesure empirique,

$$W(\mu_n, q) = \frac{1}{n} \sum_{i=1}^{n} d_\phi(X_i, q(X_i))$$

le critère empirique associé.

Tout d'abord, nous rappelons dans la Section 1.2 en quoi consistent la quantification et la question liée du clustering, en introduisant le cadre et les notations qui s'y rattachent. Nous présentons ensuite les divergences de Bregman dans la Section 1.3. Nous donnons la définition de cette classe de mesures de distorsion en dimension finie et infinie et nous développons quelques exemples, avant de rappeler des propriétés utiles de ces divergences. La Section 1.4.2 établit des conditions d'existence d'un minimum de la distorsion $W(\mu, q)$ en dimension finie et infinie, et de même pour la distorsion empirique $W(\mu_n, q)$. Puis, dans la Section 1.5, nous nous intéressons à la convergence de la distorsion. Il s'agit de voir si, $q_n^\star$ étant un minimiseur de la distorsion empirique $W(\mu_n, q)$, $W(\mu, q_n^\star)$ s'approche du minimum de $W(\mu, q)$ lorsque le nombre n d'observations devient très grand. Nous obtenons sous certaines conditions les convergence presque sûre et L^1 de $W(\mu, q_n^\star)$

vers $W(\mu, q)$, ainsi qu'une borne non-asymptotique qui ne dépend pas de la dimension de l'espace. Enfin, nous présentons quelques simulations dans la Section 1.6.

1.2. Quantification

Cette section expose le principe de la quantification, en lien avec le clustering des k-means. Pour commencer, nous donnons la définition d'un quantificateur.

Définition 1.2.1 (Quantificateur)**.** *Soit $k \geq 1$ un entier. Un k-quantificateur est une application borélienne $q : \mathcal{X} \to \mathbf{c}$, où $\mathbf{c} = \{c_1, \ldots, c_\ell\}$, $\ell \leq k$, est un sous-ensemble de $\mathcal{X}$ appelé table de codage.*

Remarque 1.2.1. Les éléments $c_1, \ldots, c_\ell$ seront également appelés les centres associés au quantificateur q.

Tout élément $x \in \mathcal{X}$ est représenté par un unique $\hat{x} = q(x) \in \mathbf{c}$. Pour mesurer l'erreur commise en représentant la variable aléatoire X par $q(X)$, on s'intéresse à $\mathscr{d}(X, q(X))$, où $\mathscr{d}$ est une fonction mesurable telle que $\mathscr{d}(x, y) \geq 0$ pour tout couple $(x, y) \in \mathcal{X}^2$. Cette fonction est appelée mesure de distorsion.

Définition 1.2.2 (Distorsion)**.** *La distorsion de q quantifiant X est la quantité*

$$W(\mu, q) = \mathbb{E}[\mathscr{d}(X, q(X))] = \int_{\mathcal{X}} \mathscr{d}(x, q(x)) \mathrm{d}\mu(x). \tag{1.1}$$

Chercher un bon k-quantificateur revient à minimiser cette distorsion.

Définition 1.2.3 (Quantificateur optimal)**.** *Soit Q_k l'ensemble de tous les k-quantificateurs, et soit*

$$W_k^\star(\mu) = \inf_{q \in Q_k} W(\mu, q).$$

Un quantificateur $q^\star \in Q_k$ est dit optimal si $W(\mu, q^\star) = W_k^\star(\mu)$.

Lorsqu'il n'y a pas d'ambiguïté, on écrira $W^\star(\mu)$, sans l'indice k. Pour représenter X de manière efficace et précise, on souhaite disposer du meilleur quantificateur possible, pour lequel l'erreur soit la plus petite possible. C'est pourquoi nous allons nous intéresser dans la suite à la question de l'existence d'un quantificateur optimal.

Remarque 1.2.2. Chaque k-quantificateur est déterminé par sa table de codage $\{c_1, \ldots, c_\ell\}$ et une partition de $\mathcal{X}$ en cellules $S_j = \{x : q(x) = c_j\}$, $j = 1, \ldots, \ell$, suivant la relation

$$q(x) = c_j \Leftrightarrow x \in S_j.$$

Un quantificateur peut donc être défini par sa table de codage et sa partition.

Envisagé sous sa version k-means, le problème du clustering, qui consiste à former des groupes d'observations de telle façon que les données soient très semblables à l'intérieur des groupes, et que les différents groupes soient aussi séparés que possible, est très proche de celui de la quantification. En effet, dans un contexte statistique, on ne connaît pas la loi μ, mais on dispose de n observations $X_1, \dots, X_n$ supposées être des variables aléatoires indépendantes toutes de loi μ. Soit μ_n la mesure empirique associée à $X_1, \dots, X_n$, définie par

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

pour tout A borélien de $\mathcal{X}$. On introduit la distorsion empirique

$$W(\mu_n, q) = \frac{1}{n} \sum_{i=1}^{n} d(X_i, q(X_i)), \tag{1.2}$$

qui est la distorsion (1.1) pour la loi μ_n. Classer les données en groupes revient à chercher un quantificateur $q_n^\star$ optimal par rapport à la distorsion empirique (1.2). Un tel quantificateur $q_n^\star$, défini par $W(\mu_n, q_n^\star) = W_k^\star(\mu_n)$, est appelé quantificateur empirique optimal.

A ce stade, il faut décider de quelle manière mesurer la proximité, c'est-à-dire fixer la fonction d. Le choix de cette notion de distance fait l'objet de la section suivante.

1.3. Divergences de Bregman

Le cadre de la quantification étant posé, la question qui se pose à présent est celle du choix de la fonction d. Souvent, c'est le carré de la distance euclidienne qui est utilisé. Néanmoins, cette distance n'est pas adaptée à tous les types de données, et en particulier ne permet pas de traiter le cas de la dimension infinie. Nous allons utiliser comme mesures de distorsion les divergences de Bregman, une famille de mesures de distorsion qui a été introduite en 1967 par Bregman [40]. Le carré de la distance euclidienne et d'autres mesures de distorsion usuelles en sont des cas particuliers. Dans un premier temps, nous donnons la définition des divergences de Bregman en dimension finie. Après avoir explicité quelques exemples, nous verrons comment la modifier pour obtenir des divergences entre fonctions. Le lecteur est invité à se reporter à l'Annexe 1.8.2 pour quelques rappels liés à la notion de gradient et de différentielle, et à l'Annexe 1.8.3 pour les définitions utiles sur les fonctions convexes et la topologie faible.

1.3.1. Définition et exemples dans $\mathbb{R}^d$

Définir une divergence de Bregman nécessite d'introduire la notion d'intérieur relatif d'un ensemble convexe (voir Rockafellar [161, Section 7]).

Définition 1.3.1. *On appelle intérieur relatif d'un convexe non vide $\mathcal{C}$ de $\mathbb{R}^d$, noté $ir(\mathcal{C})$, l'intérieur de $\mathcal{C}$ relativement à l'enveloppe affine de $\mathcal{C}$ (plus petit sous-espace affine contenant $\mathcal{C}$).*

Remarque 1.3.1. Alors que l'intérieur d'un convexe est souvent vide, l'intérieur relatif d'un convexe non vide $\mathcal{C}$ de $\mathbb{R}^d$ est non vide (et a la même dimension que $\mathcal{C}$), d'où l'intérêt d'introduire cette notion.

La Figure 1.1 donne l'exemple d'un convexe d'intérieur non vide dans $\mathbb{R}^2$ qui est d'intérieur vide dans $\mathbb{R}^3$. En effet, cet objet peut contenir une boule de $\mathbb{R}^2$, mais pas une boule de $\mathbb{R}^3$.

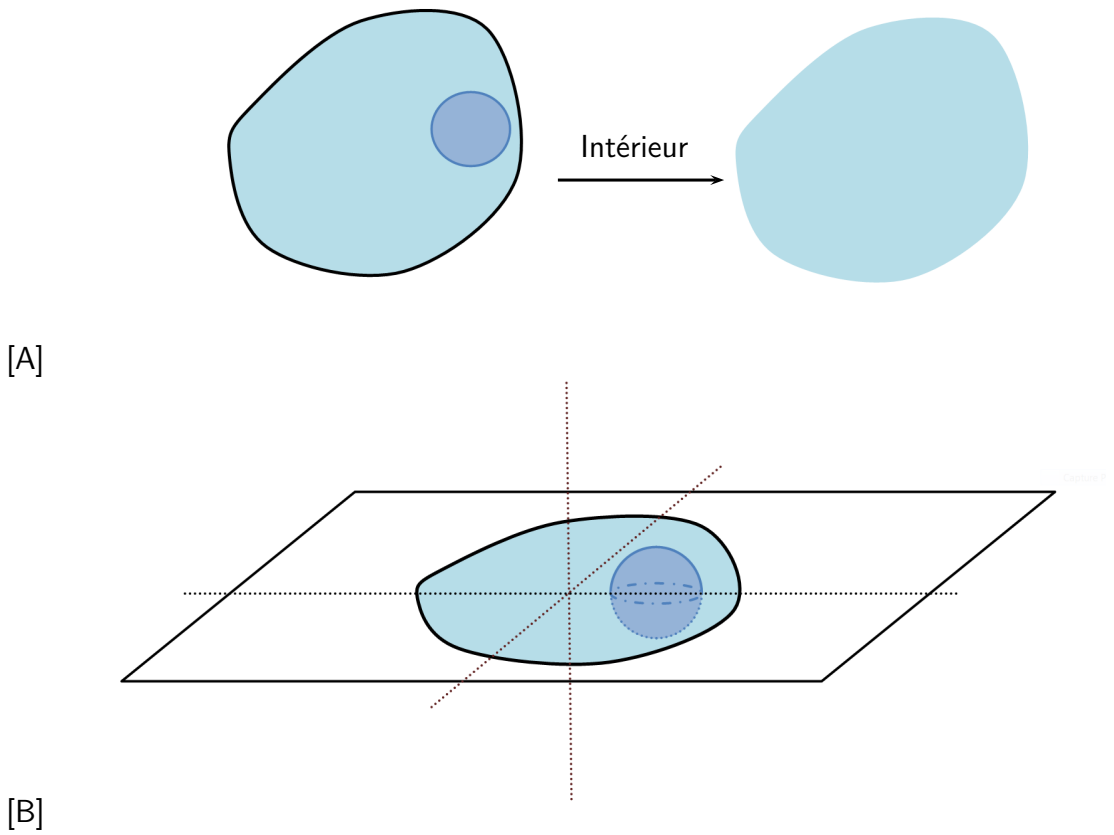


FIGURE 1.1.: Un convexe qui est [A] d'intérieur non vide dans $\mathbb{R}^2$ et [B] d'intérieur vide dans $\mathbb{R}^3$.

Nous noterons $\partial\mathcal{C}$ la frontière relative du convexe $\mathcal{C}$, c'est-à-dire le complémentaire de $ir(\mathcal{C})$ dans son adhérence $\overline{\mathcal{C}}$.

Définition 1.3.2. *Soient $\mathcal{C} \subset \mathbb{R}^d$ un convexe et $\phi : \mathcal{C} \to \mathbb{R}$ une fonction strictement convexe, différentiable sur $ir(\mathcal{C})$. La divergence de Bregman associée*

$$d_\phi : \mathcal{C} \times ir(\mathcal{C}) \to [0, +\infty[$$

est définie par

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle,$$

où $\langle \cdot, \cdot \rangle$ désigne le produit scalaire de $\mathbb{R}^d$ et $\nabla\phi(y)$ le gradient de ϕ au point y.

Remarque 1.3.2. On peut noter que la fonction ϕ est de classe C^1 sur $ir(\mathcal{C})$, car elle y est convexe et différentiable (voir par exemple Phelps [154]). De même, pour tout $y \in ir(\mathcal{C})$, les fonctions $x \mapsto d_\phi(x, y)$, dont nous verrons plus loin qu'elles sont convexes, sont C^1 sur $ir(\mathcal{C})$.

La distance euclidienne au carré et d'autres distances usuelles sont des cas particuliers de divergence de Bregman. Voici tout d'abord quelques exemples de divergences obtenues lorsque $d = 1$, mentionnés en particulier par Banerjee *et al.* [17] et Nielsen *et al.* [149]. Dans ces exemples ainsi que tous les suivants (en dimension plus grande), nous vérifions que ϕ possède les propriétés requises, en particulier la stricte convexité, et présentons les calculs qui mènent à la divergence de Bregman $d_\phi(\cdot, \cdot)$ associée.

Exemple 1.3.1

1. Distance euclidienne au carré en dimension 1. Soient $\mathcal{C} = \mathbb{R}$ et ϕ définie par $\phi(x) = x^2$. On a $\phi'(x) = 2x$ et $\phi''(x) = 2$, d'où la différentiabilité et la stricte convexité de ϕ. La divergence de Bregman associée est définie, pour tout $(x, y) \in \mathbb{R}^2$, par

$$\begin{aligned} d_\phi(x, y) &= x^2 - y^2 - 2y(x - y) \\ &= \boxed{(x - y)^2}. \end{aligned}$$

   On retrouve donc, en dimension 1, le carré de la distance euclidienne.

2. Avec un exposant $\alpha \geq 2$. Soit $\mathcal{C} = \mathbb{R}^+$. On peut choisir plus généralement ϕ telle que $\phi(x) = x^\alpha$ avec $\alpha \geq 2$ un entier. On a $\phi'(x) = \alpha x^{\alpha-1}$ et $\phi''(x) = \alpha(\alpha - 1)x^{\alpha-2}$, donc ϕ est strictement convexe. Soit $(x, y) \in \mathbb{R}^+ \times \mathbb{R}^{+*}$. On obtient la divergence de Bregman

$$\begin{aligned} d_\phi(x, y) &= x^\alpha - y^\alpha - \alpha y^{\alpha-1}(x - y) \\ &= \boxed{x^\alpha + (\alpha - 1)y^\alpha - \alpha x y^{\alpha-1}}. \end{aligned}$$

3. Distance de Kullback-Leibler généralisée en dimension 1. Prenons $\mathcal{C} = \mathbb{R}^+$, et ϕ définie par $\phi(x) = x \ln x$. On a $\phi'(x) = 1 + \ln x$ et $\phi''(x) = \frac{1}{x}$. Soit $(x, y) \in \mathbb{R}^+ \times \mathbb{R}^{+*}$. La divergence de Bregman obtenue est la suivante :

$$\begin{aligned} d_\phi(x, y) &= x \ln x - y \ln y - (x - y)(\ln y + 1) \\ &= \boxed{x \ln \frac{x}{y} - (x - y)}. \end{aligned}$$

4. Perte logistique. Soient $\mathcal{C} = [0, 1]$ et ϕ définie par

$$\phi(x) = x \ln x + (1 - x) \ln(1 - x).$$

On a $\phi'(x) = \ln(x) - \ln(1 - x)$ et $\phi''(x) = \frac{1}{x} + \frac{1}{1-x}$. Ainsi, pour tout $(x, y) \in [0, 1]\times]0, 1[$,

$$\begin{aligned} d_\phi(x, y) &= x \ln x + (1 - x) \ln(1 - x) - y \ln y - (1 - y) \ln(1 - y) \\ &\quad - (x - y)(\ln y - \ln(1 - y)) \\ &= \boxed{x \ln \frac{x}{y} + (1 - x) \ln \left(\frac{1 - x}{1 - y} \right)}. \end{aligned}$$

Plus généralement, si $N \in \mathbb{N}^*$ et $\mathcal{C} = [0, N]$,

$$\phi(x) = x \ln x + (N - x) \ln(N - x)$$

conduit, pour tout $(x, y) \in [0, N]\times]0, N[$, à

$$d_\phi(x, y) = \boxed{x \ln \frac{x}{y} + (N - x) \ln \left(\frac{N - x}{N - y} \right)}.$$

5. Distance de Itakura-Saito. Soit $\mathcal{C} = \mathbb{R}^{+*}$ et soit ϕ l'entropie de Burg, c'est-à-dire $\phi(x) = -\ln x$. On a $\phi'(x) = -\frac{1}{x}$, $\phi''(x) = \frac{1}{x^2}$, et, pour tout $(x, y) \in (\mathbb{R}^{+*})^2$,

$$\begin{aligned} d_\phi(x, y) &= -\ln x + \ln y + \frac{1}{y}(x - y) \\ &= \boxed{\frac{x}{y} - \ln \frac{x}{y} - 1}. \end{aligned}$$

6. Exponentielle. En prenant $\mathcal{C} = \mathbb{R}$ et ϕ définie par $\phi(x) = e^x$, on a la divergence de Bregman sur $\mathbb{R}^2$

$$d_\phi(x, y) = \boxed{e^x - e^y - (x - y)e^y}.$$

7. Divergence de type Hellinger. Soient $\mathcal{C} = [-1, 1]$ et ϕ définie par $\phi(x) = -\sqrt{1-x^2}$. On a $\phi'(x) = \frac{x}{\sqrt{1-x^2}}$ et $\phi''(x) = \frac{1}{(1-x^2)^{3/2}}$, ce qui montre que ϕ est strictement convexe. Appelée divergence de type Hellinger par Nielsen *et al.* [149], la divergence associée est, pour tout $(x, y) \in [-1, 1] \times]-1, 1[$,

$$\begin{aligned} d_\phi(x, y) &= -\sqrt{1-x^2} + \sqrt{1-y^2} - (x-y)\frac{y}{\sqrt{1-y^2}} \\ &= \boxed{\frac{1-xy}{\sqrt{1-y^2}} - \sqrt{1-x^2}}. \end{aligned}$$

La fonction de perte suivante est « presque » une divergence de Bregman.

8. *Hinge Loss*. Soit $\mathcal{C} = \mathbb{R}$, et soit ϕ la fonction valeur absolue : $\phi(x) = |x|$. La fonction ϕ n'est pas dérivable en 0, et elle est convexe, mais pas strictement convexe. On a

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

En écrivant, pour tout $(x, y) \in \mathbb{R}^2$,

$$\begin{aligned} d_\phi(x, y) &= |x| - |y| - \phi'(y)(x-y) \\ &= \boxed{(-2 \text{ signe}(y)x)_+}, \end{aligned}$$

où $a_+ = \max(0, a)$, on obtient une perte *Hinge Loss*, qui n'est pas une vraie divergence de Bregman, puisque la fonction valeur absolue n'est pas strictement convexe.

Dans les exemples suivants, l'espace ambiant est $\mathbb{R}^d$. Notons qu'à partir des divergences de Bregman unidimensionnelles, nous pouvons obtenir des divergences de Bregman sur $\mathbb{R}^d$ en sommant sur les coordonnées.

Exemple 1.3.2 1. Distance euclidienne au carré. Soient $\mathcal{C} = \mathbb{R}^d$ et ϕ définie par $\phi(x) = \|x\|^2$. La fonction ϕ est strictement convexe et différentiable sur $\mathbb{R}^d$. Pour tout $(x, y) \in (\mathbb{R}^d)^2$,

$$\begin{aligned} d_\phi(x, y) &= \|x\|^2 - \|y\|^2 - \langle x - y, \nabla\phi(y)\rangle \\ &= \|x\|^2 - \|y\|^2 - \langle x - y, 2y\rangle \\ &= \boxed{\|x - y\|^2}. \end{aligned}$$

On obtient ainsi le carré de la distance euclidienne.

2. Distance de Mahalanobis. Soient $\mathcal{C} = \mathbb{R}^d$, A une matrice symétrique définie positive et ϕ définie par $\phi(X) = {}^tXAX$, où ${}^tX$ désigne le vecteur transposé de X. Calculons la différentielle de ϕ au point Y notée $D_Y\phi$. On a

$$\begin{aligned}\phi(Y+H) - \phi(Y) &= {}^t(Y+H)A(Y+H) - {}^tYAY \\ &= {}^tYAH + {}^tHAY + o(H).\end{aligned}$$

Donc $D_Y\phi(H) = {}^tYAH + {}^tHAY$ et, pour tout $(x, y) \in (\mathbb{R}^d)^2$,

$$\begin{aligned}d_\phi(X,Y) &= {}^tXAX - {}^tYAY - {}^tYA(X-Y) - {}^t(X-Y)AY \\ &= \boxed{{}^t(X-Y)A(X-Y)}.\end{aligned}$$

Lorsque la matrice A est l'inverse d'une matrice de covariance, la divergence de Bregman obtenue est appelée distance de Mahalanobis.

3. Distance de Kullback-Leibler entre deux mesures positives discrètes. Soit $\mathcal{C} = (\mathbb{R}^+)^d$. Un élément $x \in \mathcal{C}$ est un vecteur constitué de d composantes positives. La fonction ϕ définie par $\phi(x) = \sum_{\ell=1}^d x_\ell \ln x_\ell$ est différentiable sur $ir(\mathcal{C})$ et strictement convexe. La divergence de Bregman obtenue avec ce choix de ϕ est la distance de Kullback-Leibler généralisée ou I-divergence : pour tout $(x, y) \in \mathcal{C} \times ir(\mathcal{C})$,

$$\begin{aligned}d_\phi(x,y) &= \sum_{\ell=1}^d x_\ell \ln x_\ell - \sum_{\ell=1}^d y_\ell \ln y_\ell - \langle x-y, \nabla\phi(y)\rangle \\ &= \sum_{\ell=1}^d x_\ell \ln x_\ell - \sum_{\ell=1}^d y_\ell \ln y_\ell - \sum_{\ell=1}^d (x_\ell - y_\ell)(\ln y_\ell + 1) \\ &= \boxed{\sum_{\ell=1}^d x_\ell \ln \frac{x_\ell}{y_\ell} - \sum_{\ell=1}^d (x_\ell - y_\ell)}.\end{aligned}$$

4. Distance de Kullback-Leibler entre deux mesures de probabilité discrètes. Soit $\mathcal{C} = (\mathbb{R}^+)^d$. On définit à nouveau ϕ par $\phi(x) = \sum_{\ell=1}^d x_\ell \ln x_\ell$. Si l'on se restreint au simplexe $\mathcal{S}_{d-1}$ de dimension $d-1$, c'est-à-dire aux vecteurs constitués de d composantes positives $x_1, \ldots, x_d$ telles que $\sum_{\ell=1}^d x_\ell = 1$, qui correspondent à des mesures de probabilité discrètes, la divergence de Bregman obtenue est la (vraie) distance de Kullback-Leibler : pour tout

$(x, y) \in [\mathcal{C} \times ir(\mathcal{C})] \cap \mathcal{S}^2_{d-1}$,

$$\begin{aligned} d_\phi(x, y) &= \sum_{\ell=1}^{d} x_\ell \ln x_\ell - \sum_{\ell=1}^{d} y_\ell \ln y_\ell - \langle x - y, \nabla\phi(y)\rangle \\ &= \sum_{\ell=1}^{d} x_\ell \ln x_\ell - \sum_{\ell=1}^{d} y_\ell \ln y_\ell - \sum_{\ell=1}^{d} (x_\ell - y_\ell)(\ln y_\ell + 1) \\ &= \boxed{\sum_{\ell=1}^{d} x_\ell \ln \frac{x_\ell}{y_\ell}}. \end{aligned}$$

1.3.2. Cas fonctionnel

Pour obtenir également des mesures de distorsion qui puissent s'appliquer à des courbes, nous considérons ici des divergences de Bregman en dimension infinie (Alber et Butnariu [4], Frigyik, Srivastava et Gupta [89]). La définition suivante est la généralisation naturelle de la définition d'une divergence de Bregman en dimension finie : le produit scalaire entre le gradient de ϕ en y et le vecteur $x - y$ est remplacé par la différentielle de Fréchet de ϕ au point y appliquée à $x - y$.

L'intérieur relatif $ir(\mathcal{C})$ peut être défini de manière similaire au cas de la dimension finie (Définition 1.3.1), en prenant l'adhérence de l'enveloppe affine du convexe $\mathcal{C}$. Il existe également d'autres notions d'intérieur relatif dans un espace de Banach de dimension infinie, appelées pseudo-intérieur relatif et quasi-intérieur relatif (voir Borwein et Goebel [38]).

Définition 1.3.3 (Divergence de Bregman fonctionnelle)**.** *Soit E un espace de Banach séparable, $\mathcal{C} \subset E$ un convexe et soit $\phi : \mathcal{C} \to \mathbb{R}$ strictement convexe, de classe C^2 sur $ir(\mathcal{C})$. La divergence de Bregman associée est définie par*

$$d_\phi(x, y) = \phi(x) - \phi(y) - D_y\phi(x - y)$$

avec $D_y\phi$ la différentielle de ϕ en y.

Observons que si nous travaillons dans $(E, \langle\cdot,\cdot\rangle)$, espace de Hilbert, la même notation qu'en dimension finie peut être utilisée.

Remarque 1.3.3. Il paraît raisonnable de construire des divergences de Bregman fonctionnelles en intégrant certaines divergences de Bregman unidimensionnelles. Les divergences de Bregman *ponctuelles* (Jones et Byrne [111], Csiszár [56]) constituent un cas particulier de divergences de Bregman fonctionnelles qui sont justement de ce type. Pour m une mesure σ-finie et f une fonction dérivable et strictement convexe sur $]0, +\infty[$, la divergence de Bregman fonctionnelle ponctuelle

associée à f est définie par

$$\begin{aligned}\tilde{d}_f(x,y) &= \int d_f(x,y)\mathrm{d}m \\ &= \int \Big[f(x) - f(y) - f'(y)(x-y)\Big]\mathrm{d}m,\end{aligned}$$

pour x, y fonctions mesurables à valeurs positives. Si la mesure m est finie, pour x une fonction bornée à valeurs positives, on peut montrer (Frigyik *et al.* [89]) que

$$\phi(x) = \int f(x(u))\mathrm{d}m(u),$$

a pour différentielle

$$D_x\phi(h) = \int f'(x(u))h(u)\mathrm{d}m(u),$$

de sorte que $d_\phi = \tilde{d}_f$.

Voici quelques exemples de divergences de Bregman fonctionnelles, introduits par Frigyik *et al.* [89] ou obtenus en intégrant certaines divergences de Bregman unidimensionnelles. La notation $L^2(I,m)$ désigne l'ensemble des fonctions de carré intégrable pour la mesure m qui sont définies sur l'intervalle I et $C^0(I)$ l'ensemble des fonctions continues sur I, tandis que les ensembles correspondants de fonctions 2π-périodiques sont notés $L^2_{2\pi}(m)$ et $C^0_{2\pi}$ respectivement. Afin d'alléger les notations, la variable et l'espace d'intégration seront omis dans les calculs.

Exemple 1.3.3 1. Distance ℓ^2 au carré. Soit $\mathcal{C} = E = \ell^2$ l'espace des suites de carré sommable. Si ϕ est donnée par $\phi(x) = \sum_{j=1}^{+\infty} x_j^2$, pour tout $(x,y) \in E^2$,

$$\begin{aligned}d_\phi(x,y) &= \sum_{j=1}^{+\infty} x_j^2 - \sum_{j=1}^{+\infty} y_j^2 - 2\sum_{j=1}^{+\infty}(x_j - y_j)y_j \\ &= \sum_{j=1}^{+\infty}(x_j - y_j)^2 \\ &= \boxed{\|x-y\|^2_{\ell^2}}.\end{aligned}$$

2. Distance L^2 au carré. Soient I un intervalle de $\mathbb{R}$, m une mesure positive borélienne et $\mathcal{C} = E = L^2(I,m)$. Soit ϕ définie par $\phi(x) = \int_I x^2(t)\mathrm{d}m(t)$. Pour tout $(y,h) \in E^2$, on a

$$\begin{aligned}\phi(y+h) - \phi(h) &= \int (y+h)^2\mathrm{d}m - \int y^2\mathrm{d}m \\ &= 2\int yh\mathrm{d}m + \int h^2\mathrm{d}m.\end{aligned}$$

Comme $\lim_{\|h\|\to 0} \frac{\int h^2 \mathrm{d}m}{\|h\|} = \lim_{\|h\|\to 0} \|h\| = 0$ et que $h \mapsto 2\int yh\mathrm{d}m$ est une application linéaire continue, on a $D_y\phi(h) = 2\int_I y(t)h(t)\mathrm{d}m(t)$. En outre, si $k \in E$,

$$\begin{aligned} D_{y+k}\phi(h) - D_y\phi(h) &= 2\int (y+k)h\mathrm{d}m - 2\int yh\mathrm{d}m \\ &= 2\int kh\mathrm{d}m. \end{aligned}$$

D'où $D_y^2\phi(k,h) = 2\int_I k(t)h(t)\mathrm{d}m(t)$ et $D_y^2\phi(h,h) = 2\|h\|^2$, ce qui montre que la fonction ϕ est strictement convexe. Finalement, pour tout $(x,y) \in E^2$,

$$\begin{aligned} d_\phi(x,y) &= \int x^2\mathrm{d}m - \int y^2\mathrm{d}m - 2\int y(x-y)\mathrm{d}m \\ &= \int (x-y)^2\mathrm{d}m \\ &= \boxed{\|x-y\|_{L^2}^2}. \end{aligned}$$

3. Biais quadratique. Frigyik *et al.* [89] notent l'importance du biais quadratique : ainsi, il nous a paru intéressant de présenter cet exemple bien qu'il ne soit pas une vraie divergence de Bregman, la fonction convexe ϕ n'étant pas strictement convexe. Soient I un intervalle de $\mathbb{R}$, m une mesure positive borélienne finie et $\mathcal{C} = E = L^2(I,m)$ (réflexif). Un élément $x \in \mathcal{C}$ est intégrable, puisque m est supposée finie, et ϕ peut donc être définie par $\phi(x) = \left[\int_I x(t)\mathrm{d}m(t)\right]^2$. Pour tout $(y,h) \in E^2$, on a

$$\begin{aligned} \phi(y+h) - \phi(h) &= \left[\int y\mathrm{d}m + \int h\mathrm{d}m\right]^2 - \left[\int y\mathrm{d}m\right]^2 \\ &= 2\int y\mathrm{d}m \int h\mathrm{d}m + \left[\int h\mathrm{d}m\right]^2. \end{aligned}$$

L'application $h \mapsto 2\int y\mathrm{d}m \int h\mathrm{d}m$ est linéaire continue et on a

$$0 \leq \frac{\left(\int h\mathrm{d}m\right)^2}{\|h\|} \leq \frac{m(I)\|h\|^2}{\|h\|} = m(I)\|h\|,$$

donc $D_y\phi(h) = 2\int_I y(t)\mathrm{d}m(t) \int_I h(t)\mathrm{d}m(t)$. Comme, pour $k \in E$,

$$\begin{aligned} D_{y+k}\phi(h) - D_y\phi(h) &= 2\int (y+k)\mathrm{d}m \int h\mathrm{d}m - 2\int y\mathrm{d}m \int h\mathrm{d}m \\ &= 2\int k\mathrm{d}m \int h\mathrm{d}m, \end{aligned}$$

on a $D_y^2\phi(h,h) = 2\left[\int_I h(t)\mathrm{d}m(t)\right]^2$, et ϕ est convexe. Elle n'est pas strictement convexe puisque $D_y^2\phi(h,h) = 0$ dès que h est d'intégrale nulle. Soit $(x,y) \in E^2$. La pseudo-divergence de Bregman associée est

$$\begin{aligned} d_\phi(x,y) &= \left[\int x\mathrm{d}m\right]^2 - \left[\int y\mathrm{d}m\right]^2 - 2\int y dm \int (x-y)\mathrm{d}m \\ &= \left[\int x\mathrm{d}m\right]^2 + \left[\int y\mathrm{d}m\right]^2 - 2\int y\mathrm{d}m \int x\mathrm{d}m \\ &= \boxed{\left[\int (x-y)\mathrm{d}m\right]^2}. \end{aligned}$$

4. Distance de Kullback-Leibler généralisée entre deux mesures positives à densité. Soient $E = C^0([0,1])$ et $\mathcal{C}$ l'ensemble des éléments positifs de E. Comme la fonction $t \mapsto x(t)\ln x(t)$ est continue sur $[0,1]$, on a $\int_0^1 |x(t)\ln x(t)|\mathrm{d}t < +\infty$. On pose $\phi(x) = \int_0^1 x(t)\ln x(t)\mathrm{d}t$. Soit $y \in ir(\mathcal{C})$, ce qui signifie que y est à valeurs strictement positives. Pour tout h de norme assez petite pour que $y + h > 0$, on a

$$\begin{aligned} \phi(y+h) - \phi(y) &= \int [(y+h)\ln(y+h) - y\ln y] \\ &= \int \left[y\ln\left(1+\frac{h}{y}\right) + h\ln(y+h)\right] \\ &= \int \left[y\ln\left(1+\frac{h}{y}\right) + h\ln y + h\ln\left(1+\frac{h}{y}\right)\right] \\ &= \int h(1+\ln y) + o(h). \end{aligned}$$

Pour tout $h \in E$, on a $\int_0^1 h(t)(1+\ln y(t))\mathrm{d}t \leq \|h\|_\infty\left(1 + \int_0^1 \ln y(t)\mathrm{d}t\right)$ avec $\int_0^1 \ln y(t)\mathrm{d}t < +\infty$, l'application $h \mapsto \int h(1+\ln y)$ est linéaire continue sur E, et on a $D_y\phi(h) = \int_0^1 h(t)(1+\ln y(t))\mathrm{d}t$. Pour $h \in E$ et k tel que $y+k > 0$, on a

$$\begin{aligned} D_{y+k}\phi(h) - D_y\phi(h) &= \int [h(1+\ln(y+k)) - h(1+\ln y)] \\ &= \int h\ln\left(1+\frac{k}{y}\right) \\ &= \int \frac{hk}{y} + o(h,k). \end{aligned}$$

Donc,

$$D_y^2\phi(h,h) = \int_0^1 \frac{h(t)^2}{y(t)}\mathrm{d}t,$$

et ϕ est strictement convexe. Pour $(x, y) \in \mathcal{C} \times ir(\mathcal{C})$, on a

$$\begin{aligned} d_\phi(x, y) &= \int [x \ln x - y \ln y + (y - x)(1 + \ln y)] \\ &= \boxed{\int_0^1 \left[x(t) \ln \frac{x(t)}{y(t)} + y(t) - x(t) \right] \mathrm{d}t}. \end{aligned}$$

Notons cependant que les résultats établis dans la suite du chapitre requièrent la réflexivité de l'espace de Banach E, de sorte qu'il serait souhaitable dans cet exemple de se placer dans $L^2([0, 1], \mathrm{d}t)$ plutôt que dans $C^0([0, 1])$. Pour ce faire, afin d'éviter un problème topologique lié au calcul de la différentielle (l'intérieur pour la norme L^2 du cônes des fonctions à valeurs positives est vide), il est commode d'utiliser la définition des divergences de Bregman fonctionnelles ponctuelles présentée dans la Remarque 1.3.3.

Soit f définie par $f(x) = x \ln x - x + 1$ pour tout $x \geq 0$. Cette fonction est dérivable et strictement convexe sur $]0, +\infty[$. On a $f'(x) = \ln(x)$ et $f''(x) = \frac{1}{x}$. Pour $(x, y) \in L^2([0, 1], \mathrm{d}t)$, fonctions à valeurs positives, la divergence de Bregman fonctionnelle ponctuelle obtenue est

$$\begin{aligned} \tilde{d}_f(x, y) &= \int_0^1 [f(x(t)) - f(y(t)) - f'(y(t))(x(t) - y(t))] \mathrm{d}t \\ &= \int_0^1 [x \ln x - x + 1 - (y \ln y - y + 1) + \ln y (y - x)] \\ &= \int_0^1 [x \ln x - y \ln y + (1 + \ln y)(y - x)] \\ &= \boxed{\int_0^1 \left[x(t) \ln \frac{x(t)}{y(t)} + y(t) - x(t) \right] \mathrm{d}t}. \end{aligned}$$

5. Distance de Kullback-Leibler entre deux mesures de probabilité à densité. Comme dans l'exemple précédent, soient $E = C^0([0, 1])$, $\mathcal{C}$ l'ensemble des éléments de E à valeurs positives et ϕ définie par $\phi(x) = \int_0^1 x(t) \ln x(t) \mathrm{d}t$. La divergence de Bregman associée, restreinte aux densités de probabilité, est

$$d_\phi(x, y) = \boxed{\int_0^1 x(t) \ln \frac{x(t)}{y(t)} \mathrm{d}t}.$$

En se plaçant dans $L^2([0, 1], \mathrm{d}t)$, et en calculant la divergence de Bregman ponctuelle correspondant à $f(x) = x \ln x - x + 1$, on obtient de même

$$\tilde{d}_f(x, y) = \boxed{\int_0^1 x(t) \ln \frac{x(t)}{y(t)} \mathrm{d}t}.$$

6. Distance de Itakura-Saito. C'est une mesure de distorsion entre densités spectrales. Soient $E = C^0_{2\pi}$ et $\mathcal{C}$ l'ensemble des fonctions de E à valeurs strictement positives. Sur le segment $[-\pi, \pi]$, la fonction continue $\ln \circ x$ est bornée, donc on peut définir ϕ par

$$\phi(x) = -\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln(x(\theta))\mathrm{d}\theta.$$

Pour h assez petit, on a

$$\begin{aligned}\phi(y+h) - \phi(y) &= -\frac{1}{2\pi}\int \ln(y+h) + \frac{1}{2\pi}\int \ln y \\ &= -\frac{1}{2\pi}\int \ln\left(\frac{y+h}{y}\right) \\ &= -\frac{1}{2\pi}\int \ln\left(1+\frac{h}{y}\right) \\ &= -\frac{1}{2\pi}\int \frac{h}{y} + o(h).\end{aligned}$$

On en déduit que

$$D_y\phi(h) = -\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{h(\theta)}{y(\theta)}\mathrm{d}\theta.$$

Pour k assez petit, on a

$$\begin{aligned}D_{k+y}\phi(h) - D_y\phi(h) &= -\frac{1}{2\pi}\int \left(\frac{h}{k+y} - \frac{h}{y}\right) \\ &= \frac{1}{2\pi}\int \frac{h}{y}\left(1 - \frac{1}{1+\frac{k}{y}}\right) \\ &= \frac{1}{2\pi}\int \frac{hk}{y^2} + o(h,k).\end{aligned}$$

Ainsi,

$$D^2_y(h,h) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\frac{h(\theta)}{y(\theta)}\right)^2 \mathrm{d}\theta,$$

ce qui assure que ϕ est strictement convexe. Finalement, pour tout $(x,y) \in \mathcal{C}^2$,

$$d_\phi(x,y) = \boxed{-\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\ln\frac{x(\theta)}{y(\theta)} - \frac{x(\theta)}{y(\theta)} + 1\right)\mathrm{d}\theta}.$$

Comme précédemment, on peut également utiliser la définition ponctuelle, ce qui permet de se placer dans $L^2_{2\pi}(\mathrm{d}\theta)$. Soit f définie par $f(x) = x - \ln x - 1$

pour tout $x > 0$. Cette fonction est dérivable et strictement convexe sur $]0, +\infty[$. On a $f'(x) = 1 - \frac{1}{x}$ et $f''(x) = \frac{1}{x^2}$. On obtient, pour tout $(x, y) \in L^2_{2\pi}(\mathrm{d}\theta)$ à valeurs strictement positives,

$$\begin{aligned}\tilde{d}_f(x, y) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x(\theta)) - f(y(\theta)) - f'(y(\theta))(x(\theta) - y(\theta))\mathrm{d}\theta \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[x - \ln x - 1 - (y - \ln y - 1) - \left(1 - \frac{1}{y}\right)(x - y)\right] \\ &= \boxed{-\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\ln\frac{x(\theta)}{y(\theta)} - \frac{x(\theta)}{y(\theta)} + 1\right)\mathrm{d}\theta}.\end{aligned}$$

1.3.3. Quelques propriétés des divergences de Bregman

Les divergences de Bregman vérifient différentes propriétés, que nous énonçons dans cette section. Positivité, séparation, convexité et linéarité, ainsi que la formule de Pythagore généralisée et les propriétés de séparateur linéaire et de classes d'équivalence sont établies dans un cadre fonctionnel par Frigyik, Srivastava et Gupta [88] (voir aussi Bregman [40] et Nielsen *et al.* [149] pour le cas de la dimension finie). Par souci de clarté, nous en redonnons ici les preuves en détaillant ou simplifiant certains points.

Positivité et séparation

Proposition 1.3.1 (Positivité et séparation)**.** *Toute divergence de Bregman $d_\phi(\cdot, \cdot)$ vérifie $d_\phi(x, y) \in \mathbb{R}^+$ pour tout $(x, y) \in \mathcal{C} \times ir(\mathcal{C})$ et on a $d_\phi(x, y) = 0$ si, et seulement si, $x = y$.*

Démonstration. La positivité résulte de la convexité de ϕ et la propriété de séparation est due à sa stricte convexité. Soit $(x, y) \in \mathcal{C} \times ir(\mathcal{C})$. Posons $\tilde{\phi}(t) = \phi(tx+(1-t)y)$ pour tout $t \in [0, 1]$. Vérifions que $d_\phi(x, y)$ est positive. Par convexité de ϕ,

$$\tilde{\phi}(t) \leq t\tilde{\phi}(1) + (1 - t)\tilde{\phi}(0), \tag{1.3}$$

ce qui implique, pour tout $t \in [0, 1[$,

$$\frac{\tilde{\phi}(t) - \tilde{\phi}(0)}{t} \leq \tilde{\phi}(1) - \tilde{\phi}(0) = \phi(x) - \phi(y).$$

En faisant tendre t vers 0, on obtient

$$\tilde{\phi}'(0) = D_y f(x - y) \leq \phi(x) - \phi(y),$$

et donc $d_\phi(x,y) \geq 0$. Pour démontrer la propriété de séparation, supposons que $d_\phi(x,y) = 0$. Il en résulte

$$\tilde{\phi}'(0) = \tilde{\phi}(1) - \tilde{\phi}(0). \tag{1.4}$$

Remarquons que la fonction $\tilde{\phi}$ est convexe, de sorte que pour tout $t \in]0,1[$,

$$\tilde{\phi}'(0) \leq \frac{\tilde{\phi}(t) - \tilde{\phi}(0)}{t},$$

ce qui se récrit

$$\begin{aligned}\tilde{\phi}(t) &\geq \tilde{\phi}(0) + t\tilde{\phi}'(0)\\ &\geq t\tilde{\phi}(1) + (1-t)\tilde{\phi}(0),\end{aligned}$$

en utilisant l'égalité (1.4). Avec l'inégalité (1.3), nous obtenons

$$\tilde{\phi}(t) = t\tilde{\phi}(1) + (1-t)\tilde{\phi}(0),$$

c'est-à-dire

$$\phi(tx + (1-t)y) = t\phi(x) + (1-t)\phi(y).$$

La stricte convexité de ϕ entraîne $x = y$. □

Convexité en la première variable

Proposition 1.3.2 (Convexité)**.** *Une divergence de Bregman $d_\phi(\cdot,\cdot)$ est strictement convexe en la première variable.*

Démonstration. Soit $z \in ir(\mathcal{C})$. Pour tout $(x,y) \in \mathcal{C}^2$, $x \neq y$, et tout $t \in]0,1[$,

$$\begin{aligned}&d_\phi(tx + (1-t)y, z)\\ &= \phi(tx + (1-t)y) - \phi(z) - D_z\phi(tx + (1-t)y - z)\\ &< t\phi(x) + (1-t)\phi(y) - \phi(z) - tD_z\phi(x) - (1-t)D_z\phi(y) + D_z\phi(z)\\ &= t(\phi(x) - \phi(z) - D_z\phi(x-z)) + (1-t)(\phi(y) - \phi(z) - D_z\phi(y-z))\\ &= td_\phi(x,z) + (1-t)d_\phi(y,z),\end{aligned}$$

où l'inégalité stricte résulte de la stricte convexité de ϕ. □

Remarque 1.3.4. La fonction $y \mapsto d_\phi(x,y)$ n'est pas nécessairement convexe.

Exemple 1.3.1.

1. La distance de Itakura-Saito définie par $d_\phi(x,y) = \frac{x}{y} - \ln\frac{x}{y} - 1$ n'est pas convexe en y.
2. Un autre exemple est le suivant, donné par Banerjee *et al.* [17] : on prend ϕ définie sur $\mathbb{R}^+$ par $\phi(x) = x^3$, de sorte que $d_\phi(x,y) = x^3 - y^3 - 3(x-y)y^2$. La dérivée seconde de $y \mapsto d_\phi(x,y)$ est $y \mapsto 12y - 6x$ qui n'est pas positive sur $\mathbb{R}^+$ entier. La fonction $y \mapsto d_\phi(x,y)$ n'est donc pas convexe.

Linéarité

On peut vérifier facilement qu'une divergence de Bregman est linéaire dans le sens suivant :

Proposition 1.3.3 (Linéarité)**.** *Soient ϕ_1, ϕ_2 des fonctions strictement convexes de $\mathcal{C}$ dans $\mathbb{R}$ et α, β des constantes. Alors, on a $d_{(\alpha\phi_1+\beta\phi_2)}(\cdot,\cdot) = \alpha d_{\phi_1}(\cdot,\cdot) + \beta d_{\phi_2}(\cdot,\cdot)$.*

Démonstration. Pour tout $(x, y) \in \mathcal{C} \times ir(\mathcal{C})$,

$$\begin{aligned} d_{(\alpha\phi_1+\beta\phi_2)}(x,y) &= (\alpha\phi_1 + \beta\phi_2)(x) - (\alpha\phi_1 + \beta\phi_2)(y) - D_y(\alpha\phi_1 + \beta\phi_2)(x-y) \\ &= \alpha d_{\phi_1}(x,y) + \beta d_{\phi_2}(x,y). \end{aligned}$$

□

Formule de Pythagore généralisée

Un calcul élémentaire montre que les divergences de Bregman vérifient une égalité de Pythagore généralisée.

Proposition 1.3.4. *Soit $d_\phi(\cdot,\cdot)$ une divergence de Bregman. Pour $x \in \mathcal{C}$, $(y, z) \in [ir(\mathcal{C})]^2$, on a*

$$d_\phi(x,y) = d_\phi(x,z) + d_\phi(z,y) + D_z\phi(x-z) - D_y\phi(x-z). \tag{1.5}$$

Démonstration. On a

$$\begin{aligned} d_\phi(x,z) + d_\phi(z,y) &= \phi(x) - \phi(z) - D_z\phi(x-z) + \phi(z) - \phi(y) - D_y\phi(z-y) \\ &= \phi(x) - \phi(y) - D_y\phi(x-y) + D_y\phi(x-z) - D_z\phi(x-z) \\ &= d_\phi(x,y) + D_y\phi(x-z) - D_z\phi(x-z), \end{aligned}$$

d'où la formule (1.5). □

Observons que si ϕ est le carré d'une norme hilbertienne, on a

$$D_z\phi(x-z) - D_y\phi(x-z) = 2\langle z-y, x-z\rangle,$$

et l'égalité (1.5) devient

$$\|x-y\|^2 = \|x-z\|^2 + \|z-y\|^2 - 2\langle y-z, x-z\rangle,$$

qui n'est autre que la formule d'Al-Kashi. De plus, le produit scalaire $\langle y-z, x-z\rangle$ est nul si, et seulement si, $y-z$ et $x-z$ sont orthogonaux, et on retrouve ainsi le théorème de Pythagore.

Hyperplan séparateur

Proposition 1.3.5. *L'ensemble des éléments de $\mathcal{C}$ équidistants, relativement à une divergence de Bregman, de deux éléments de $ir(\mathcal{C})$ est l'intersection de $\mathcal{C}$ avec un hyperplan affine de E. En d'autres termes, si $(y,z) \in [ir(\mathcal{C})]^2$, l'ensemble des $x \in \mathcal{C}$ tels que $d_\phi(x,y) = d_\phi(x,z)$ est l'intersection de $\mathcal{C}$ avec un hyperplan affine de E.*

Démonstration. Soit $(y,z) \in [ir(\mathcal{C})]^2$. Il s'agit de montrer que l'ensemble des $x \in \mathcal{C}$ tels que $d_\phi(x,y) = d_\phi(x,z)$ est l'intersection de $\mathcal{C}$ avec le noyau d'une forme affine non constante. Or $d_\phi(x,y) = d_\phi(x,z)$ équivaut à $D_z\phi(x) - D_y\phi(x) = \phi(y) - \phi(z) + D_z\phi(z) - D_y\phi(y)$, ce qui se récrit $L(x) = c$ avec $c = \phi(y) - \phi(z) + D_z\phi(z) - D_y\phi(y)$ et L la forme linéaire définie par $L(x) = D_z\phi(x) - D_y\phi(x)$, d'où le résultat. □

Classes d'équivalence

Une relation d'équivalence $\sim$ peut être définie sur les fonctions strictement convexes en posant que deux fonctions appartiennent à la même classe lorsqu'elles conduisent à la même divergence de Bregman.

Proposition 1.3.6. *Les fonctions strictement convexes ϕ_1 et ϕ_2 appartiennent à une même classe pour la relation $\sim$ si et seulement si elles ne diffèrent que par un terme de la forme $R(x) = L(x) + c$ où L est une forme linéaire et c une constante.*

Démonstration. Tout d'abord, si ϕ_2 est définie par $\phi_2(x) = \phi_1(x) + L(x) + c$, on a pour tout $(x,y) \in \mathcal{C} \times ir(\mathcal{C})$

$$\begin{aligned} d_{\phi_2}(x,y) &= \phi_2(x) - \phi_2(y) - D_y\phi_2(x-y) \\ &= \phi_1(x) - \phi_1(y) + L(x) - L(y) - D_y\phi_1(x-y) - L(x-y) \\ &= \phi_1(x) - \phi_1(y) - D_y\phi_1(x-y) \\ &= d_{\phi_1}(x,y), \end{aligned}$$

donc ϕ_1 et ϕ_2 définissent la même divergence de Bregman. Inversement, supposons que pour tout $(x,y) \in \mathcal{C} \times ir(\mathcal{C})$, $d_{\phi_1}(x,y) = d_{\phi_2}(x,y)$. On a alors

$$\phi_2(x) - \phi_2(y) - D_y\phi_2(x-y) = \phi_1(x) - \phi_1(y) - D_y\phi_1(x-y)$$

En fixant y et en posant $c = \phi_2(y) - D_y\phi_2(y) - \phi_1(y) + D_y\phi_1(y)$, il vient, pour tout $x \in \mathcal{C}$,

$$\phi_2(x) = \phi_1(x) + D_y\phi_2(x) - D_y\phi_1(x) + c,$$

donc ϕ_2 est de la forme cherchée avec $L = D_y\phi_2 - D_y\phi_1$. □

Remarque 1.3.5 (Divergences de Bregman fonctionnelles ponctuelles). Les propriétés que nous venons d'énoncer sont également vérifiées (en remplaçant $D_y\phi(x)$ par $\int xf'(y)\mathrm{d}m$ le cas échéant) par les divergences de Bregman fonctionnelles ponctuelles, de la forme $\tilde{d}_f(x,y) = \int d_f(x,y)\mathrm{d}m$. Pour retrouver les différents résultats, il suffit d'utiliser le fait que $d_f(x,y)$ est une divergence de Bregman unidimensionnelle ainsi que les propriétés de l'intégrale (positivité, linéarité).

Lien avec l'inégalité de Jensen

Nous savons qu'une fonction convexe ϕ vérifie l'inégalité de Jensen. Pour une variable aléatoire X,

$$\mathbb{E}[\phi(X)] \geq \phi(\mathbb{E}[X]).$$

Cette inégalité peut s'exprimer en matière de divergence de Bregman associée à ϕ. Dans le cas d'une divergence de Bregman en dimension finie, ce résultat est établi par Banerjee, Guo et Wang [15]. Nous proposons de le retrouver en dimension quelconque. Supposons que $\mathbb{E}[X] \in ir(\mathcal{C})$. Si $\mathbb{E}|D_{\mathbb{E}[X]}\phi(X)| < +\infty$, on a

$$\mathbb{E}[D_{\mathbb{E}[X]}\phi(X - \mathbb{E}[X])] = D_{\mathbb{E}[X]}\phi(\mathbb{E}[X] - \mathbb{E}[X]) = 0,$$

en échangeant l'espérance et la différentielle, puisque celle-ci est une forme linéaire continue (voir par exemple Arendt, Batty, Hieber et Neubrander [9, Proposition 1.1.6]). Alors,

$$\begin{aligned}\mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X]) &= \mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X]) - \mathbb{E}[D_{\mathbb{E}[X]}\phi(X - \mathbb{E}[X])] \\ &= \mathbb{E}[\phi(X) - \phi(\mathbb{E}[X]) - D_{\mathbb{E}[X]}\phi(X - \mathbb{E}[X])] \\ &= \mathbb{E}[d_\phi(X, \mathbb{E}[X])] \geq 0.\end{aligned}$$

La différence entre les deux termes de l'inégalité de Jensen correspond à l'espérance de la divergence de Bregman entre X et $\mathbb{E}[X]$, qui est positive.

Relation entre divergences de Bregman et familles exponentielles

Banerjee *et al.* [17] ont établi une relation entre divergences de Bregman et familles exponentielles. Ce résultat, qui repose sur la notion de dualité des fonctions convexes (voir par exemple Rockafellar [161]), fait l'objet de l'Annexe 1.8.4. Son intérêt pratique est d'apporter un élément de réponse à la question : quelle divergence de Bregman utiliser pour quel type d'observations ?

1.3.4. Projection de Bregman

Dans un espace de Banach E réflexif, il est possible de définir une projection de Bregman sur un convexe fermé qui généralise la projection hilbertienne (Alber

et Butnariu [4]). En dimension finie, la notion de projection basée sur une divergence de Bregman est développée entre autres par Bregman [40] et Nielsen *et al.* [149], tandis que le cas d'une divergence de Bregman ponctuelle a été traité par Csiszár [56]. Dans un souci de clarté et de cohérence des notations, nous incluons ici la présentation et les preuves de la méthode de projection d'Alber et Butnariu [4].

Définition 1.3.4 (Projection de Bregman)**.** *Soient $\mathcal{C}$ et $\mathcal{K}$ deux convexes de E tels que $\mathcal{K} \subset ir(\mathcal{C})$ et soit $d_\phi : \mathcal{C} \times ir(\mathcal{C}) \to [0, +\infty[$ une divergence de Bregman. Pour tout $y \in ir(\mathcal{C})$, comme $x \mapsto d_\phi(x, y)$ est strictement convexe, l'ensemble $\arg\min_{x\in\mathcal{K}} d_\phi(x, y)$ contient au plus un élément $\overline{y}$. Si $\overline{y}$ existe, on l'appelle la projection de Bregman de y sur $\mathcal{K}$ associée à la fonction ϕ.*

Notons

$$d_\phi(\mathcal{K}, y) = \inf_{x\in\mathcal{K}} d_\phi(x, y).$$

Proposition 1.3.7 (Existence de la projection)**.** *Soient $d_\phi : \mathcal{C} \times ir(\mathcal{C}) \to [0, +\infty[$ une divergence de Bregman et $\mathcal{K} \subset ir(\mathcal{C})$ un convexe fermé borné de l'espace de Banach réflexif E. Alors pour tout $y \in ir(\mathcal{C})$, la projection $\overline{y}$ existe.*

Démonstration. Par définition de la borne inférieure, il existe une suite $(x_n)_{n\in\mathbb{N}}$ d'éléments de $\mathcal{K}$ telle que

$$\lim_{n\to+\infty} d_\phi(x_n, y) = d_\phi(\mathcal{K}, y).$$

Comme $\mathcal{K}$ est borné, la suite $(x_n)_{n\in\mathbb{N}}$ est également bornée. Donc, E étant réflexif, elle possède une sous-suite extraite, encore notée $(x_n)_{n\in\mathbb{N}}$, qui converge pour la topologie faible $\sigma(E, E')$ vers un $x \in E$ (Annexe 1.8.3, Théorème 1.8.6). Comme $\mathcal{K}$ est un convexe fortement fermé, il est aussi faiblement fermé (Proposition 1.8.5), et ainsi $x \in \mathcal{K}$. Pour tout $y \in ir(\mathcal{C})$, la fonction $z \mapsto d_\phi(z, y)$ est convexe et continue sur $\mathcal{K} \subset ir(\mathcal{C})$. Ainsi, son épigraphe $\{(z, \lambda) \in \mathcal{K} \times \mathbb{R}, d_\phi(z, y) \leq \lambda\}$ est un convexe fortement fermé de $E \times \mathbb{R}$ (Propositions 1.8.3 et 1.8.4), donc aussi faiblement fermé. On en déduit que la limite faible de la suite $(x_n, d_\phi(x_n, y))$ appartient à l'épigraphe de la fonction $z \mapsto d_\phi(z, y)$, c'est-à-dire $d_\phi(x, y) \leq d_\phi(\mathcal{K}, y)$. On pose alors $\overline{y} = x$. □

Proposition 1.3.8. *Soient $d_\phi(\cdot, \cdot)$ une divergence de Bregman, $\mathcal{K} \subset ir(\mathcal{C})$ un convexe fermé borné de E, $y \in ir(\mathcal{C})$ et $\overline{y}$ sa projection de Bregman sur $\mathcal{K}$. On a*

$$\forall x \in \mathcal{K}, d_\phi(x, \overline{y}) \leq d_\phi(x, y) - d_\phi(\overline{y}, y) \leq d_\phi(x, y).$$

Démonstration. Le fait qu'une divergence de Bregman soit à valeurs positives implique l'inégalité de droite. Pour celle de gauche, écrivons

$$\begin{aligned} d_\phi(x, \overline{y}) + d_\phi(\overline{y}, y) &= \phi(x) - \phi(y) - D_{\overline{y}}\phi(x - \overline{y}) - D_y\phi(\overline{y} - y) \\ &= d_\phi(x, y) + D_y\phi(x - y) - D_{\overline{y}}\phi(x - \overline{y}) - D_y\phi(\overline{y} - y) \\ &= d_\phi(x, y) + D_y\phi(x - \overline{y}) - D_{\overline{y}}\phi(x - \overline{y}). \end{aligned}$$

Montrons que $D_y\phi(x-\overline{y}) \leq D_{\overline{y}}\phi(x-\overline{y})$, ce qui établira l'inégalité. Soit $a \in]0,1]$. On pose $u(a) = \overline{y} + a(x-\overline{y})$. Comme $\mathcal{K}$ est convexe, $u(a) \in \mathcal{K}$ et ainsi $d_\phi(\overline{y}, y) \leq d_\phi(u(a), y)$. On a

$$\begin{aligned} 0 &\leq d_\phi(u(a), y) - d_\phi(\overline{y}, y) \\ &\leq D_{u(a)} d_\phi(u(a) - \overline{y}, y) \\ &= D_{u(a)} d_\phi(a(x-\overline{y}), y) \\ &= a(D_{u(a)}\phi(x-\overline{y}) - D_y\phi(x-\overline{y})), \end{aligned}$$

où la deuxième inégalité découle de la convexité de d_ϕ en la première variable. Donc,

$$D_y\phi(x-\overline{y}) \leq D_{u(a)}\phi(x-\overline{y}).$$

Comme ϕ est de classe C^1, on peut faire tendre a vers 0 pour obtenir

$$D_y\phi(x-\overline{y}) \leq D_{\overline{y}}\phi(x-\overline{y}).$$

□

Remarque 1.3.6. Au lieu de prendre $\mathcal{K} \subset ir(\mathcal{C})$ dans la Définition 1.3.4 et la Proposition 1.3.7, nous aurions pu faire l'hypothèse plus générale $\mathcal{K} \cap ir(\mathcal{C}) \neq \emptyset$. Notons que, dans ce cas, la projection $\overline{y}$ d'un élément $y \in ir(\mathcal{C})$ peut ne plus appartenir à $ir(\mathcal{C})$. Or, la Proposition 1.3.8 n'a de sens que lorsque $\overline{y} \in ir(\mathcal{C})$. Cependant, si ϕ est une fonction de Legendre (Annexe 1.8.3), alors $\overline{y} \in ir(\mathcal{C})$ (Bauschke, Borwein et Combettes [26]).

Puisque que les divergences de Bregman ont été définies et leurs principales propriétés rappelées, revenons à présent à notre problème de quantification, en prenant pour mesure de distorsion d une divergence de Bregman d_ϕ.

1.4. Choix d'un bon quantificateur

Soit $(E, \|\cdot\|)$ un espace de Banach réflexif et séparable. S'il s'agit d'un espace de Hilbert, son produit scalaire est noté $\langle\cdot,\cdot\rangle$. Nous considérons désormais une variable aléatoire X de loi μ à valeurs dans un convexe $\mathcal{C} \subset E$. Tout au long de ce chapitre, nous faisons les hypothèses suivantes :

1. $\mathbb{E}\|X\| < +\infty$.
2. $\mathbb{E}[X] \in ir(\mathcal{C})$.
3. $\mathbb{E}|\phi(X)| < +\infty$ et, pour tout $c \in ir(\mathcal{C})$, $\mathbb{E}|D_c\phi(X)| < +\infty$. Ceci implique en particulier $\mathbb{E}[d_\phi(X,c)] < +\infty$ pour tout c.

Dans ce contexte, un k-quantificateur est une application borélienne $q : \mathcal{C} \subset E \to \mathbf{c}$, où la table de codage $\mathbf{c} = \{c_1, \ldots, c_\ell\}$, $\ell \leq k$, est un sous-ensemble de $ir(\mathcal{C})$.

1.4.1. Quantificateur des plus proches voisins

Rappelons que la précision d'un quantificateur q est évaluée grâce à la distorsion $W(\mu, q)$. Pour obtenir le meilleur quantificateur possible (au sens de la distorsion), nous devons en principe déterminer la meilleure table de codage $\mathbf{c} = \{c_1, \ldots, c_\ell\}$ et la meilleure partition $\{S_1, \ldots, S_\ell\}$, $\ell \leq k$, puisqu'un quantificateur est caractérisé par sa partition et sa table de codage. Or, lorsque d est le carré de la distance euclidienne, on connaît, à même table de codage, la meilleure partition, et à partition donnée, la meilleure table de codage (voir par exemple Linder [131]). Nous allons voir qu'il en va de même dans le cas présent, où d est plus généralement une divergence de Bregman d_ϕ.

La définition suivante décrit un type de partition important, la partition de Voronoi.

Définition 1.4.1 (Partition de Voronoi)**.** *Etant donné une table de codage* $\mathbf{c} = \{c_j\}_{j=1}^{\ell}$*, une partition* $\{S_j\}_{j=1}^{\ell}$ *vérifiant*

$$S_1 = \{x \in \mathcal{C}, d_\phi(x, c_1) \leq d_\phi(x, c_p), p = 1 \ldots, \ell\},$$

et pour $j = 2, \ldots, \ell$*,*

$$S_j = \{x \in \mathcal{C}, d_\phi(x, c_j) \leq d_\phi(x, c_p), p = 1, \ldots, \ell\} \setminus \bigcup_{m=1}^{j-1} S_m,$$

est appelée partition de Voronoi.

Remarque 1.4.1. Retirer $\bigcup_{m=1}^{j-1} S_m$ est une manière d'éviter les ambiguïtés au bord des cellules : si $d_\phi(x, c_i) = d_\phi(x, c_j)$, on affecte x à la cellule dont l'indice est le plus petit.

En couplant chaque table de codage à sa partition de Voronoi, on obtient une classe particulières de quantificateurs.

Définition 1.4.2 (Quantificateur des plus proches voisins)**.** *Un quantificateur de table de codage* $\mathbf{c} = \{c_j\}_{j=1}^{\ell}$ *qui admet comme partition la partition de Voronoi associée à* $\mathbf{c}$ *est appelé quantificateur des plus proches voisins.*

Le résultat suivant nous montre que s'il existe un k-quantificateur optimal, c'est nécessairement un quantificateur des plus proches voisins.

Lemme 1.4.1 (Meilleure partition)**.** *Soient* q *un* k*-quantificateur de table de codage* $\{c_j\}_{j=1}^{\ell}$ *et* q' *le quantificateur des plus proches voisins ayant même table de codage. Alors, on a*

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

Démonstration. Par définition de q', on a

$$d_\phi(x, q'(x)) = \min_{y \in \mathbf{c}} d_\phi(x, y).$$

Donc

$$\begin{aligned} W(\mu, q) &= \mathbb{E}[d_\phi(X, q(X))] \\ &= \int_{\mathcal{C}} d_\phi(x, q(x)) \mathrm{d}\mu(x) \\ &= \sum_{j=1}^{\ell} \int_{S_j} d_\phi(x, c_j) \mathrm{d}\mu(x) \\ &\geq \sum_{j=1}^{\ell} \int_{S_j} \min_{c \in \mathbf{c}} d_\phi(x, c) \mathrm{d}\mu(x) \\ &= \mathbb{E}[d_\phi(X, q'(X))] \\ &= W(\mu, q'). \end{aligned}$$

□

Pour construire un bon quantificateur, il suffit par conséquent de considérer les quantificateurs des plus proches voisins. Ainsi, notre but est de trouver la table de codage optimale, en minimisant la distorsion récrite en fonction de $\mathbf{c}$,

$$W(\mu, \mathbf{c}) = \mathbb{E}\left[\min_{j=1,\dots,k} d_\phi(X, c_j)\right].$$

Comparons à présent les quantificateurs de même partition, pour déterminer quelle est la meilleure table de codage. Pour cela, le point important consiste à montrer que la fonction

$$c \mapsto \mathbb{E}[d_\phi(X, c) | X \in S],$$

avec $S \subset \mathcal{C}$, atteint son minimum en un élément $c \in ir(\mathcal{C})$. Pour une divergence de Bregman en dimension finie, ce résultat est dû à Banerjee, Guo et Wang [16]. La preuve s'adapte en dimension infinie, comme nous allons le voir.

Proposition 1.4.1. *Soit $S \subset \mathcal{C}$ mesurable tel que $\mu(S) > 0$ et $\mathbb{E}[X|X \in S] \in ir(\mathcal{C})$. Alors, la fonction $c \mapsto \mathbb{E}[d_\phi(X, c)|X \in S]$ atteint son minimum sur $ir(\mathcal{C})$ en un unique élément $\mathbb{E}[X|X \in S]$.*

Remarque 1.4.2. Ce n'est pas le cas lorsque la mesure de distorsion est une norme L^1 (contexte considéré par Laloë [125]). En effet, il faut alors considérer la médiane et non l'espérance (Kemperman [114]).

Démonstration. Nous allons vérifier que $\mathbb{E}[X|X \in S]$ minimise $\mathbb{E}[d_\phi(X,c)|X \in S]$ en c et qu'il s'agit du seul élément de $ir(\mathcal{C})$ ayant cette propriété. Pour tout $c \in ir(\mathcal{C})$,

$$\begin{aligned}
&\mathbb{E}[d_\phi(X,c)|X \in S] - \mathbb{E}[d_\phi(X,\mathbb{E}[X|X \in S])|X \in S] \\
&= \mathbb{E}[\phi(X) - \phi(c) - D_c\phi(X-c) - \phi(X) + \phi(\mathbb{E}[X|X \in S]) \\
&\quad + D_{\mathbb{E}[X|X\in S]}\phi(X - \mathbb{E}[X|X \in S])|X \in S] \\
&= \phi(\mathbb{E}[X|X \in S]) - \phi(c) - D_c\phi(\mathbb{E}[X|X \in S] - c) \\
&= d_\phi(\mathbb{E}[X|X \in S], c),
\end{aligned}$$

où la seconde égalité découle de l'interversion entre l'espérance et la différentielle, qui est une forme linéaire continue. Or, la propriété de positivité et séparation des divergences de Bregman implique

$$d_\phi(\mathbb{E}[X|X \in S], c) \geq 0$$

et $d_\phi(\mathbb{E}[X|X \in S], c) = 0$ si, et seulement si, $c = \mathbb{E}[X|X \in S]$, d'où

$$\mathbb{E}[d_\phi(X,c)|X \in S] \geq \mathbb{E}[d_\phi(X,\mathbb{E}[X|X \in S])|X \in S]$$

avec égalité si, et seulement si, $c = \mathbb{E}[X|X \in S]$. Donc, $\mathbb{E}[X|X \in S]$ est l'unique minimiseur de la fonction $c \mapsto \mathbb{E}[d_\phi(X,c)|X \in S]$ sur $ir(\mathcal{C})$. □

Notons que Frigyik *et al.* [89], avec d'autres hypothèses sur la fonction ϕ, proposent une preuve différente, basée sur le calcul différentiel.

Nous pouvons à présent décrire la meilleure table de codage pour un quantificateur de partition donnée.

Lemme 1.4.2 (Meilleure table de codage)**.** *Soit q un quantificateur de partition associée $\{S_j\}_{j=1}^\ell$ avec $\mu(S_j) > 0$ et $\mathbb{E}[X|X \in S_j] \in ir(\mathcal{C})$ pour $j = 1, \ldots, \ell$. Si q' est un quantificateur de même partition, dont la table de codage $\{c'_1, \ldots, c'_\ell\}$ est définie par*

$$c'_j \in \arg\min_{c \in ir(\mathcal{C})} \mathbb{E}[d_\phi(X,c)|X \in S_j] \quad \textit{pour } j = 1, \ldots, \ell,$$

c'est-à-dire

$$c'_j = \mathbb{E}[X|X \in S_j] \quad \textit{pour } j = 1, \ldots, \ell,$$

alors

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

Démonstration. On a

$$\begin{aligned} W(\mu, q) &= \mathbb{E}[d_\phi(X, q(X))] \\ &= \sum_{j=1}^{\ell} \mathbb{E}[d_\phi(X, c_j)|X \in S_j]\mu(S_j) \\ &\geq \sum_{j=1}^{\ell} \mathbb{E}[d_\phi(X, c'_j)|X \in S_j]\mu(S_j) \\ &= \mathbb{E}[d_\phi(X, q'(X))] \\ &= W(\mu, q'), \end{aligned}$$

d'où le résultat annoncé, puisque la Proposition 1.4.1 entraîne $c'_j = \mathbb{E}[X|X \in S_j]$ pour tout $j = 1, \ldots, \ell$. □

Remarque 1.4.3 (Divergences de Bregman ponctuelles). Les propriétés énoncées dans les Lemmes 1.4.1 et 1.4.2 s'adaptent au cas d'une divergence de Bregman ponctuelle $\tilde{d}_f$.

Dans la section suivante, nous allons voir que sous certaines conditions, l'existence d'un quantificateur optimal est assurée. En pratique, trouver un minimiseur exact de la distorsion est un problème que l'on ne peut résoudre en temps polynomial, mais les Lemmes 1.4.1 et 1.4.2 montrent que la solution peut être approchée à l'aide d'un algorithme itératif, reposant sur deux étapes, au cours desquelles la table de codage et la partition sont actualisées successivement. Il s'agit, pour le carré de la distance euclidienne, de l'algorithme des k-means (Lloyd [132], Steinhaus [174]), que Linde, Buzo et Gray [129] généralisent à certaines autres mesures de distorsion, et que les deux lemmes permettent d'étendre aux divergences de Bregman. Explicitons l'algorithme, lorsque μ est inconnue, c'est-à-dire dans le cas du clustering. A partir d'une table de codage initiale $\{c_{0,1}, \ldots, c_{0,\ell}\}$, $\mathcal{C}$ est partitionné en cellules de Voronoi $S_{0,1}, \ldots, S_{0,\ell}$ en affectant chaque donnée X_i au centre $c_{0,j}$ le plus proche au sens de la divergence de Bregman considérée. Ensuite, les nouveaux centres $c_{1,1}, \ldots, c_{1,\ell}$ sont calculés en effectuant la moyenne des X_i tombés dans la cellule S_j, et ces deux étapes sont itérées, jusqu'au moment où la table de codage demeure inchangée, ce qui signifie qu'un minimum local a été atteint.

1.4.2. Existence d'un minimiseur de la distorsion

Nous cherchons à présent des conditions qui garantissent l'existence d'un quantificateur optimal $q^\star$, c'est-à-dire tel que $W(\mu, q^\star) = W^\star(\mu)$. D'après ce qui précède, nous savons qu'un quantificateur optimal est à rechercher parmi les quantificateurs des plus proches voisins. Un k-quantificateur des plus proches voisins q étant caractérisé par sa table de codage $\mathbf{c} = (c_1, \ldots, c_k)$, notre but est de démontrer

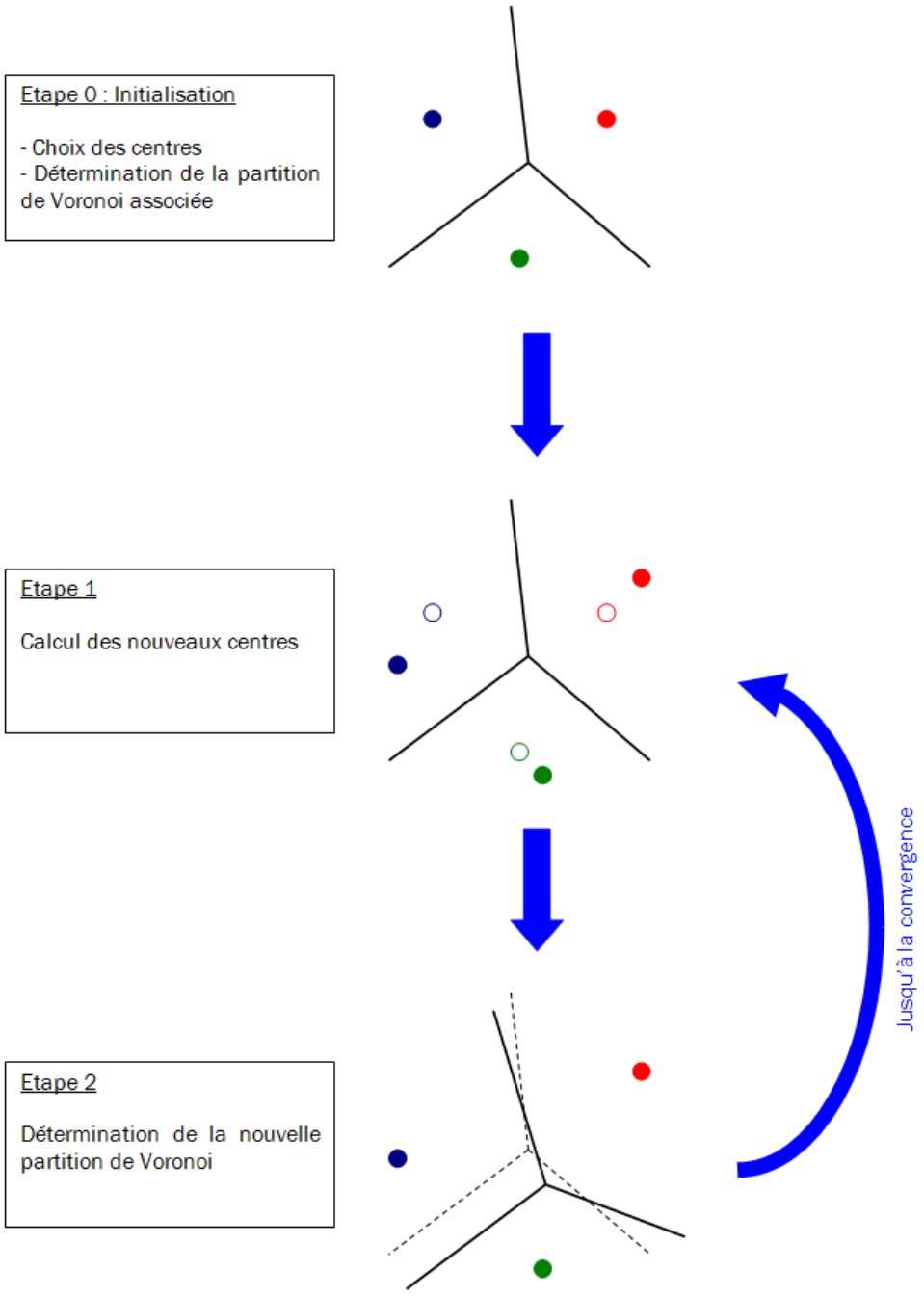


FIGURE 1.2.: Etapes de l'algorithme des k-means.

l'existence d'une table de codage $\mathbf{c}^\star$ optimale, autrement dit une table de codage $\mathbf{c}^\star$ telle que

$$W(\mu, \mathbf{c}^\star) = W^\star(\mu).$$

Le fait que le minimum de la distorsion soit atteint repose sur un argument de compacité. Nous distinguons le cas fini-dimensionnel (Théorème 1.4.1) du cas général (Théorème 1.4.2). En dimension finie, nous démontrons le résultat en utilisant une idée de Sabin et Gray [165], basée sur la compactification d'Alexandroff. Les résultats utiles liées à la compacité sont rappelés en Annexe 1.8.3, ainsi que la définition et quelques propriétés des fonctions semi-continues inférieurement.

Théorème 1.4.1 (Cas fini-dimensionnel)**.** *Supposons que $\mathcal{C}$ est inclus dans un sous-espace affine de dimension finie et que la divergence de Bregman $d_\phi(\cdot,\cdot)$ vérifie les propriétés suivantes :*

1. *Pour tout $x \in \mathcal{C}$, la fonction $y \mapsto d_\phi(x,y)$ est semi-continue inférieurement sur $ir(\mathcal{C})$.*
2. *Pour tout $(x,y) \in \mathcal{C} \times ir(\mathcal{C})$, $d_\phi(x,y) \leq \liminf_{z \in ir(\mathcal{C}) \to \tilde{z}} d_\phi(x,z)$ pour tout $\tilde{z} \in \partial\mathcal{C}$.*
3. *Pour tout $(x,y) \in \mathcal{C} \times ir(\mathcal{C})$, $d_\phi(x,y) \leq \liminf_{\|z\| \to +\infty} d_\phi(x,z)$.*

Alors, il existe une table de codage optimale $\mathbf{c}^\star$, c'est-à-dire telle que

$$W(\mu, \mathbf{c}^\star) = W^\star(\mu).$$

Remarquons que l'hypothèse 1 n'est pas restrictive, car $y \mapsto d_\phi(x,y)$ est continue pour la plupart des divergences de Bregman usuelles. Comme ϕ et $y \mapsto D_y\phi$ sont continues sur $ir(\mathcal{C})$, cette condition pourrait être remplacée par la semi-continuité inférieure de $y \mapsto D_y\phi(y)$. Le rôle des hypothèses 2 et 3 est d'empêcher un possible minimiseur de se trouver à l'infini. Observons enfin que la condition 3 est vide lorsque $\mathcal{C}$ est borné. Dans ce cas, $\overline{\mathcal{C}}$ est compact, et l'existence d'une table de codage optimale se démontre facilement sans recourir à la compactification d'Alexandroff.

Démonstration. En posant $d_\phi(x,\tilde{z}) = \liminf_{z \to \tilde{z} \in \partial\mathcal{C}} d_\phi(x,z)$ pour tout $x \in \mathcal{C}$ et tout $\tilde{z} \in \partial\mathcal{C}$, $d_\phi(\cdot,\cdot)$ se prolonge en une fonction semi-continue inférieurement $\overline{\mathcal{C}} \to [0,+\infty]$. On compactifie $\overline{\mathcal{C}}$ en lui ajoutant un point à l'infini ω. Notons $\tilde{\mathcal{C}} = \overline{\mathcal{C}} \cup \{\omega\}$ le compactifié d'Alexandroff de $\overline{\mathcal{C}}$. D'après le Théorème de Tychonoff (Théorème 1.8.3 de l'Annexe 1.8.3), le produit $\tilde{\mathcal{C}}^k$ est lui aussi compact. On pose, pour tout $x \in \mathcal{C}$, $d_\phi(x,\omega) = \lim_{\|z\| \to +\infty} d_\phi(x,z)$. D'après les hypothèses, pour tout $x \in \mathcal{C}$, $y \mapsto d_\phi(x,y)$ de $\tilde{\mathcal{C}}$ dans $[0,+\infty]$ est semi-continue inférieurement, ce qui signifie que l'ensemble de niveau $\{c \in \tilde{\mathcal{C}}, d_\phi(x,c) \leq \lambda\}$ est fermé pour tout $\lambda \in \mathbb{R}$. Comme $\{\mathbf{c} \in \tilde{\mathcal{C}}^k, \min_{j=1,\dots,k} d_\phi(x,c_j) \leq \lambda\} = \bigcup_{j=1}^k \{\mathbf{c} \in \tilde{\mathcal{C}}^k, d_\phi(x,c_j) \leq \lambda\}$, les

ensembles de niveau de $\mathbf{c} \mapsto \min_{j=1,\dots,k} d_\phi(x, c_j)$ sont fermés également, et ainsi, cette fonction est semi-continue inférieurement. Alors, pour tout $\mathbf{c} \in \tilde{\mathcal{C}}^k$,

$$\begin{aligned}
\liminf_{\mathbf{c}' \to \mathbf{c}} W(\mu, \mathbf{c}') &= \liminf_{\mathbf{c}' \to \mathbf{c}} \int \min_{j=1,\dots,k} d_\phi(x, c'_j) \mathrm{d}\mu(x) \\
&\geq \int \liminf_{\mathbf{c}' \to \mathbf{c}} \min_{j=1,\dots,k} d_\phi(x, c'_j) \mathrm{d}\mu(x) \\
&\geq \int \min_{j=1,\dots,k} d_\phi(x, c_j) \mathrm{d}\mu(x) \\
&= W(\mu, \mathbf{c}),
\end{aligned}$$

où la première inégalité découle du Lemme de Fatou et la seconde de la semi-continuité inférieure de la fonction $\mathbf{c} \mapsto \min_{j=1,\dots,k} d_\phi(x, c_j)$. Ainsi, $\mathbf{c} \mapsto W(\mu, \mathbf{c})$ est semi-continue inférieurement sur le compact $\tilde{\mathcal{C}}$ et donc y atteint son minimum en une table de codage $\mathbf{c}^\star$. Par les hypothèses 2 et 3, on peut supposer que $\mathbf{c}^\star \in ir(\mathcal{C})^k$, quitte à remplacer les composantes appartenant à $\partial\mathcal{C}$ ou égales à ω par des éléments de $ir(\mathcal{C})$. Finalement, l'existence d'une table de codage optimale $\mathbf{c}^\star$ est établie. □

Lorsque l'espace E est potentiellement de dimension infinie et $\mathcal{C}$ est un convexe quelconque de E, nous ne pouvons pas procéder de la même manière. En effet, la compactification d'Alexandroff s'applique aux espaces localement compacts alors que le Théorème de Riesz affirme qu'un espace vectoriel normé de dimension infinie n'est jamais localement compact (Théorème 1.8.2 de l'Annexe 1.8.3). Cependant, comme E est réflexif, un convexe fermé borné de E est compact pour la topologie faible $\sigma(E, E')$ (Corollaire 1.8.2). De plus, une fonction faiblement semi-continue inférieurement atteint sont minimum sur un ensemble faiblement compact. Donc, si nous savons d'avance que $\mathbf{c}^\star$ est à rechercher dans un ensemble compact pour la topologie faible, il suffit d'une hypothèse de continuité pour assurer l'existence du minimum. Désormais, $\mathcal{C}_R \subset ir(\mathcal{C})$ désigne un convexe fermé (borné) inclus dans $B(0, R) = \{x \in E, \|x\| \leq R\}$, la boule fermée de centre 0 et de rayon $R > 0$. Une propriété intéressante qui sera utilisée est que, si $X \in \mathcal{C}_R$, alors par projection (Proposition 1.3.8), si $\mathbf{c}^\star$ existe, on a aussi $\mathbf{c}^\star \in \mathcal{C}_R$.

Exemple 1.4.1 Voici quelques exemples de classes de variables aléatoires à valeurs dans l'espace de Banach E (de dimension infinie) telles que

$$\mathbb{P}\{\|X\| \leq R\} = 1. \tag{1.6}$$

1. Variable aléatoire tronquée. Pour toute variable aléatoire X à valeurs dans E, la variable $X\mathbf{1}_{\{\|X\| \leq R\}}$, $R > 0$, vérifie l'hypothèse (1.6).
2. Série dans un espace de Hilbert. Si E est un espace de Hilbert (séparable) et $(\psi_k)_{k \geq 1}$ désigne une base hilbertienne de E, les variables aléatoires de la forme $\sum_{k=1}^{+\infty} A_k \psi_k$, où $\sum_{k=1}^{+\infty} A_k^2 \leq R^2$, conviennent.

3. Fonctions bruitées. Les variables aléatoires modélisant l'évolution temporelle de quantités physiques mesurées avec un bruit fournissent un autre exemple. Si X est la somme d'une fonction g, $\|g\| \leq M$, et d'une variable aléatoire ε, centrée et à support compact, modélisant le bruit, la condition (1.6) est satisfaite.

Théorème 1.4.2 (Cas général)**.** *Supposons qu'il existe un réel $R > 0$ tel que $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$ et que, pour tout $x \in \mathcal{C}$, $y \mapsto d_\phi(x, y)$ est faiblement semi-continue inférieurement. Alors, il existe un k-quantificateur des plus proches voisins de table de codage optimale $\mathbf{c}^\star$, c'est-à-dire*

$$W(\mu, \mathbf{c}^\star) = W^\star(\mu).$$

Exemple 1.4.1. Comme exemples de fonctions semi-continues inférieurement pour la topologie faible $\sigma(E, E')$, on peut citer les fonctions convexes semi-continues inférieurement pour la norme (Corollaire 1.8.1).

Puisque la topologie faible et la topologie forte coïncident en dimension finie (Remarque 1.8.2), le mot "faiblement" dans le Théorème 1.4.2 peut être omis si E est de dimension finie.

Démonstration du Théorème 1.4.2. L'hypothèse $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$ entraîne qu'il suffit de chercher un minimiseur $\mathbf{c}^\star$ de la distorsion sur $\mathcal{C}_R^k$. En effet, d'après la Proposition 1.3.8, on a

$$\forall c \in ir(\mathcal{C}), d_\phi(X, c) \geq d_\phi(X, \overline{c}),$$

où $\overline{c}$ désigne la projection de Bregman de c sur $\mathcal{C}_R$. Par conséquent, pour toute table de codage $\mathbf{c}$, en notant $\overline{\mathbf{c}} = (\overline{c}_1, \ldots, \overline{c}_k)$ le vecteur des projections sur $\mathcal{C}_R$, on a $\mathbb{E}\left[\min_{j=1,\ldots,k} d_\phi(X, c_j)\right] \geq \mathbb{E}\left[\min_{j=1,\ldots,k} d_\phi(X, \overline{c}_j)\right]$, c'est-à-dire $W(\mu, \mathbf{c}) \geq W(\mu, \overline{\mathbf{c}})$, ce qui montre que l'on réduit la distorsion en projetant sur le convexe fermé borné $\mathcal{C}_R$. Comme E est réflexif, $\mathcal{C}_R$ est compact pour la topologie faible $\sigma(E, E')$, donc $\mathcal{C}_R^k$ également. Montrons que $W(\mu, \cdot)$ est faiblement semi-continue inférieurement. Par hypothèse, pour tout $x \in \mathcal{C}$, $d_\phi(x, \cdot)$ est semi-continue inférieurement pour la topologie faible, ce qui signifie que les ensembles de niveau $\{c \in \mathcal{C}_R, d_\phi(x, c) \leq \lambda\}$, $\lambda \in \mathbb{R}$, sont faiblement fermés. Puisque $\{\mathbf{c} \in \mathcal{C}_R^k, \min_{j=1,\ldots,k} d_\phi(x, c_j) \leq \lambda\} = \bigcup_{j=1}^k \{\mathbf{c} \in \mathcal{C}_R^k, d_\phi(x, c_j) \leq \lambda\}$, les ensembles de niveau de la fonction $\mathbf{c} \mapsto \min_{j=1,\ldots,k} d_\phi(x, c_j)$ sont eux aussi faiblement fermés et ainsi elle est faiblement semi-continue inférieurement. Si $\mathbf{c}'$ converge faiblement vers $\mathbf{c}$, on a

$$\begin{aligned}
\liminf_{\mathbf{c}' \to \mathbf{c}} W(\mu, \mathbf{c}') &= \liminf_{\mathbf{c}' \to \mathbf{c}} \int \min_{j=1,\ldots,k} d_\phi(x, c'_j) \mathrm{d}\mu(x) \\
&\geq \int \liminf_{\mathbf{c}' \to \mathbf{c}} \min_{j=1,\ldots,k} d_\phi(x, c'_j) \mathrm{d}\mu(x) \\
&\geq \int \min_{j=1,\ldots,k} d_\phi(x, c_j) \mathrm{d}\mu(x) = W(\mu, \mathbf{c}),
\end{aligned}$$

où la première inégalité découle du lemme de Fatou et la seconde de la semi-continuité de $\mathbf{c} \mapsto \min_{j=1,\dots,k} d_\phi(x, c_j)$. Il en résulte que $W(\mu, \cdot)$ est faiblement semi-continue inférieurement sur un ensemble faiblement compact, ce qui implique qu'elle y atteint sa borne inférieure. Autrement dit, il existe $\mathbf{c}^\star \in \mathcal{C}_R^k$, tel que $W(\mu, \mathbf{c}^\star) = W^\star(\mu)$. □

Remarque 1.4.4 (Divergences de Bregman ponctuelles). Le Théorème 1.4.2 se démontre de la même manière pour une divergence de Bregman ponctuelle $\tilde{d}_f$.

Lorsque nous avons seulement $\mathcal{C}_R \cap ir(\mathcal{C}) \neq 0$ (au lieu de $\mathcal{C}_R \subset ir(\mathcal{C})$), notons que si ϕ est une fonction de Legendre, le projeté d'un élément de $ir(\mathcal{C})$ appartient à $ir(\mathcal{C})$ (Remarque 1.3.6), de sorte qu'il est encore possible d'utiliser la projection de Bregman pour démontrer l'existence d'un quantificateur optimal, à condition de pouvoir prolonger les fonctions $y \mapsto d(x, y)$ en des fonctions faiblement semi-continues inférieurement sur le convexe fermé borné $\mathcal{C}_R \cap \overline{\mathcal{C}}$.

Le Lemme 1.4.3 ci-dessous, dont la preuve est donnée dans l'Annexe 1.7, assure que dans le cas particulier où $d_\phi(\cdot, \cdot)$ est la distance au carré induite par le produit scalaire d'un espace de Hilbert, chercher un quantificateur optimal revient à le chercher sur une boule.

Lemme 1.4.3. *Soit d_ϕ une divergence de Bregman. On suppose que la différentielle seconde de $\phi : E \to \mathbb{R}$ est uniformément coercive, c'est-à-dire qu'il existe $m = m(\phi) > 0$ tel que pour tout c, $D_c^2\phi(x, x) \geq m\|x\|^2$, et qu'il existe $M = M(\phi)$ tel que pour tout c, on ait $\|D_c^2\phi\| \leq M$. Alors,*

$$\inf_{\mathbf{c} \in E^k} W(\mu, \mathbf{c}) = \inf_{\mathbf{c} \in B_R^k} W(\mu, \mathbf{c})$$

pour un certain $R > 0$.

Le Théorème 1.4.2 admet par conséquent le corollaire suivant.

Corollaire 1.4.1. *Soit E un espace de Hilbert. Si $\phi = \|\cdot\|^2$, il existe un quantificateur optimal associé à la divergence de Bregman $d_\phi(\cdot, \cdot)$.*

La dernière partie de cette section est consacrée à la question de l'existence d'un quantificateur empirique optimal. En d'autres termes, nous cherchons un minimiseur $\mathbf{c}_n^\star$ de la distorsion empirique

$$W(\mu_n, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,k} d_\phi(X_i, c_j).$$

Comme le support de la mesure empirique μ_n contient au plus n points, il est inclus dans une boule fermée B_R. Ainsi, le Théorème 1.4.2 entraîne le résultat suivant.

Corollaire 1.4.2. *Supposons que pour tout $x \in \mathcal{C}$, $y \mapsto d_\phi(x, y)$ est faiblement semi-continue inférieurement. Alors, il existe une table de codage optimale $\mathbf{c}_n^\star$.*

Comme précédemment, le mot « faiblement » peut être omis dès que E est de dimension finie.

Démonstration du Corollaire 1.4.2. Le support de la mesure empirique μ_n, formé d'au plus n points, est contenu dans une boule fermée B_R. Donc, par projection, comme dans la démonstration du Théorème 1.4.2, il suffit de chercher la table de codage optimale dans cette boule. L'existence de $\mathbf{c}_n^\star$ résulte de la compacité faible de B_R, comme pour le Théorème 1.4.2. □

1.5. Convergence

1.5.1. Convergence vers le minimum de distorsion

Supposons qu'il existe une table de codage $\mathbf{c}_n^\star$ qui réalise le minimum de la distorsion empirique $W(\mu_n, \mathbf{c})$. Pour évaluer la qualité du quantificateur correspondant, nous nous intéressons à la « vraie » distorsion $W(\mu, \mathbf{c})$, prise en $\mathbf{c} = \mathbf{c}_n^\star$. Plus précisément, il s'agit de déterminer si $W(\mu, \mathbf{c}_n^\star)$ s'approche de la distorsion minimale $W^\star(\mu)$ lorsque le nombre d'observations n devient grand.

Remarque 1.5.1. Dans tout ce qui suit, $\mathbf{c}_n^\star$ pourrait être remplacé par un δ_n-minimiseur de la distorsion empirique, c'est-à-dire une table de codage $\mathbf{c}_n$ telle que $W(\mu_n, \mathbf{c}_n) < W^\star(\mu_n) + \delta_n$, avec $\lim_{n\to+\infty} \delta_n = 0$.

En supposant l'existence de $\mathbf{c}^\star$, on a

$$\begin{aligned} W(\mu, \mathbf{c}_n^\star) - W^\star(\mu) &= W(\mu, \mathbf{c}_n^\star) - W(\mu, \mathbf{c}^\star) \\ &= W(\mu, \mathbf{c}_n^\star) - W(\mu_n, \mathbf{c}_n^\star) + W(\mu_n, \mathbf{c}_n^\star) - W(\mu, \mathbf{c}^\star) \\ &\leq W(\mu, \mathbf{c}_n^\star) - W(\mu_n, \mathbf{c}_n^\star) + W(\mu_n, \mathbf{c}^\star) - W(\mu, \mathbf{c}^\star) \\ &\leq 2 \sup_{\mathbf{c} \in ir(\mathcal{C})^k} |W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c})| \end{aligned}$$

Pour montrer que $W(\mu, \mathbf{c}_n^\star)$ converge vers $W^\star(\mu)$, il suffit donc de prouver que la quantité $\sup_{\mathbf{c} \in ir(\mathcal{C})^k} |W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c})|$ tend vers 0 lorsque n tend vers l'infini.

Comme dans la section précédente, nous distinguons le cas fini-dimensionnel (Théorème 1.5.1) du cas général 1.5.2).

Théorème 1.5.1 (Cas fini-dimensionnel)**.** *Supposons que $\mathcal{C}$ est inclus dans un espace affine de dimension finie et que les propriétés suivantes sont vérifiées :*

1. *La divergence de Bregman $d_\phi(\cdot,\cdot)$ est continue sur $\mathcal{C} \times ir(\mathcal{C})$.*
2. *Pour tout $x \in \mathcal{C}$ et tout $\tilde{z} \in \partial\mathcal{C}$, $\lim_{z \in ir(\mathcal{C}) \to \tilde{z}} d_\phi(x,z) = +\infty$.*
3. *Pour tout $x \in \mathcal{C}$, $\lim_{\|z\| \to +\infty} d_\phi(x,z) = +\infty$.*
4. *Pour tout $x \in \mathcal{C}$, la fonction $y \mapsto d_\phi(x,y)$ est convexe sur $ir(\mathcal{C})$.*

Alors, si $\mathbf{c}_n^\star$ est un minimiseur de la distorsion empirique, on a

$$\lim_{n \to +\infty} W(\mu, \mathbf{c}_n^\star) = W^\star(\mu) \quad p.s.$$

Remarquons que l'existence de $\mathbf{c}_n^\star$ (et $\mathbf{c}^\star$) est assurée sous ces hypothèses.

D'après la définition de ϕ, la condition 1 pourrait être remplacée par la continuité de $(x,y) \mapsto D_y\phi(x-y)$. Comme nous l'avons mentionné plus haut, l'hypothèse 4 n'est pas vérifiée pour toute divergence de Bregman.

Démonstration du Théorème 1.5.1. Il s'agit de prouver que $W(\mu_n, \cdot)$ converge uniformément vers $W(\mu, \cdot)$ en dehors d'un ensemble de probabilité nulle. La méthode employée est à nouveau inspirée de Sabin et Gray [175]. Comme dans la démonstration du Théorème 1.4.1, nous définissons la divergence de Bregman $d_\phi(\cdot,\cdot)$ sur $\mathcal{C} \times \tilde{\mathcal{C}}$, où $\tilde{\mathcal{C}}$ est le compactifié d'Alexandroff de $\overline{\mathcal{C}}$. Les hypothèses entraînent que la fonction ainsi prolongée $d_\phi(\cdot,\cdot)$ est continue. D'après la Proposition 1.8.2 de l'Annexe 1.8.3, comme $\tilde{\mathcal{C}}^k$ est compact, il suffit de montrer que si $(\mathbf{c}_n)_{n \in \mathbb{N}}$ est une suite de points de $\tilde{\mathcal{C}}^k$ convergeant vers $\mathbf{c}$, alors

$$\lim_{n \to +\infty} W(\mu_n, \mathbf{c}_n) = W(\mu, \mathbf{c}) \quad p.s.$$

D'après un théorème de Varadarayan (voir par exemple Dudley [79, Théorème 11.4.1]), la mesure empirique μ_n converge étroitement vers μ presque sûrement. Le Théorème de Représentation de Skorohod (Dudley [79, Théorème 11.7.2]) assure l'existence de variables aléatoires Y et Y_n définies sur un même espace de probabilité telles que Y soit de loi μ, Y_n de loi μ_n, et Y_n converge vers Y presque sûrement. Comme la fonction prolongée $d_\phi(\cdot,\cdot)$ est continue, $\min_{j=1,\dots,k} d_\phi(x_n, c_{nj})$ converge vers $\min_{j=1,\dots,k} d_\phi(x, c_j)$ lorsque $(x_n, \mathbf{c}_n)$ converge vers $(x, \mathbf{c})$. Ainsi, lorsque $\mathbf{c}_n$ converge vers $\mathbf{c}$, $\min_{j=1,\dots,k} d_\phi(Y_n, c_{nj})$ converge presque sûrement (et donc en loi) vers $\min_{j=1,\dots,k} d_\phi(Y, c_j)$. De plus, pour tout c, $d_\phi(Y_n, c)$ converge vers $d_\phi(Y, c)$ presque sûrement, donc aussi en loi.

Si pour tout $j = 1, \dots, k$, on a $c_j = \omega$ ou $c_j \in \partial\mathcal{C}$, alors $W(\mu, \mathbf{c}) = +\infty$. Par ailleurs, d'après le Lemme de Fatou,

$$\liminf_{n \to +\infty} W(\mu_n, \mathbf{c}_n) = \liminf_{n \to +\infty} \mathbb{E}\Big[\min_{j=1\dots,k} d_\phi(Y_n, c_{nj})\Big] \geq \mathbb{E}\Big[\min_{j=1\dots,k} d_\phi(Y, c_j)\Big] = W(\mu, \mathbf{c}).$$

Donc, $\lim_{n\to+\infty} W(\mu_n, \mathbf{c}_n) = +\infty = W(\mu, \mathbf{c})$.

Sinon, soit c_m un élément de $\mathbf{c}$ appartenant à $ir(\mathcal{C})$. Il existe dans $ir(\mathcal{C})$ un polyèdre convexe régulier centré en c_m, contenant les c_{nm} pour n assez grand (par exemple, un hypercube de dimension s centré en c_m, où s désigne la dimension du sous-espace affine engendré par $ir(\mathcal{C})$). Soit $\mathcal{V}$ l'ensemble fini de ses sommets. La fonction $y \mapsto d_\phi(x, y)$ étant supposée convexe, on a, pour n assez grand,

$$\min_{j=1,\dots,k} d_\phi(x, c_{nj}) \leq d_\phi(x, c_{nm}) \leq \sum_{v\in\mathcal{V}} d_\phi(x, v). \tag{1.7}$$

Par la loi forte des grands nombres, presque sûrement, pour tout $v \in \mathcal{V}$,

$$\mathbb{E}[d_\phi(Y_n, v)] = \int d_\phi(x, v)\mathrm{d}\mu_n(x) = \frac{1}{n}\sum_{i=1}^{n} d_\phi(X_i, v)$$

tend lorsque $n \to +\infty$ vers

$$\mathbb{E}[d_\phi(X, v)] = \mathbb{E}[d_\phi(Y, v)].$$

D'après Billingsley [34, Théorème 3.6], pour tout $v \in \mathcal{V}$, les $d_\phi(Y_n, v)$ sont uniformément intégrables. Ceci implique, par l'inégalité (1.7), que les $\min_{j=1,\dots,k} d_\phi(Y_n, c_{nj})$ sont également uniformément intégrables. Par [34, Théorème 3.5], $W(\mu_n, \mathbf{c}_n) = \mathbb{E}\left[\min_{j=1,\dots,k} d_\phi(Y_n, c_{nj})\right]$ tend donc presque sûrement vers $\mathbb{E}\left[\min_{j=1,\dots,k} d_\phi(Y, c_j)\right] = W(\mu, \mathbf{c})$. □

Pour passer à la dimension quelconque, nous supposons que $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$, comme pour l'existence d'un quantificateur optimal.

Théorème 1.5.2 (Cas général)**.** *Supposons que pour tout $x \in \mathcal{C}$, $y \mapsto d_\phi(x, y)$ est faiblement semi-continue inférieurement, de sorte qu'il existe un minimiseur $\mathbf{c}_n^\star$ de la distorsion empirique. S'il existe $R > 0$ tel que $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$, et $M = M(\phi, R) \geq 0$ tel que, pour tout $c \in \mathcal{C}_R$, $\|D_c\phi\| \leq M$, alors*

$$\lim_{n\to+\infty} W(\mu, \mathbf{c}_n^\star) = W^\star(\mu) \quad p.s.$$

et

$$\lim_{n\to+\infty} \mathbb{E}[W(\mu, \mathbf{c}_n^\star)] = W^\star(\mu).$$

Démonstration. Comme $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$, les centroïdes restent dans le convexe fermé borné $\mathcal{C}_R$ comme le montre la preuve du Théorème 1.4.2. Soient Y de loi

μ et Y_n de loi μ_n les variables aléatoires données par le Théorème de Skorohod (Dudley [79, Théorème 11.7.2]). Alors, pour toute table de codage $\mathbf{c}$,

$$\begin{aligned} W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c}) &= \mathbb{E}\Big[\min_{j=1,\dots,k} d_\phi(Y_n, c_j)\Big] - \mathbb{E}\Big[\min_{j=1,\dots,k} d_\phi(Y, c_j)\Big] \\ &= \mathbb{E}\Big[\min_{j=1,\dots,k} \big(\phi(Y_n) - \phi(c_j) - D_{c_j}\phi(Y_n - c_j)\big)\Big] \\ &\quad - \mathbb{E}\Big[\min_{j=1,\dots,k} \big(\phi(Y) - \phi(c_j) - D_{c_j}\phi(Y - c_j)\big)\Big] \\ &\leq \mathbb{E}[\phi(Y_n)] - \mathbb{E}[\phi(Y)] + \mathbb{E}\Big[\min_{j=1,\dots,k} D_{c_j}\phi(Y_n - Y)\Big] \\ &\leq \mathbb{E}[\phi(Y_n)] - \mathbb{E}[\phi(Y)] + M\mathbb{E}\|Y_n - Y\|. \end{aligned}$$

Or, on a

$$\mathbb{E}[\phi(Y_n)] = \int \phi(x)\mathrm{d}\mu_n(x) = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i).$$

Donc, par la loi forte des grands nombres, $\mathbb{E}[\phi(Y_n)]$ converge vers $\mathbb{E}[\phi(X)] = \mathbb{E}[\phi(Y)]$ presque sûrement. D'après l'inégalité triangulaire, $\|Y\|+\|Y_n\|-\|Y_n-Y\| \geq 0$. Par le Lemme de Fatou,

$$\liminf_{n\to+\infty} \mathbb{E}\left[\|Y\| + \|Y_n\| - \|Y_n - Y\|\right] \geq \mathbb{E} \lim_{n\to+\infty} \left[\|Y\| + \|Y_n\| - \|Y_n - Y\|\right] = 2\mathbb{E}\|Y\|.$$

De plus, par la loi des grands nombres, $\mathbb{E}\|Y_n\|$ converge vers $\mathbb{E}\|Y\|$ presque sûrement, ce qui implique

$$\mathbb{E}\|Y - Y_n\| \underset{n\to+\infty}{\longrightarrow} 0 \quad p.s. \tag{1.8}$$

Donc, presque sûrement,

$$\sup_{\mathbf{c}\in\mathcal{C}_R^k} |W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c})| \underset{n\to+\infty}{\longrightarrow} 0.$$

Ceci termine la démonstration du premier point.

Pour la seconde assertion, l'inégalité suivante montre qu'il suffit de prouver que $\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k}(W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c}))\right]$ tend vers 0 lorsque n tend vers l'infini (voir Devroye, Györfi et Lugosi [71]) :

$$\begin{aligned} &\mathbb{E}[W(\mu, \mathbf{c}_n^\star)] - \inf_{\mathbf{c}\in\mathcal{C}_R^k} W(\mu, \mathbf{c}) \\ &\qquad \leq \mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k} (W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c}))\right] + \mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c}))\right]. \end{aligned}$$

Nous avons vu plus haut que, pour tout $\mathbf{c} \in \mathcal{C}_R^k$,

$$W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c}) = \mathbb{E}[\phi(Y_n)] - \mathbb{E}[\phi(Y)] + M\mathbb{E}\|Y_n - Y\|.$$

De plus,

$$\mathbb{E}[\phi(Y_n)] - \mathbb{E}[\phi(Y)] = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i) - \mathbb{E}[\phi(X)],$$

et en prenant l'espérance par rapport aux X_i, on a

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\phi(X_i) - \mathbb{E}[\phi(X)]\right] = 0.$$

Il reste à montrer que l'espérance (par rapport aux X_i) de $\mathbb{E}\|Y_n - Y\|$ tend vers 0 lorsque n tend vers l'infini, ce qui découle du théorème de convergence dominée, car $\mathbb{E}\|Y - Y_n\|$ tend vers 0 presque sûrement d'après ce qui précède et, de plus, $\mathbb{E}\|Y_n - Y\| \leq 2R$. Finalement,

$$\lim_{n\to+\infty}\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k}(W(\mu_n,\mathbf{c}) - W(\mu,\mathbf{c}))\right] = 0.$$

□

Remarquons que les convergences

$$\lim_{n\to+\infty} W(\mu,\mathbf{c}_n^\star) = W^\star(\mu) \quad p.s.$$

et

$$\lim_{n\to+\infty}\mathbb{E}[W(\mu,\mathbf{c}_n^\star)] = W^\star(\mu)$$

ont toujours lieu dès que ϕ est le carré de la norme de l'espace de Banach séparable et réflexif E (Biau, Devroye et Lugosi [32]).

Remarque 1.5.2 (Divergences de Bregman ponctuelles). En remplaçant la fonction strictement convexe ϕ par $x \mapsto \int f(x)\mathrm{d}m$ et l'application linéaire $D_c\phi$ par $x \mapsto \int x f'(c)\mathrm{d}m$ dans la démonstration du Théorème 1.5.2, on obtient le résultat de convergence de la distorsion pour une divergence de Bregman ponctuelle.

Explicitons les résultats d'existence d'un quantificateur optimal et de convergence de la distorsion sur quelques exemples.

Exemple 1.5.1 1. Distance de Kullback-Leibler généralisée en dimension 1. Ici, $E = \mathbb{R}$, $\mathcal{C} = \mathbb{R}^+$ et $d_\phi(x,y) = x\ln\frac{x}{y} - (x-y)$. Soit $x \in \mathcal{C}$. La fonction $y \mapsto x\ln\frac{x}{y} - (x-y)$ est continue et convexe sur $ir(\mathcal{C}) = \mathbb{R}^{+*}$ (sa dérivée seconde est $\frac{x}{y^2} \geq 0$) et tend vers $+\infty$ en 0 et en $+\infty$. Donc il existe un quantificateur dont la table de codage réalise le minimum de la distorsion $W(\mu,\mathbf{c})$ (Théorème 1.4.1) ainsi qu'un quantificateur empirique optimal (Corollaire 1.4.2). En outre, si $\mathbf{c}_n^\star$ minimise la distorsion empirique, la convergence presque sûre de $W(\mu,\mathbf{c}_n^\star)$ vers $W^\star(\mu)$ est garantie (Théorème 1.5.1).

2. Perte exponentielle. Soient $\mathcal{C} = E = \mathbb{R}$ et $\phi(x) = e^x$, ce qui donne $d_\phi(x, y) = e^x - e^y - (x - y)e^y$. La fonction $y \mapsto e^x - e^y - (x - y)e^y$ est continue sur $\mathbb{R}$. Si $\mathbb{P}\{|X| \leq R\} = 1$, le Théorème 1.4.2 assure l'existence d'un quantificateur optimal, et comme $\phi'(x) = e^x \leq e^R$ sur $[-R, R]$, $W(\mu, \mathbf{c}_n^\star)$ converge presque sûrement et dans L^1 vers $W^\star(\mu)$ par le Théorème 1.5.2.

3. Distance euclidienne au carré. Lorsque $d_\phi(\cdot, \cdot)$ est le carré de la distance euclidienne, l'existence d'un quantificateur optimal et la convergence presque sûre et L^1 de la distorsion sont assurées (cas particulier des normes hilbertiennes).

4. Distance de Kullback-Leibler entre mesures de probabilité discrètes. Ici, $E = \mathbb{R}^d$ et $\mathcal{C} = (\mathbb{R}^+)^d$. Soit $\mathcal{S}_{d-1}$ le simplexe de dimension $d - 1$. Pour $(x, y) \in [\mathcal{C} \times ir(\mathcal{C})] \cap \mathcal{S}_{d-1}^2$, $d_\phi(x, y) = \sum_{\ell=1}^d x_\ell \ln \frac{x_\ell}{y_\ell}$. La fonction $y = (y_1, \ldots, y_d) \mapsto \sum_{\ell=1}^d x_\ell \ln \frac{x_\ell}{y_\ell}$ est continue et convexe sur $\mathcal{S}_{d-1} \cap (\mathbb{R}^{+*})^d$ et tend vers $+\infty$ lorsque l'un des y_ℓ tend vers 0. Donc il existe un quantificateur optimal et l'on a convergence presque sûre de la distorsion.

5. Norme L^2 au carré. Soit $\mathcal{C} = E = L^2([0, 1], \mathrm{d}t)$ et $d_\phi(x, y) = \int_0^1 (x(t) - y(t))^2 \mathrm{d}t$. Comme il s'agit d'une norme hilbertienne, l'existence d'un minimiseur de la distorsion et la convergence sont garanties.

6. Distance de Kullback-Leibler généralisée. Soit $E = L^2([0, 1], \mathrm{d}t)$. On a $\tilde{d}_f(x, y) = \int_0^1 [x(t) \ln \frac{x(t)}{y(t)} + y(t) - x(t)]\mathrm{d}t$ (définition ponctuelle). La fonction $y \mapsto \tilde{d}_f(x, y)$ est semi-continue inférieurement et convexe donc semi-continue inférieurement pour la topologie faible. Supposons que $\mathbb{P}\{r \leq \|X\| \leq R\} = 1$ $(r > 0)$. Alors, il existe un quantificateur optimal. De plus, on a convergence presque sûre et L^1 de la distorsion.

1.5.2. Vitesse de convergence

Les résultats de convergence de la section précédente signifient que $W(\mu, \mathbf{c}_n^\star)$ s'approche de la distorsion minimale lorsque le nombre d'observations devient grand. Cependant, ils ne donnent pas d'indication permettant d'évaluer la taille n de l'échantillon nécessaire pour que $W(\mu, \mathbf{c}_n^\star)$ soit effectivement très proche de $W^\star(\mu)$. C'est pourquoi nous nous intéressons à présent à la vitesse de convergence.

Remarquons tout d'abord que minimiser

$$W(\mu, \mathbf{c}) = \mathbb{E}\left[\min_{j=1,\ldots,k} d_\phi(X, c_j)\right] = \mathbb{E}\left[\min_{j=1,\ldots,k} \Big(\phi(X) - \phi(c_j) - D_{c_j}\phi(X - c_j)\Big)\right]$$

équivaut à minimiser

$$\overline{W}(\mu, \mathbf{c}) = \mathbb{E}\left[\min_{j=1,\ldots,k} \Big(-\phi(c_j) - D_{c_j}\phi(X - c_j)\Big)\right].$$

De même, nous associons à la distorsion empirique $W(\mu_n, \mathbf{c})$ la quantité

$$\overline{W}(\mu_n, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,k} \left(-\phi(c_j) - D_{c_j}\phi(X_i - c_j) \right).$$

Comme

$$W(\mu, \mathbf{c}_n^\star) - \inf_{\mathbf{c} \in ir(\mathcal{C})^k} W(\mu, \mathbf{c}) = \overline{W}(\mu, \mathbf{c}_n^\star) - \inf_{\mathbf{c} \in ir(\mathcal{C})^k} \overline{W}(\mu, \mathbf{c})$$

et

$$\begin{aligned} &\mathbb{E}\left[\overline{W}(\mu, \mathbf{c}_n^\star)\right] - \inf_{\mathbf{c} \in ir(\mathcal{C})^k} \overline{W}(\mu, \mathbf{c}) \\ &\leq \mathbb{E}\left[\sup_{\mathbf{c} \in ir(\mathcal{C})^k} \left(\overline{W}(\mu_n, \mathbf{c}) - \overline{W}(\mu, \mathbf{c}) \right) \right] + \mathbb{E}\left[\sup_{\mathbf{c} \in ir(\mathcal{C})^k} \left(\overline{W}(\mu, \mathbf{c}) - \overline{W}(\mu_n, \mathbf{c}) \right) \right] \end{aligned} \tag{1.9}$$

(voir Devroye, Györfi et Lugosi [71, Lemme 8.2]), on cherche à majorer la déviation maximale en espérance

$$\mathbb{E}\left[\sup_{\mathbf{c} \in \mathcal{C}_R^k} \left(\overline{W}(\mu_n, \mathbf{c}) - \overline{W}(\mu, \mathbf{c}) \right) \right].$$

Le second terme du membre de droite de l'inégalité (1.9) peut en effet être majoré par une borne du premier terme. Pour construire cette borne (Théorème 1.5.3 ci-dessous), nous utilisons les moyennes de Rademacher comme mesure de complexité pour une classe de fonctions (Bartlett, Boucheron et Lugosi [22], Koltchinskii [122]). L'espace de Banach E est supposé de type 2. La définition d'un espace de Banach de type p, $1 \leq p \leq 2$, ainsi que quelques propriétés intéressantes des moyennes de Rademacher sont rappelées dans l'Annexe A.

Théorème 1.5.3. *Supposons que l'espace de Banach E est de type 2, avec une constante T_2. Pour $\mathcal{C}_R \subset ir(\mathcal{C})$,*

$$\begin{aligned} &\mathbb{E}\left[\sup_{\mathbf{c} \in \mathcal{C}_R^k} \left(\overline{W}(\mu_n, \mathbf{c}) - \overline{W}(\mu, \mathbf{c}) \right) \right] \\ &\leq \frac{2k}{\sqrt{n}} \left[\sup_{c \in \mathcal{C}_R} | -\phi(c) + D_c\phi(c)| + T_2 \sup_{c \in \mathcal{C}_R} \|D_c\phi\| (\mathbb{E}\|X\|^2)^{1/2} \right]. \end{aligned}$$

Ce théorème résulte du Lemme 1.5.1, qui est prouvé dans l'Annexe 1.7.

Lemme 1.5.1. *Soient $\varepsilon_1, \ldots, \varepsilon_n$ des variables aléatoires de Rademacher indépendantes, indépendantes des X_i. Pour $c \in \mathcal{C}_R$, notons ℓ_c la fonction à valeurs réelles définie par*

$$\ell_c(x) = -\phi(c) - D_c\phi(x - c), \quad x \in \mathcal{C}.$$

Alors,

1. $$\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\left(\overline{W}(\mu_n, \mathbf{c}) - \overline{W}(\mu, \mathbf{c})\right)\right] \leq 2\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n \varepsilon_i \min_{j=1,\ldots,k} \ell_{c_j}(X_i)\right].$$

2. $$\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n \varepsilon_i \min_{j=1,\ldots,k} \ell_{c_j}(X_i)\right]$$
$$\leq k\left(\mathbb{E}\left[\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n \varepsilon_i D_c\phi(X_i)\right] + \frac{1}{\sqrt{n}}\sup_{c\in\mathcal{C}_R}|-\phi(c) + D_c\phi(c)|\right).$$

3. *Si E est de type 2, avec une constante T_2,*

$$\mathbb{E}\left[\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n \varepsilon_i D_c\phi(X_i)\right] \leq \frac{T_2}{\sqrt{n}}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|(\mathbb{E}\|X\|^2)^{1/2}.$$

Finalement, nous sommes en mesure d'établir le résultat de vitesse de convergence suivant pour notre problème de quantification avec une divergence de Bregman.

Corollaire 1.5.1. *Supposons que l'espace de Banach E est de type 2, avec une constante T_2, et que pour tout $x \in \mathcal{C}$, $y \mapsto d_\phi(x, y)$ est faiblement semi-continue inférieurement, ce qui garantit l'existence d'un quantificateur optimal $\mathbf{c}_n^\star$. Supposons de plus qu'il existe $R > 0$ tel que $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$. Alors, si $|-\phi(c) + D_c\phi(c)|$ et $\|D_c\phi\|$ sont uniformément bornés sur $\mathcal{C}_R$ par $M_1 = M_1(\phi, R) \geq 0$ et $M_2 = M_2(\phi, R) \geq 0$ respectivement, on a*

$$\mathbb{E}[W(\mu, \mathbf{c}_n^\star)] - W^\star(\mu) \leq \frac{4k}{\sqrt{n}}\left(M_1 + T_2M_2(\mathbb{E}\|X\|^2)^{1/2}\right),$$

et donc

$$\mathbb{E}[W(\mu, \mathbf{c}_n^\star)] - W^\star(\mu) \leq \frac{4k}{\sqrt{n}}\left(M_1 + T_2M_2R\right). \tag{1.10}$$

La borne (1.10), qui décroît vers 0 à la vitesse $1/\sqrt{n}$, ne fait pas intervenir la dimension de l'espace ambiant. Cette caractéristique est très intéressante puisque, comme nous l'avons mentionné plus haut, l'objectif peut être de classer des données de dimension élevée, voire infinie.

Remarquons que lorsque $\phi(x) = x^2$ ou $\phi(x) = e^x$, pour la distance euclidienne ou une distance L^2 (par exemple), $|-\phi(c)+D_c\phi(c)|$ et $\|D_c\phi\|$ sont bornées sur une boule fermée B_R. Notons, d'autre part, que $T_2 = 1$ si E est un espace de Hilbert.

Remarque 1.5.3 (Divergences de Bregman ponctuelles). En transposant aux fonctions $x \mapsto \int f(x)\mathrm{d}m$ et $x \mapsto \int xf'(c)\mathrm{d}m$ les hypothèses relatives à ϕ et $D_c\phi$, l'inégalité (1.10) est obtenue également pour une divergence de Bregman ponctuelle.

Exemple 1.5.2 Nous donnons ici les bornes obtenues pour quelques divergences de Bregman usuelles. Tout au long de cet exemple, nous supposons qu'il existe $R > 0$ tel que $\mathbb{P}\{\|X\| \leq R\} = 1$.

1. Distance euclidienne au carré en dimension 1. Pour $\phi(x) = x^2$, on a

$$\mathbb{E}[W(\mu, \mathbf{c}_n^\star)] - W^\star(\mu) \leq \frac{4k}{\sqrt{n}}\left(R^2 + 2R(\mathbb{E}|X|^2)^{1/2}\right),$$

   et donc

$$\mathbb{E}[W(\mu, \mathbf{c}_n^\star)] - W^\star(\mu) \leq \frac{12kR^2}{\sqrt{n}}.$$

2. Perte exponentielle. Pour $\phi(x) = e^x$,

$$\mathbb{E}[W(\mu, \mathbf{c}_n^\star)] - W^\star(\mu) \leq \frac{4k(2R-1)e^R}{\sqrt{n}}.$$

3. Distance euclidienne au carré. Pour la norme euclidienne au carré $\phi(x) = \|x\|^2$,

$$\mathbb{E}[W(\mu, \mathbf{c}_n^\star)] - W^\star(\mu) \leq \frac{12kR^2}{\sqrt{n}}.$$

4. Distance de Mahalanobis. Pour $\phi(x) = {}^t xAx$ avec A définie positive,

$$\mathbb{E}[W(\mu, \mathbf{c}_n^\star)] - W^\star(\mu) \leq \frac{12k\|A\|R^2}{\sqrt{n}}.$$

5. Distance L^2 au carré. Lorsque ϕ est une norme L^2 au carré,

$$\mathbb{E}[W(\mu, \mathbf{c}_n^\star)] - W^\star(\mu) \leq \frac{12kR^2}{\sqrt{n}}.$$

Remarque 1.5.4. Si nous considérons un δ_n-minimiseur de la distorsion empirique (voir Remarque 1.5.1) à la place de $\mathbf{c}_n^\star$, les majorations demeurent inchangées, à ceci près que s'ajoute le terme δ_n.

Concluons cette section en commentant les hypothèses. La première observation est relative à l'hypothèse de bornitude de la différentielle sur $\mathcal{C}_R$, et la seconde concerne plus généralement la condition $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$.

Remarque 1.5.5. Certaines divergences de Bregman, en particulier la divergence de Kullback-Leibler, font intervenir la fonction logarithme. La différentielle n'est alors pas uniformément bornée sur une boule B_R puisqu'un problème se pose au voisinage de 0. Pour contourner cette difficulté, une solution possible est de considérer une classe d'éléments de E vérifiant la condition :

- En dimension 1, $0 < r \leq x \leq R < +\infty \quad p.s.$
- En dimension d ($2 \leq d \leq \infty$), lorsque le logarithme apparaît dans une somme ou une intégrale, $\sum_{\ell=1}^{d} \ln^2(x_\ell) \leq M(R)$ ou $\int \ln^2(x(t))\mathrm{d}t \leq M(R)$.

On peut trouver plusieurs conditions de ce type dans la littérature sur la divergence de Kullback-Leibler. Par exemple, Jordan, Nguyen et Wainwright [112], qui développent un procédé d'estimation pour la divergence de Kullback-Leibler, se restreignent à une classe $\mathcal{F}$ de fonctions vérifiant la condition d'enveloppe $\int \sup_{f\in\mathcal{F}} |\ln f(t)|\mathrm{d}t < +\infty$ ou aux fonctions à la fois minorées et majorées.

A titre d'illustration, soit $d_\phi(x,y) = \int_0^1 x(t) \ln \frac{x(t)}{y(t)}\mathrm{d}t$. Si nous supposons qu'il existe $R > 0$ tel que $\mathbb{P}\{\|X\| \leq R\} = 1$ et que $\int \ln^2(X(t))\mathrm{d}t \leq R^2$, alors, en considérant des centres appartenant à la même classe de fonctions que X, nous obtenons

$$\mathbb{E}[W(\mu, \mathbf{c}_n^\star)] - W^\star(\mu) \leq \frac{2kR}{\sqrt{n}}(1+R).$$

Remarque 1.5.6. L'hypothèse

$$\mathbb{P}\{X \in \mathcal{C}_R\} = 1 \tag{1.11}$$

est fréquemment rencontrée dans la littérature sur le clustering. Elle apparaît en particulier dans l'étude de la vitesse de convergence, pour établir une borne non-asymptotique.

Dans ce chapitre, cette hypothèse est initialement introduite, dans le cas de la dimension infinie, dans le but de créer de la compacité afin d'obtenir l'existence d'un quantificateur optimal. Il est cependant inutile de supposer (1.11) pour prouver l'existence d'un quantificateur empirique optimal.

Observons aussi que, pour la convergence de la distorsion en dimension quelconque, nous n'utilisons pas directement l'hypothèse (1.11), mais plutôt le fait que

la différentielle soit uniformément bornée (la dernière étape de la démonstration pourrait en effet être effectuée différemment). Toutefois, par (1.11), il suffit de supposer $\|D_c\phi\|$ uniformément bornée sur $\mathcal{C}_R$, hypothèse plus souple que $\|D_c\phi\|$ uniformément bornée.

Pour la vitesse de convergence, la condition (1.11) est utilisée à la fois explicitement et au travers d'hypothèses sur ϕ et sa différentielle. Il est possible qu'elle puisse être affaiblie, par exemple en posant une condition du type $\mathbb{E}[X\mathbf{1}_{\{\|X\|\leq R\}}] \leq \varepsilon$, comme le font Merhav et Ziv [145], ou une condition de moment exponentiel, comme Cadre et Paris dans [44] (voir Biau, Devroye et Lugosi [32] et Laloë [125] pour d'autres résultats dans cette direction). Néanmoins, le lien entre (1.11) et les hypothèses sur la fonction ϕ et sa différentielle complique la situation.

1.6. Simulations

Dans cette section, nous proposons d'illustrer par quelques simulations le clustering avec différentes divergences de Bregman.

Jusqu'à présent, le nombre de groupes était fixé, et nous pouvons nous demander comment procéder pour le choisir en pratique. Différents moyens de sélectionner k ont été envisagés dans la littérature. Souvent, ces méthodes s'appuient sur le fait que la distorsion décroît plus fortement tant qu'incrémenter k permet de séparer deux vraies classes. Puis, lorsque la valeur de k augmente davantage, conduisant à éclater un groupe existant, la distorsion continue de décroître, mais moins significativement. Remarquons que les méthodes reposant sur un critère pénalisé de type BIC (Schwarz [167]) ou ICL (Biernacki, Celeux et Govaert [33]), utilisées dans le cas des modèles de mélange, ne peuvent s'appliquer ici puisqu'il n'est fait aucune hypothèse paramétrique sur la loi μ. Nous reviendrons sur la question du choix du nombre de classes dans le Chapitre 3.

Dans les simulations de cette section, réalisées avec le logiciel R, nous avons décidé d'utiliser la méthode de la *Gap Statistic* de Tibshirani, Walther et Hastie [181], qui présente l'avantage de s'appliquer à une méthode de clustering quelconque. Cette démarche est basée sur la quantité

$$w_k = \sum_{j=1}^{k} \frac{1}{2n_j} \sum_{(X_{i_1}, X_{i_2}) \in S_j^2} d_\phi(X_{i_1}, X_{i_2}),$$

où $n_j = |S_j|$ pour tout j.

Remarque 1.6.1. Pour le carré d'une norme hilbertienne, pour $j = 1, \dots, k$,

$$\begin{aligned}\sum_{i=1}^{n} \|X_i - c_j\|^2 \mathbf{1}_{\{X_i \in S_j\}} &= \sum_{X_i \in S_j} \|X_i - c_j\|^2 \\ &= \sum_{X_{i_1} \in S_j} \|X_{i_1} - \frac{1}{n_j} \sum_{X_{i_2} \in S_j} X_{i_2}\|^2 \\ &= \frac{1}{2n_j} \sum_{(X_{i_1}, X_{i_2}) \in S_j^2} \|X_{i_1} - X_{i_2}\|^2,\end{aligned}$$

de sorte que w_k est égal à n fois la distorsion empirique.

La méthode de la *Gap Statistic* consiste à comparer la courbe de $\ln(w_k)$ avec la courbe de l'espérance de $\ln(w_k^\star)$, équivalent de $\ln(w_k)$ pour des observations uniformément distribuées (qu'il n'est donc pas pertinent de classer en groupes). En pratique, cette espérance est estimée en simulant m jeux de n données uniformément distribuées sur l'ensemble dans lequel la variable aléatoire X prend ses valeurs. Dans toutes les simulations, nous prenons $m = 20$ comme proposé par Hastie, Tibshirani et Friedman [105]. La *Gap Statistic* est la quantité

$$\mathrm{Gap}(k) = \frac{1}{m} \sum_{\ell=1}^{m} \ln(w_{k\ell}^\star) - \ln(w_k),$$

où $w_{k1}^\star, \dots, w_{km}^\star$ désignent les jeux de données simulées de loi uniforme. En notant alors

$$v = \frac{1}{m} \sum_{\ell=1}^{m} \Big(\ln(w_{k\ell}^\star) - \frac{1}{m} \sum_{\ell=1}^{m} \ln(w_{k\ell}^\star) \Big)^2$$

et $s_k = \sqrt{v(1 + \frac{1}{m})}$, le $\hat{k}$ choisi est la plus petite valeur de k telle que

$$\mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - s_{k+1}.$$

Le terme s_{k+1} a ici pour fonction de corriger l'erreur de simulation dans le calcul de l'espérance de $\ln(w_k^\star)$ (voir Tibshirani, Walther et Hastie [181] pour les justifications).

A l'aide du logiciel R, nous allons illustrer le choix du nombre de groupes au moyen de la *Gap Statistic* par deux exemples, l'un en dimension finie, l'autre en dimension infinie.

Un exemple en dimension finie : données gaussiennes

Dans cet exemple, les observations sont des réalisations de 3 vecteurs gaussiens de même matrice de variance-covariance égale à la matrice identité et de moyennes différentes $(5, 5)$, $(8, 8)$ et $(10, 3)$ (Figure 1.3).

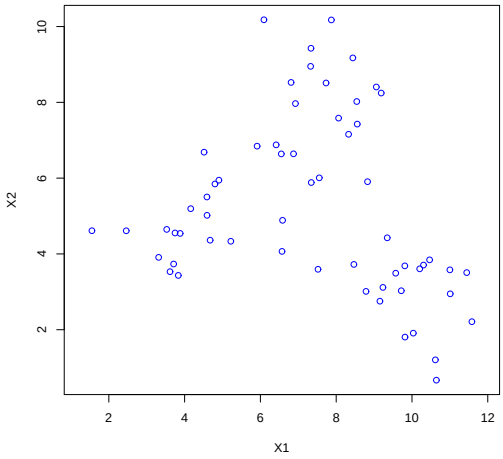


FIGURE 1.3.: Un ensemble d'observations dans le plan (k=3, n=60).

La Figure 1.4 montre les courbes de la *Gap Statistic* obtenues pour différentes divergences de Bregman. Chacune de ces courbes correspond à la différence entre $\ln(w_k)$ pour les données $X_1, \ldots, X_n$ et l'espérance de cette même quantité pour des données uniformément distribuées. Les nombres de groupes estimés, identiques sur plusieurs essais, sont donnés dans le Tableau 1.1.

Divergence de Bregman	Euclidien	I-divergence	Itakura-Saito	Logistique
Nombre de groupes k	3	3	1	3

TABLEAU 1.1.: Résultats de la méthode de la *Gap Statistic* pour les données gaussiennes (k=3, n=60).

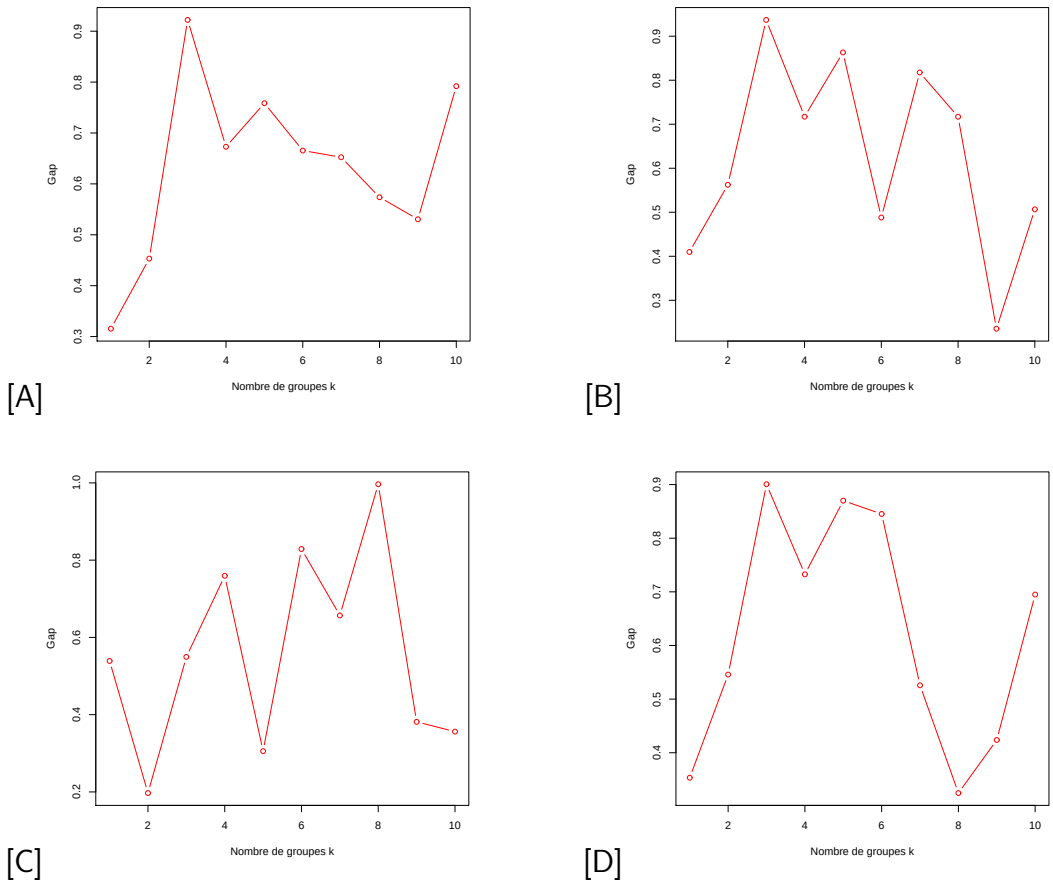


FIGURE 1.4.: Courbes de la *Gap Statistic*. [A] Carré de la distance euclidienne. [B] Distance de Kullback-Leibler généralisée ou I-divergence. [C] Distance de Itakura-Saito. [D] Perte logistique.

Un exemple en dimension infinie : sinusoïdes bruitées

Les courbes de la Figure 1.5 ont été construites comme 3 ensembles de 15 sinusoïdes de phase 0, $\pi/8$ et $\pi/4$ respectivement. La Figure 1.6 et le Tableau 1.2 présentent les résultats de la *Gap Statistic*.

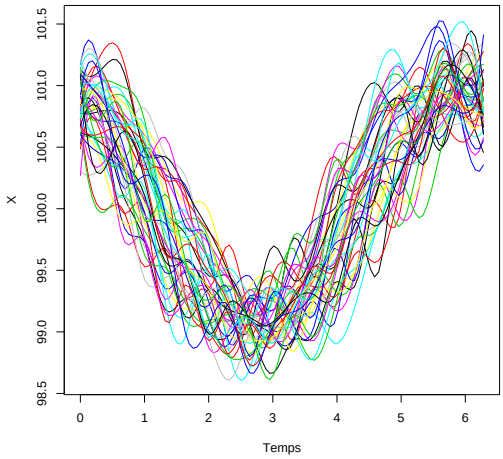


FIGURE 1.5.: Une collection de sinusoïdes bruitées (k=3,n=45).

Norme 2	Biais quadratique	Itakura-Saito	I-divergence
3	6	3	3

TABLEAU 1.2.: Résultats de la méthode de la *Gap Satistic* pour les sinusoïdes (k=3, n=45).

Sur ces exemples, on constate une certaine stabilité par rapport aux différentes divergences de Bregman dans le nombre de groupes k sélectionné. Dans nos simulations, la valeur de k utilisée sera celle choisie pour la plupart des divergences considérées.

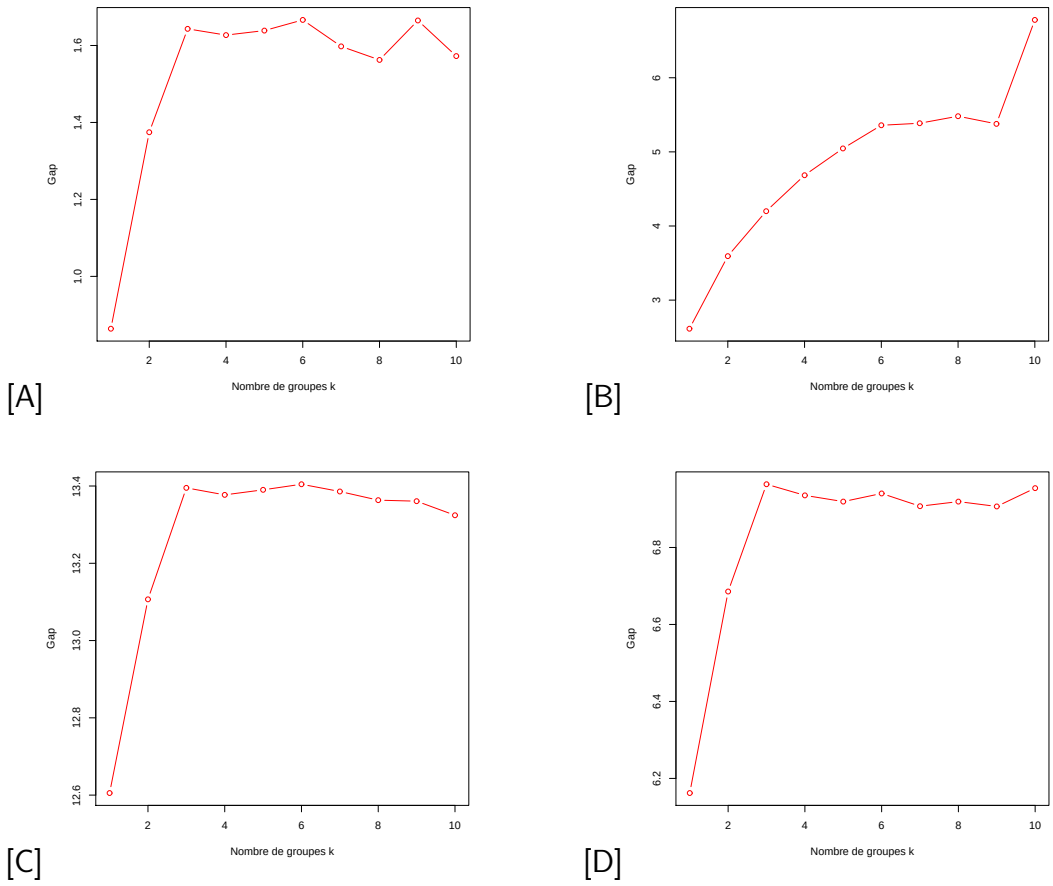


FIGURE 1.6.: Courbes de la *Gap Statistic*. [A] Carré de la distance L^2. [B] Biais quadratique. [C] Distance de Itakura-Saito. [D] Distance de Kullback-Leibler généralisée.

Par ailleurs, si la distorsion donne une idée de la qualité du clustering effectué, elle ne peut servir à comparer les partitions obtenues en utilisant différentes divergences de Bregman. En effet, comme la distorsion dépend de la divergence de Bregman choisie, il est possible qu'une divergence conduise à des groupes meilleurs qu'une autre, bien que l'erreur associée soit plus grande. Pour cette raison, Banerjee *et al.* [17] utilisent pour mesurer la qualité des partitions l'information mutuelle normalisée, critère proposé par Strehl et Ghosh dans [175]. Cette quantité permet d'évaluer la corrélation entre deux partitions S et S'. Si, dans les simulations, on prend pour S la partition attendue, et pour S' celle retournée par l'ordinateur, nous pouvons ainsi juger de la précision de la méthode de clustering. Evidemment, dans notre contexte d'apprentissage non supervisé, le fait de se donner une partition de référence semble inadapté. Cependant, si l'on veut comparer les performances de plusieurs divergences de Bregman, il n'est guère possible de procéder différemment. En notant $n_{j,\ell}$ le nombre de données appartenant à la fois à S_j et à S'_ℓ, $n_j = |S_j|$ et $n_\ell = |S'_\ell|$ les cardinaux des cellules, l'information mutuelle normalisée s'écrit

$$\frac{\sum_{j=1}^{k}\sum_{\ell=1}^{k} n_{j,\ell} \ln\left(\frac{n_{j,\ell} n}{n_j n_\ell}\right)}{\sqrt{\left(\sum_{j=1}^{k} n_j \ln\frac{n_j}{n}\right)\left(\sum_{\ell=1}^{k} n_\ell \ln\frac{n_\ell}{n}\right)}}.$$

Le résultat du clustering est d'autant meilleur que ce coefficient de corrélation est proche de 1.

1.6.1. Simulations en dimension finie

Cette sous-section propose une première série d'exemples en dimension 1 ou dans le plan $\mathbb{R}^2$.

Lois gaussienne, binomiale et loi Poisson

Comme nous l'avons indiqué au paragraphe 1.3.3, Banerjee *et al.* [17] ont montré qu'il existe une relation entre familles exponentielles et divergences de Bregman (voir l'Annexe 1.8.4). Le Tableau 1.3 illustre par exemple cette propriété pour la loi gaussienne, la loi binomiale et la loi de Poisson.

Carré de la distance euclidienne	$\leftrightarrow$	Loi gaussienne
Perte logistique	$\leftrightarrow$	Loi binomiale
I-divergence	$\leftrightarrow$	Loi de Poisson

TABLEAU 1.3.: Exemples de correspondances entre lois de la famille exponentielle et divergences de Bregman.

Dans la première simulation destinée à illustrer la relation entre lois et divergences de Bregman, nous allons comparer comme Banerjee *et al.* [17] les partitions obtenues avec les lois gaussienne, binomiale, et la loi de Poisson, en utilisant les divergences de Bregman correspondantes : distance euclidienne, perte logistique, et distance de Kullback-Leibler généralisée ou I-divergence. Pour chacune des lois, 90 observations ont été simulées, réparties en 3 ensembles de 30 données centrées en 10, 20 et 40 respectivement. En prenant la variance égale à 25 pour la loi gaussienne, et le nombre d'épreuves égal à 100 pour la loi binomiale, les trois modèles obtenus ont une variance similaire.

Remarque 1.6.2. Une réalisation gaussienne peut être négative (en particulier si la moyenne est 10), alors que la perte logistique et la distance de Kullback-Leibler généralisée ne sont pas définies pour de tels éléments. La solution utilisée pour contourner ce problème consiste à n'accepter que les vecteurs formés de réalisations positives. En réalité, la loi considérée est donc une loi gaussienne conditionnelle.

Le Tableau 1.4 présente pour chaque association possible Loi – Divergence de Bregman la moyenne de l'information mutuelle normalisée sur 100 essais, tandis que le Tableau 1.5 donne le nombre de fois où une divergence a conduit pour une loi à la meilleure partition « au sens large » (la somme des nombres d'une ligne du tableau est supérieure à 100).

	Euclidien	Logistique	I-divergence
Gaussienne	**0.689**	0.685	0.672
Binomiale	0.791	**0.813**	0.806
Poisson	0.702	0.728	**0.732**

TABLEAU 1.4.: Information mutuelle normalisée (100 essais).

	Euclidien	Logistique	I-divergence
Gaussienne	**52**	42	35
Binomiale	38	**62**	57
Poisson	37	56	**63**

TABLEAU 1.5.: Nombre de cas (sur 100) où chaque divergence donne le meilleur résultat (au sens large).

Dans cet exemple, nous retrouvons le fait que la distance euclidienne est la plus adaptée pour des données de loi gaussienne, la perte logistique pour la loi binomiale et la distance de Kullback-Leibler pour la loi de Poisson.

Un exemple visuel : bande et cercle

Nous cherchons ici à classer 100 observations obtenues comme suit : 50 sont distribuées uniformément sur un cercle de centre $(3, 10)$ et de rayon 1 et les 50 autres sur la bande située entre les droites d'équation $x = 0$ et $x = 1$ (rectangle de hauteur 20 et de largeur 1). Le résultat du clustering pour différentes divergences de Bregman est visible dans la Figure 1.7 et l'information mutuelle normalisée moyenne sur 50 essais est donnée dans le Tableau 1.6.

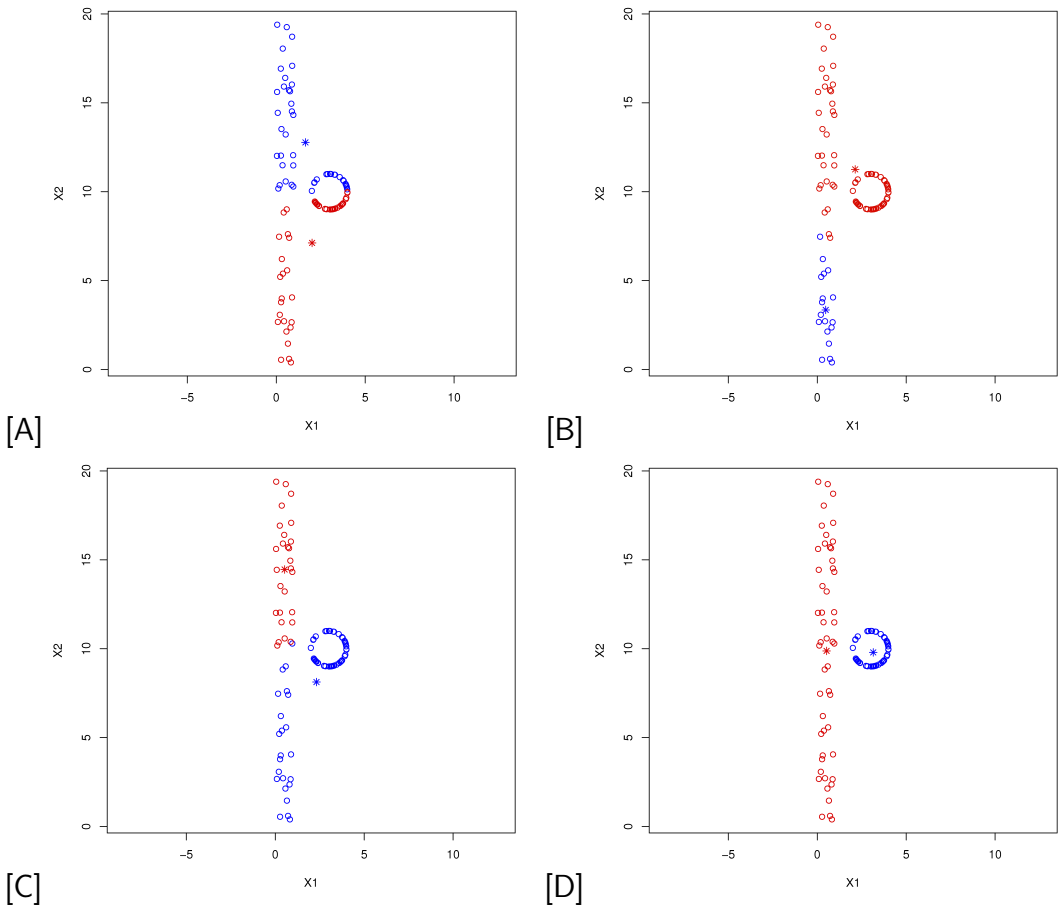


FIGURE 1.7.: Clustering de données distribuées uniformément sur une bande ou un cercle (k=2, n=100). [A] Distance euclidienne au carré. [B] Distance de Kullback-Leibler généralisée. [C] Perte logistique. [D] Distance de Itakura-Saito.

Euclidien	I-divergence	Logistique	Itakura-Saito
0.302	0.327	0.320	**0.986**

TABLEAU 1.6.: Information mutuelle normalisée (50 essais).

Nous constatons que la distance de Itakura-Saito, non symétrique et non convexe en la deuxième variable, sépare la bande et le cercle, alors que les autres divergences de Bregman coupent les données différemment.

Données sur le simplexe

Nous avons simulé 45 observations distribuées sur le simplexe de dimension 2 suivant une loi de Dirichlet. Pour mémoire, une loi de Dirichlet de paramètres (a_1, a_2, a_3), où $a_i > 0$ pour tout $i = 1, \ldots, 3$, est définie par

$$\mathbb{P}\left\{P_1 = p_1, P_2 = p_2, P_3 = p_3\right\} = \frac{\Gamma\left(\sum_{\ell=1}^{3} a_\ell\right)}{\prod_{\ell=1}^{3} \Gamma(a_\ell)} \prod_{\ell=1}^{3} p_\ell^{a_\ell - 1},$$

où $p_i > 0$ pour tout $i = 1, \ldots, 3$ et $p_1 + p_2 + p_3 = 1$ (proportions). Rappelons que la fonction Γ est donnée, pour $x > 0$, par $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} \mathrm{d}t$. Ici, 3 groupes de 15 données suivant respectivement la loi de Dirichlet de paramètres $(10, 10, 2)$, $(5, 5, 5)$ et $(2, 2, 10)$ ont été construits. Le Tableau 1.7 indique que la distance de Kullback-Leibler est la plus appropriée pour retrouver les groupes que forment ces données appartenant au simplexe. Ce résultat est cohérent avec l'utilisation habituelle de cette divergence en classification de documents (Banerjee *et al.* [17]). En effet, ces observations suivant une loi de Dirichlet peuvent s'interpréter comme un cas simple de classification de textes basé sur 3 mots ou expressions. Chacun des paramètres de la loi est lié à la fréquence moyenne de l'un de ces mots.

Euclidien	Kullback-Leibler	Logistique	Itakura-Saito
0.674	**0.714**	0.689	0.673

TABLEAU 1.7.: Information mutuelle normalisée (100 essais).

Distance de Mahalanobis/Norme euclidienne

Lorsque nous effectuons le clustering des observations représentées dans la Figure 1.8, la distance de Mahalanobis avec la matrice

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 8 \end{pmatrix}^{-1}$$

et la distance euclidienne au carré produisent des groupes très différents.

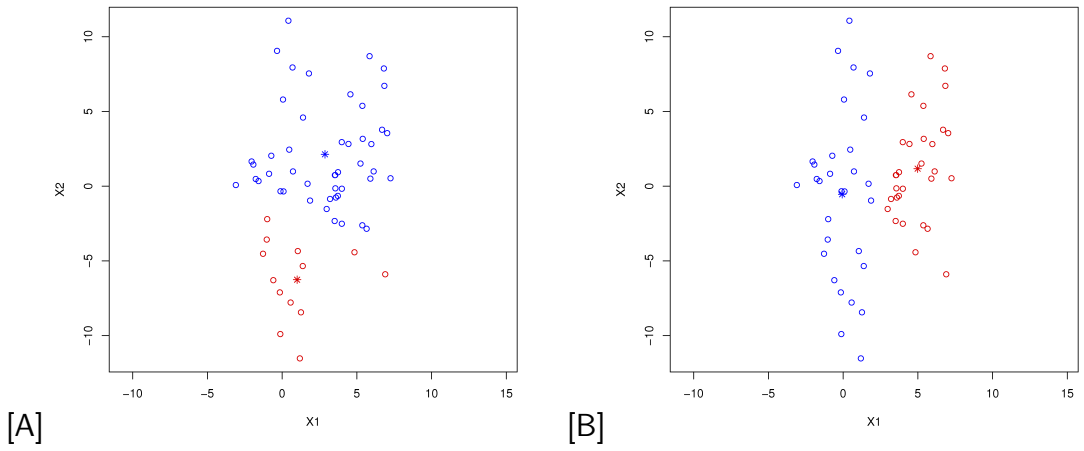


FIGURE 1.8.: Clustering d'observations gaussiennes (k=2, n=60). [A] Carré de la distance euclidienne. [B] Distance de Mahalanobis.

En fait, les deux ellipses allongées sont formées de réalisations de deux vecteurs gaussiens de matrice de covariance

$$\begin{pmatrix} 2 & 1 \\ 1 & 8 \end{pmatrix}.$$

On retrouve le fait que la distance de Mahalanobis adaptée à ces données est celle prenant comme paramètre l'inverse de leur matrice de covariance. Le résultat de la simulation sur le plan de l'information mutuelle normalisée est donné par le Tableau 1.8.

Euclidien	Mahalanobis
0.445	**0.793**

TABLEAU 1.8.: Information mutuelle normalisée (100 essais).

Notons qu'il existe des méthodes permettant d'estimer la matrice de covariance à partir des données de façon à choisir la distance de Mahalanobis la plus appropriée (Art, Gnanadesikan, et Kettenring [12], Tarsitano [179]).

1.6.2. Simulations en dimension infinie

Donnons à présent quelques illustrations utilisant des divergences de Bregman de dimension infinie.

Courbes gaussiennes

Dans les deux exemples suivants, nous nous intéressons à un ensemble d'observations constitué de 40 courbes de densité gaussienne.

Dans le premier cas, il s'agit plus précisément de 2 groupes de 20 courbes gaussiennes centrées en 22.5 et 24.5 respectivement, avec un écart-type choisi uniformément entre 2 et 5. Les résultats sont présentés dans la Figure 1.9 et le Tableau 1.9. Il s'avère que la distance de Kullback-Leibler généralisée est celle qui sépare le mieux les 2 groupes.

Remarque 1.6.3. Pour éviter un problème lorsque les courbes s'approchent de 0, nous n'avons pas intégré sur toute la droite réelle, mais seulement sur $[15, 32]$. En conséquence, comme les fonctions considérées ne sont pas d'intégrale 1 sur cet intervalle, la distance de Kullback-Leibler pour les densités de probabilité a été remplacée par la distance de Kullback-Leibler généralisée.

Dans le deuxième exemple, nous disposons de 20 courbes gaussiennes d'écart-type 4 et 20 d'écart-type 5, avec une moyenne choisie uniformément entre 21 et 26. Cette fois-ci, retrouver les 2 groupes signifie classer les courbes selon la variance de la gaussienne, alors que dans ce qui précède la propriété caractéristique d'un groupe était la moyenne. Comme le montrent la Figure 1.10 et le Tableau 1.10, la conclusion est inversée, et la divergence qui conduit au meilleur résultat est le biais quadratique.

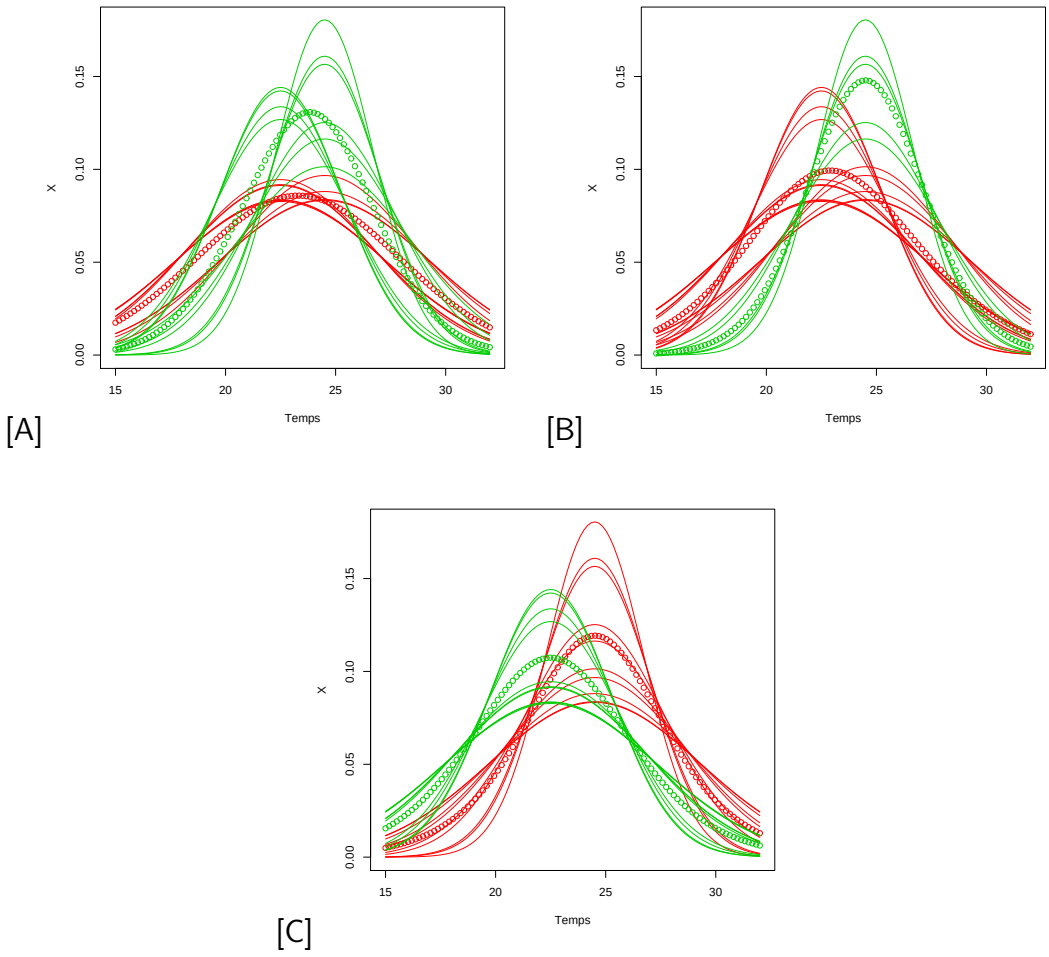


FIGURE 1.9.: Clustering de courbes de densité gaussienne (k=2,n=40). [A] Biais quadratique. [B] Carré de la distance L^2. [C] Distance de Kullback-Leibler généralisée.

I-divergence	Norme 2	Biais quadratique
0.910	0.799	0.017

TABLEAU 1.9.: Information mutuelle normalisée (30 essais).

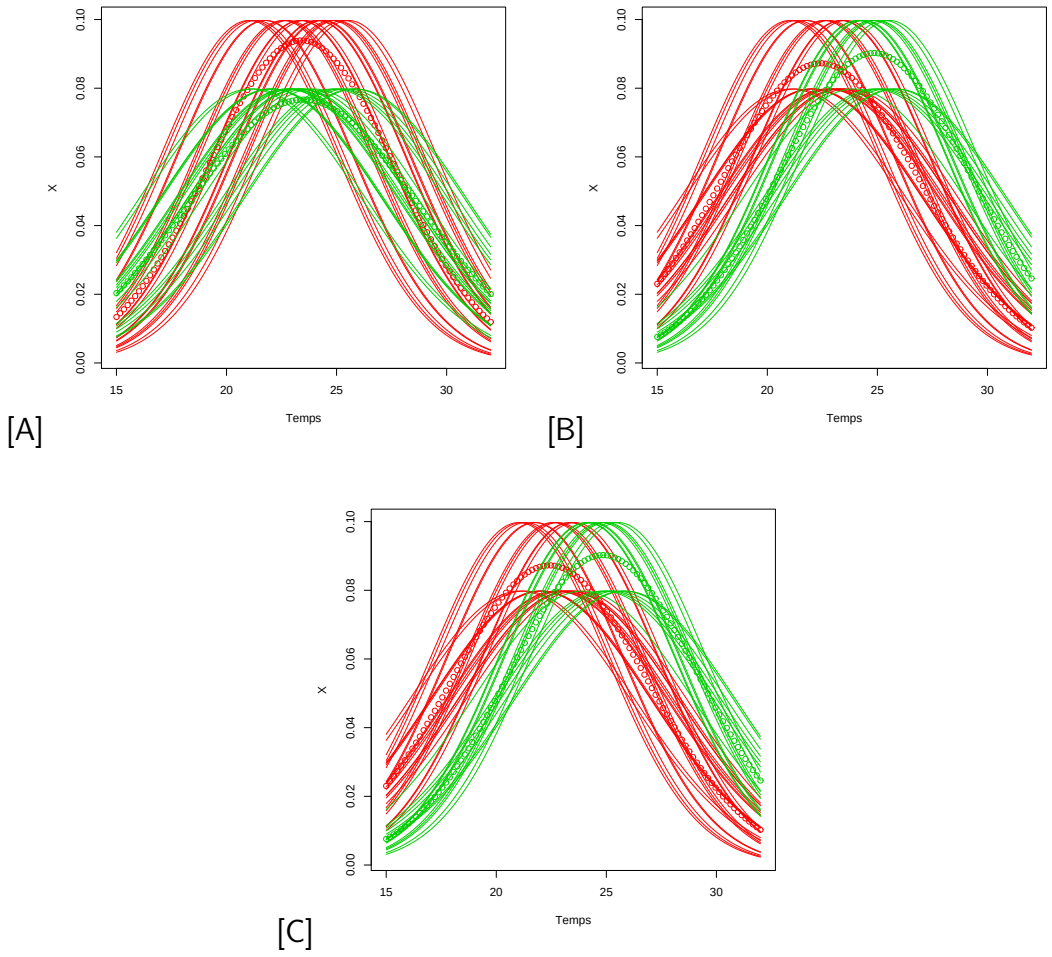


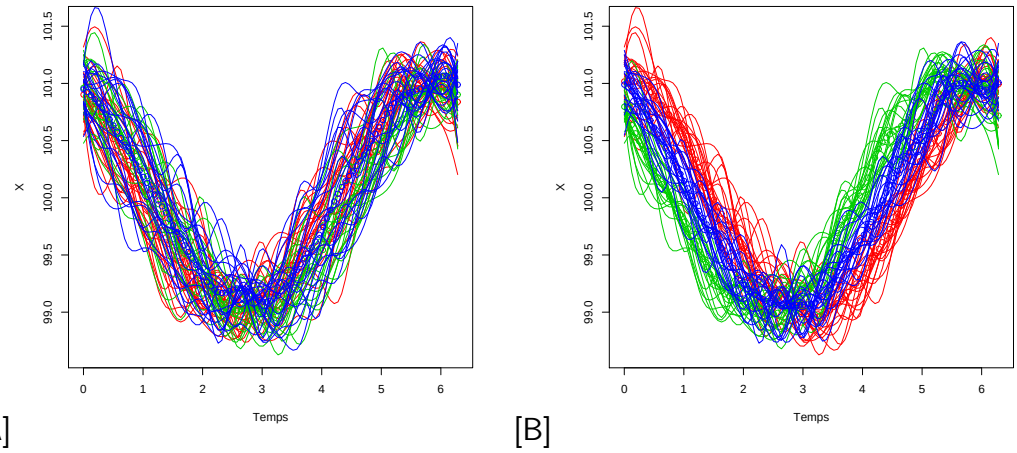
FIGURE 1.10.: Clustering de courbes de densité gaussienne (k=2,n=40). [A] Biais quadratique. [B] Carré de la distance L^2. [C] Distance de Kullback-Leibler généralisée.

I-divergence	Norme 2	Biais quadratique
0.017	0.022	**1**

TABLEAU 1.10.: Information mutuelle normalisée (30 essais).

Sinusoïdes bruitées

Dans ce dernier exemple, nous considérons des observations formant 3 groupes de sinusoïdes bruitées, correspondant à 3 phases différentes. La variance du bruit gaussien est 0.1. D'après la Figure 1.11 (phases 0, $\pi/8$, $\pi/4$) et le Tableau 1.11 (phases 0, $\pi/24$, $\pi/12$), la distance L^2 au carré et la distance de Itakura-Saito semblent toutes deux appropriées pour distinguer les sinusoïdes selon leur phase.



[A] [B]

FIGURE 1.11.: Clustering de sinusoïdes bruitées (k=3,n=45). [A] Biais quadratique. [B] Carré de la distance L^2.

Norme 2	Biais quadratique	Itakura-Saito
0.858	0.043	**0.853**

TABLEAU 1.11.: Information mutuelle normalisée (30 essais).

1.7. Preuves de deux lemmes

1.7.1. Preuve du Lemme 1.4.3

La preuve de ce lemme est proche de la première partie de la démonstration du Théorème 1 de Linder [131] (voir aussi Pollard [157]).

Remarquons tout d'abord que, par la formule de Taylor, il existe z appartenant

au segment ouvert xy tel que

$$\phi(x) = \phi(y) + D_y\phi(x-y) + \frac{1}{2}D_z^2\phi(x-y, x-y).$$

Donc,

$$d_\phi(x,y) = \frac{1}{2}D_z^2\phi(x-y, x-y),$$

ce qui entraîne, par hypothèse,

$$\frac{m}{2}\|x-y\|^2 \leq d_\phi(x,y) \leq \frac{M}{2}\|x-y\|^2.$$

Pour $\ell \in \mathbb{N}^\star$ et $\mathbf{c}^\ell = (c_1, \ldots, c_\ell)$, notons

$$w_\ell(\mathbf{c}^\ell) = \mathbb{E}\Big[\min_{j=1,\ldots,\ell} d_\phi(x, c_j)\Big].$$

Rappelons que $W_\ell^\star(\mu)$ désigne la distorsion optimale pour les ℓ-quantificateurs. Comme le Lemme 1.4.2 assure l'existence d'un quantificateur optimal lorsque $k = 1$, nous considérons le cas $k \geq 2$. Nous pouvons de plus supposer que le support de μ contient au moins k points (sinon, s'intéresser à un k-quantificateur n'a pas beaucoup de sens), de sorte que $W_k^\star(\mu) < W_{k-1}^\star(\mu)$. Soit $\varepsilon > 0$ tel que

$$\varepsilon < \frac{1}{2}(W_{k-1}^\star(\mu) - W_k^\star(\mu)) \tag{1.12}$$

et soient $0 < r_1 < r_2$ tels que

$$\frac{m}{2}(r_2 - r_1)^2\mu(B_{r_1}) > W_k^\star(\mu) + \varepsilon \tag{1.13}$$

et

$$2M\int_{B_{2r_2}^c} \|x\|^2 \mathrm{d}\mu(x) < \varepsilon. \tag{1.14}$$

Choisissons une table de codage $\mathbf{c}^k = (c_1, \ldots, c_k)$ telle que

$$w_k(\mathbf{c}^k) < W_k^\star(\mu) + \varepsilon.$$

Ainsi,

$$w_k(\mathbf{c}^k) < W_{k-1}^\star(\mu) - \varepsilon,$$

ce qui implique que les $c_1, \ldots, c_k$ sont distincts. Supposons que ces éléments sont rangés par ordre croissant, c'est-à-dire $\|c_1\| \leq \cdots \leq \|c_k\|$. Alors, $\|c_1\| \leq r_2$. En effet, si $\|c_1\| > r_2$, alors $\|c_j\| > r_2$ pour tout j. Donc, pour $x \in B_{r_1}$,

$$\begin{aligned}
\min_{j=1,\ldots,k} d_\phi(x, c_j) &\geq \frac{m}{2}\min_{j=1,\ldots,k}\|x - c_j\|^2 \\
&\geq \frac{m}{2}\min_{j=1,\ldots,k}(\|c_j\| - \|x\|)^2 \\
&\geq \frac{m}{2}(r_2 - r_1)^2.
\end{aligned}$$

Ainsi, $W_k^\star(\mu) + \varepsilon > \frac{m}{2}(r_2 - r_1)^2 \mu(B_{r_1})$, ce qui est exclu d'après l'inégalité (1.13). Montrons à présent que, pour tout j, $\|c_j\| \leq C r_2$ où $C = 2 + 3\sqrt{\frac{M}{m}} > 0$. Supposons que $\|c_k\| > C r_2$. Pour $x \in B_{2r_2}$, on a

$$d_\phi(x, c_1) \leq \frac{M}{2}(\|x\| + \|c_1\|)^2 \leq \frac{9}{2} M r_2^2$$

et

$$d_\phi(x, c_k) \geq \frac{m}{2}(\|c_k\| - \|x\|)^2 > \frac{m}{2}(C r_2 - 2r_2)^2 = \frac{9}{2} M r_2^2,$$

donc

$$d_\phi(x, c_1) \leq d_\phi(x, c_k).$$

Pour $x \in B_{2r_2}^c$,

$$d_\phi(x, c_1) \leq \frac{M}{2}(\|x\| + \|c_1\|)^2 \leq 2M\|x\|^2.$$

Donc

$$d_\phi(x, c_1) \leq d_\phi(x, c_k) + 2M\|x\|^2 \mathbf{1}_{\{x \in B_{2r_2}^c\}}. \tag{1.15}$$

Notons $\mathbf{c}^{k-1} = (c_1, \ldots, c_{k-1})$ et $S_1, \ldots, S_k$ les cellules de Voronoi associées aux composantes de $\mathbf{c}^k$. Nous obtenons

$$\begin{aligned} w_{k-1}(\mathbf{c}^{k-1}) &= \sum_{j=1}^{k} \int_{S_j} \min_{j=1,\ldots,k-1} d_\phi(x, c_j) \mathrm{d}\mu(x) \\ &\leq \sum_{j=1}^{k-1} \int_{S_j} d_\phi(x, c_j) \mathrm{d}\mu(x) + \int_{S_k} d_\phi(x, c_1) \mathrm{d}\mu(x) \\ &\leq \sum_{j=1}^{k} \int_{S_j} d_\phi(x, c_j) \mathrm{d}\mu(x) + 2M \int_{B_{2r2}^c} \|x\|^2 \mathrm{d}\mu(x). \end{aligned}$$

Ce dernier point provient de l'inégalité (1.15). Alors, par les inégalités (1.14) et (1.12),

$$\begin{aligned} w_{k-1}(\mathbf{c}^{k-1}) &\leq w_k(\mathbf{c}^k) + \varepsilon \\ &\leq W_k^\star(\mu) + 2\varepsilon \\ &< W_{k-1}^\star(\mu), \end{aligned}$$

ce qui est absurde d'après la définition de $W_{k-1}^\star(\mu)$ comme borne inférieure. Ainsi $w_k(\mathbf{c}^k) < W_k^\star(\mu) + \varepsilon$ implique $(c_1, \ldots, c_k) \in (B_{Cr_2})^k$ et, en posant $R = C r_2$,

$$W_k^\star(\mu) = \inf_{\mathbf{c}^k \in B_R^k} w_k(\mathbf{c}^k).$$

1.7.2. Preuve du Lemme 1.5.1

La preuve du lemme utilise les propriétés des moyennes de Rademacher rappelées dans l'Annexe A.

(i) Soient $X'_1, \ldots, X'_n$ des copies indépendantes de $X_1, \ldots, X_n$, indépendantes des variables de Rademacher indépendantes $\varepsilon_1, \ldots, \varepsilon_n$. La preuve de l'inégalité repose sur un argument de symétrisation dû à Giné et Zinn [96]. On a

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\left(\overline{W}(\mu_n,\mathbf{c})-\overline{W}(\mu,\mathbf{c})\right)\right]\\
&\quad=\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\left(\frac{1}{n}\sum_{i=1}^n\min_{j=1,\ldots,k}\ell_{c_j}(X_i)-\mathbb{E}\left[\min_{j=1,\ldots,k}\ell_{c_j}(X)\right]\right)\right]\\
&\quad=\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\left(\frac{1}{n}\sum_{i=1}^n\min_{j=1,\ldots,k}\ell_{c_j}(X_i)-\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\min_{j=1,\ldots,k}\ell_{c_j}(X'_i)\Big|X_1,\ldots,X_n\right]\right)\right].
\end{aligned}$$

Par l'inégalité de Jensen, et en utilisant le fait que les $X_1, \ldots, X_n, X'_1, \ldots, X'_n$ sont indépendants et de même loi,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\left(\overline{W}(\mu_n,\mathbf{c})-\overline{W}(\mu,\mathbf{c})\right)\right]\\
&\quad\le\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\left(\min_{j=1,\ldots,k}\ell_{c_j}(X_i)-\min_{j=1,\ldots,k}\ell_{c_j}(X'_i)\right)\right]\\
&\quad=\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\left(\min_{j=1,\ldots,k}\ell_{c_j}(X_i)-\min_{j=1,\ldots,k}\ell_{c_j}(X'_i)\right)\right]\\
&\quad\le\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\min_{j=1,\ldots,k}\ell_{c_j}(X_i)\right]+\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n(-\varepsilon_i)\min_{j=1,\ldots,k}\ell_{c_j}(X'_i)\right]\\
&\quad=2\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\min_{j=1,\ldots,k}\ell_{c_j}(X_i)\right].
\end{aligned}$$

(ii) Pour obtenir la majoration (ii), on raisonne par récurrence sur k. Pour $k = 1$,

on a

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\ell_c(X_i)\Big]\\
&\quad=\mathbb{E}\Big[\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_i(-\phi(c)-D_c\phi(X_i-c))\Big]\\
&\quad\le\mathbb{E}\Big[\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_iD_c\phi(X_i)\Big]+\mathbb{E}\Big[\sup_{c\in\mathcal{C}_R}\frac{|-\phi(c)+D_c\phi(c)|}{n}\Big|\sum_{i=1}^n\varepsilon_i\Big|\Big]\\
&\quad\le\mathbb{E}\Big[\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_iD_c\phi(X_i)\Big]+\frac{1}{n}\sup_{c\in\mathcal{C}_R}|-\phi(c)+D_c\phi(c)|\left(\mathbb{E}\Big[\sum_{i=1}^n\varepsilon_i\Big]^2\right)^{1/2}\\
&\quad=\mathbb{E}\Big[\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_iD_c\phi(X_i)\Big]+\frac{1}{\sqrt{n}}\sup_{c\in\mathcal{C}_R}|-\phi(c)+D_c\phi(c)|,
\end{aligned}$$

en utilisant l'indépendance des ε_i. Supposons l'énoncé (ii) vrai au rang $k-1$, et montrons qu'alors, il l'est encore au rang k. Soit $\mathbf{c}^{k-1}=(c_1,\dots,c_{k-1})$.

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\min_{j=1,\dots,k}\ell_{c_j}(X_i)\Big]\\
&\quad=\mathbb{E}\Big[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\min\Big(\ell_{c_k}(X_i),\min_{j=1,\dots,k-1}\ell_{c_j}(X_i)\Big)\Big]\\
&\quad=\mathbb{E}\Big[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{2n}\sum_{i=1}^n\varepsilon_i\Big(\ell_{c_k}(X_i)+\min_{j=1,\dots,k-1}\ell_{c_j}(X_i)-|\ell_{c_k}(X_i)-\min_{j=1,\dots,k-1}\ell_{c_j}(X_i)|\Big)\Big],
\end{aligned}$$

car $\min(a,b)=(a+b)/2-|a-b|/2$. Par les propriétés des moyennes de Rademacher,

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\min_{j=1,\dots,k}\ell_{c_j}(X_i)\Big]\\
&\quad\le\mathbb{E}\Big[\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\ell_c(X_i)\Big]+\mathbb{E}\Big[\sup_{\mathbf{c}^{k-1}\in(\mathcal{C}_R)^{k-1}}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\min_{j=1,\dots,k-1}\ell_{c_j}(X_i)\Big]\\
&\quad=\mathbb{E}\Big[\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_iD_c\phi(X_i)\Big]+\frac{1}{\sqrt{n}}\sup_{c\in\mathcal{C}_R}|-\phi(c)+D_c\phi(c)|\\
&\qquad+(k-1)\left(\mathbb{E}\Big[\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_iD_c\phi(X_i)\Big]+\frac{1}{\sqrt{n}}\sup_{c\in\mathcal{C}_R}|-\phi(c)+D_c\phi(c)|\right)\\
&\quad=k\left(\mathbb{E}\Big[\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_iD_c\phi(X_i)\Big]+\frac{1}{\sqrt{n}}\sup_{c\in\mathcal{C}_R}|-\phi(c)+D_c\phi(c)|\right),
\end{aligned}$$

qui est la majoration attendue au rang k.

(iii) On a

$$\begin{aligned}\mathbb{E}\Big[\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_i D_c\phi(X_i)\Big] &= \mathbb{E}\Big[\sup_{c\in\mathcal{C}_R}\frac{1}{n}D_c\phi\Big(\sum_{i=1}^n\varepsilon_i X_i\Big)\Big]\\ &\le \frac{1}{n}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|\mathbb{E}\Big\|\sum_{i=1}^n\varepsilon_i X_i\Big\|\\ &\le \frac{1}{n}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|\Big(\mathbb{E}\Big\|\sum_{i=1}^n\varepsilon_i X_i\Big\|^2\Big)^{1/2}\end{aligned}$$

Comme E est de type 2, et puisque les X_i ont même loi,

$$\begin{aligned}\mathbb{E}\Big[\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_i D_c\phi(X_i)\Big] &\le \frac{1}{n}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|T_2\Big[\sum_{i=1}^n\mathbb{E}\|X_i\|^2\Big]^{1/2}\\ &= \frac{1}{n}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|T_2(n\mathbb{E}\|X\|^2)^{1/2}\\ &= \frac{T_2}{\sqrt{n}}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|(\mathbb{E}\|X\|^2)^{1/2}.\end{aligned}$$

1.8. Annexe

1.8.1. Variables aléatoires dans un espace de Banach

Cette section constitue un rappel sur la manière de définir l'espérance d'une variable aléatoire à valeurs dans un espace de Banach. Pour un exposé plus détaillé sur ce sujet, le lecteur pourra se reporter à Bharucha-Reid [31, Chapitre 1].

Soit $(\Omega,\mathcal{A},P)$ un espace de probabilités complet et $(E,\mathcal{B})$ un espace mesurable, où E est un espace de Banach et $\mathcal{B}$ la tribu des boréliens de E.

Définition 1.8.1 (Variable aléatoire). *1. Une application $X:\Omega\to E$ est appelée variable aléatoire à valeurs dans E si pour tout $B\in\mathcal{B}$, $X^{-1}(B)\in\mathcal{A}$.*

2. Une application $X:\Omega\to E$ est appelée variable aléatoire forte (ou de Bochner) s'il existe une suite de variables aléatoires étagées convergeant presque sûrement vers X.

3. Une application $X:\Omega\to E$ est appelée variable aléatoire faible (ou de Pettis) si pour toute forme linéaire continue $L\in E'$, $L(X)$ est une variable aléatoire réelle.

Si E est séparable, ces différentes définitions de variable aléatoire coïncident.

Définition 1.8.2 (Intégrale de Pettis)**.** *Une variable aléatoire X est dite intégrable au sens de Pettis si, pour tout $A \in \mathcal{A}$, il existe un élément $m_A \in E$ tel que, pour tout $L \in E'$,*

$$L(m_A) = \int_A L(X)\mathrm{d}P.$$

Pour tout $A \in \mathcal{A}$, l'intégrale de Pettis est alors définie par

$$\int_A^{(\mathrm{P})} X\mathrm{d}P = m_A.$$

L'espérance de X au sens de Pettis est donnée par

$$\mathbb{E}^{(\mathrm{P})}[X] = \int_\Omega^{(\mathrm{P})} X\mathrm{d}P.$$

Afin de définir l'intégrabilité au sens de Bochner, il faut tout d'abord en donner la caractérisation dans le cas des variables aléatoires étagées.

Une variable aléatoire étagée $X = \sum_{j=1}^m x_j \mathbf{1}_{A_j}$ est dite intégrable au sens de Bochner si $\|X\|$ est intégrable. Alors, pour tout $A \in \mathcal{A}$, $\int_A^{(\mathrm{B})} X\mathrm{d}P = \sum_{j=1}^m x_j P(A \cap A_j)$.

Définition 1.8.3 (Intégrale de Bochner)**.** *Une variable aléatoire X est dite intégrable au sens de Bochner s'il existe une suite $(X_n)_{n\geq 1}$ de variables aléatoires étagées convergeant presque sûrement vers X et vérifiant*

$$\lim_{n\to+\infty} \int_\Omega \|X_n - X\| dP = 0.$$

Pour tout $A \in \mathcal{A}$, l'intégrale de Bochner est alors définie par

$$\int_A^{(\mathrm{B})} X\mathrm{d}P = \lim_{n\to+\infty} \int_A^{(\mathrm{B})} X_n\mathrm{d}P.$$

L'espérance de X au sens de Bochner est donnée par

$$\mathbb{E}^{(\mathrm{B})}[X] = \int_\Omega^{(\mathrm{B})} X\mathrm{d}P.$$

Deux résultats intéressants

- Si une variable aléatoire est intégrable au sens de Bochner, elle est également intégrable au sens de Pettis, et les deux intégrales ont même valeur.
- De plus, si E est séparable, l'espérance $\mathbb{E}^{(\mathrm{B})}[X]$ existe si et seulement si $\mathbb{E}\|X\| < +\infty$.

1.8.2. Quelques rappels de calcul différentiel

Nous rassemblons ici quelques éléments utiles de calcul différentiel (voir, par exemple, Cartan [48]).

Soient $(E, \|\cdot\|_E)$ et $(F, \|\cdot\|_F)$ des espaces de Banach et U un ouvert non vide de E. Notons $\mathcal{L}(E, F)$ l'ensemble des applications linéaires continues de E dans F.

Proposition 1.8.1. *Muni de la norme d'opérateur $\|\cdot\|_{\mathcal{L}(E,F)}$ définie par*

$$\|L\|_{\mathcal{L}(E,F)} = \sup_{x \in E \setminus \{0\}} \frac{\|L(x)\|_F}{\|x\|_E},$$

l'ensemble $\mathcal{L}(E, F)$ est un espace de Banach.

Pour éviter d'alourdir les notations, nous omettrons désormais l'indice dans l'écriture des normes.

Définition 1.8.4 (Application différentiable)**.** *Une application $f : U \to F$ est dite différentiable (au sens de Fréchet) au point a de U s'il existe une application linéaire continue $L : E \to F$ telle que,*

$$\|f(a+h) - f(a) - L(h)\| = o(\|h\|).$$

Si elle existe, l'application L est unique. Elle est appelée différentielle de f au point a et notée $D_a f$. On dit que f est différentiable dans U si elle est différentiable en tout point de U.

Notons que l'écriture $f(a+h)$ sous-entend que h est pris suffisamment petit pour que $a+h$ appartienne encore à U. C'est pour cela que U est supposé ouvert.

Soit Df l'application

$$\begin{aligned} U &\to \mathcal{L}(E, F) \\ a &\mapsto D_a f. \end{aligned}$$

Définition 1.8.5 (Application de classe C^1)**.** *On dit que $f : U \to F$ est de classe C^1 si f est différentiable dans U et si l'application $Df : U \to \mathcal{L}(E, F)$ est continue.*

Lorsqu'une application f est différentiable sur U, on peut se demander si Df est elle-même différentiable.

Définition 1.8.6 (Différentielle d'ordre 2)**.** *On dit que f est deux fois différentiable au point $a \in U$ si l'application Df est différentiable en a. On note alors $D_a^2(f)$ la différentielle en a de Df. On dit que f est deux fois différentiable dans U si elle est deux fois différentiable en tout point de U.*

Soit $D^2 f$ l'application

$$U \to \mathcal{L}(E, \mathcal{L}(E, F))$$
$$a \mapsto D_a^2 f.$$

Définition 1.8.7 (Application de classe C^2)**.** *On dit que $f : U \to F$ est de classe C^2 si f est deux fois différentiable dans U et si l'application $D^2 f$ est continue.*

Dans un espace de Hilbert, le Théorème de Représentation de Riesz permet d'introduire la notion de gradient.

Théorème 1.8.1 (Représentation de Riesz)**.** *Soient $(E, \langle \cdot, \cdot, \rangle)$ un espace de Hilbert et L une forme linéaire continue sur E, c'est-à-dire $L \in \mathcal{L}(E, \mathbb{R})$. Alors, il existe un unique $v \in E$ tel que pour tout $h \in E$, $L(h) = \langle h, v \rangle$.*

Définition 1.8.8 (Gradient)**.** *Soient $(E, \langle \cdot, \cdot, \rangle)$ un espace de Hilbert et $f : E \to \mathbb{R}$ une application différentiable en $a \in E$. On appelle gradient de f en a et on note $\nabla f(a)$ l'unique vecteur de E tel que pour tout $h \in E$,*

$$D_a f(h) = \langle h, \nabla f(a) \rangle.$$

1.8.3. Des résultats utiles de topologie

Compacité

Nous rappelons quelques résultats concernant la notion de compacité. Pour davantage de précisions, on pourra se reporter à Hirsch et Lacombe [108]. Enonçons tout d'abord le Théorème de Riesz, qui exprime le « défaut de compacité » en dimension infinie.

Théorème 1.8.2 (Riesz)**.** *Soit E un espace vectoriel normé. Les propriétés suivantes sont équivalentes :*

1. *E est de dimension finie.*
2. *E est localement compact.*
3. *La boule unité fermée de E est compacte.*

Théorème de Tychonoff

Soit $(E_\ell, d_\ell)_{\ell \in \mathbb{N}}$ une suite d'espaces métriques et soit $E = \Pi_{\ell \in \mathbb{N}} E_\ell$. La fonction $d : E \times E \to \mathbb{R}^+$ définie par

$$d(x, y) = \sum_{\ell=0}^{+\infty} 2^{-\ell} \min(d_\ell(x_\ell, y_\ell), 1)$$

est une distance sur E appelée distance produit. Ainsi, (E, d) est un espace métrique. Une suite $(x^n)_{n\in\mathbb{N}}$ d'éléments de E converge vers $x \in E$ pour la distance d si, et seulement si, pour tout $\ell \in \mathbb{N}$, la suite $(x^n_\ell)_{n\in\mathbb{N}}$ d'éléments de E_ℓ converge vers $x_\ell \in E_\ell$.

Le Théorème de Tychonoff assure qu'un produit d'espaces compacts est compact.

Théorème 1.8.3 (Tychonoff)**.** *Soit $(E_\ell)_{\ell\in\mathbb{N}}$ une suite d'espaces métriques compacts, et soit $E = \Pi_{\ell\in\mathbb{N}} E_\ell$ muni de la distance produit. Alors E est compact.*

Compactifié d'Alexandroff

Soient (E, d) un espace métrique localement compact séparable et $\tilde{E} = E \cup \{\omega\}$, où ω est un point n'appartenant pas à E, appelé point à l'infini. Muni de la topologie formée des ouverts de E et des ensembles de la forme $\{\omega\} \cup K^c$ avec K compact de E, $\tilde{E}$ est alors compact. Cet ensemble est appelé compactifié d'Alexandroff de E.

Exemple 1.8.1.
1. Le compactifié d'Alexandroff de $\mathbb{R}$ est un cercle.
2. Le compactifié d'Alexandroff du plan $\mathbb{R}^2$ est la sphère de Riemann $\mathbb{S}^2$.

Remarque 1.8.1. La compactification d'Alexandroff ne s'applique pas au cadre général des espaces de Banach, puisque le Théorème de Riesz (Théorème 1.8.2) indique qu'aucun espace vectoriel normé de dimension infinie n'est localement compact.

Convergence continue sur un compact

La proposition suivante, dont la preuve est par exemple présentée dans Lojasiewicz [134, Théorème 3.1.9], établit l'équivalence de deux modes de convergence de suites de fonctions définies sur un compact.

Proposition 1.8.2. *Soit $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions à valeurs réelles définies sur un compact K. Les assertions suivantes sont équivalentes :*

1. *Pour toute suite $(x_n)_{n\in\mathbb{N}}$ d'éléments de K qui converge vers x, $(f_n(x_n))_{n\in\mathbb{N}}$ converge vers $f(x)$.*
2. *La suite de fonctions $(f_n)_{n\in\mathbb{N}}$ converge uniformément vers f sur K.*

Fonctions semi-continues inférieurement et fonctions convexes

Les définitions et résultats suivants concernent deux familles de fonctions particulières, les fonctions semi-continues et les fonctions convexes (voir Brezis [41]

et Rockafellar [161] pour les démonstrations). Notons que ces fonctions peuvent prendre la valeur $+\infty$. Néanmoins, nous ne considérons que des fonctions *propres*, c'est-à-dire non identiquement égales à $+\infty$.

Définition 1.8.9 (Epigraphe)**.** *Soit f une fonction de E dans $]-\infty,+\infty]$. L'épigraphe de f est l'ensemble $\{(x,\lambda) \in E \times \mathbb{R}, f(x) \leq \lambda\}$.*

Définition 1.8.10 (Fonction semi-continue inférieurement)**.** *Une fonction $f : E \to]-\infty,+\infty]$ est dite semi-continue inférieurement en x_0 si*

$$\liminf_{x \to x_0} f(x) \geq f(x_0).$$

La fonction f est semi-continue inférieurement sur E si elle l'est en tout point de E.

Proposition 1.8.3. *Les assertions suivantes sont équivalentes :*

1. *La fonction $f : E \to]-\infty,+\infty]$ est semi-continue inférieurement.*
2. *Pour tout $\lambda \in \mathbb{R}$, l'ensemble $\{x \in E, f(x) \leq \lambda\}$ est fermé.*
3. *L'épigraphe de f est fermé.*

Définition 1.8.11 (Fonction convexe)**.** *Une fonction $f : E \to]-\infty,+\infty]$ est dite convexe si pour tous $x, y \in E$ et pour tout $\lambda \in [0,1]$, on a*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)y).$$

Proposition 1.8.4. *Une fonction $f : E \to]-\infty,+\infty]$ est convexe si, et seulement si, son épigraphe est convexe.*

Les fonctions de Legendre forment une classe de fonctions convexes possédant de bonnes propriétés.

Définition 1.8.12 (Fonction de Legendre)**.** *Une paire $(\mathcal{C}, f)$ est dite de Legendre si $\mathcal{C}$ est un convexe ouvert non vide de $\mathbb{R}^d$ et si f est une fonction strictement convexe et différentiable sur $\mathcal{C}$ telle que $|\nabla f(x_n)|$ tende vers l'infini lorsque $(x_n)_{n\in\mathbb{N}}$ tend vers $x \in \partial\mathcal{C}$.*

Définition 1.8.13 (Fonction convexe conjuguée)**.** *Soit f une fonction convexe semi-continue inférieurement sur $\mathcal{C}$. La fonction convexe conjuguée ou duale de Fenchel-Legendre $f^\star$ de f est définie par $f^\star(y) = \sup_{x\in\mathcal{C}}\{\langle x, y\rangle - f(x)\}$.*

Théorème 1.8.4. *Soient $f : \mathcal{C} \to]-\infty,+\infty]$ une fonction convexe semi-continue inférieurement et $f^\star : \mathcal{C}^\star \to]-\infty,+\infty]$ sa fonction convexe conjuguée. Alors $(\mathcal{C}, f)$ est de Legendre si, et seulement si, $(\mathcal{C}^\star, f^\star)$ est de Legendre.*

Il est intéressant de noter que Bauschke, Borwein et Combettes [26] généralisent les fonctions de Legendre au cadre des espaces de Banach.

Topologie faible

Cette section présente la notion de topologie faible. Dans la suite, l'espace de Banach $\mathcal{L}(E, \mathbb{R})$ des formes linéaires continues sur l'espace de Banach E, appelé espace dual de E, sera noté E' (pour davantage de détails, voir Brezis [41]).

Définition 1.8.14 (Topologie faible)**.** *La topologie faible $\sigma(E, E')$ sur E est la topologie la moins fine sur E — c'est-à-dire avec le minimum d'ouverts — rendant continues toutes les formes linéaires $L \in E'$.*

Remarque 1.8.2. En dimension finie, la topologie faible et la topologie usuelle coïncident, alors qu'en dimension infinie la topologie faible est strictement moins fine.

L'intérêt principal de la topologie faible est le suivant : une topologie qui contient moins d'ouverts possède davantage de compacts. Le Théorème de Riesz (Théorème 1.8.2) affirme que la boule unité fermée d'un espace de Banach E de dimension infinie n'est jamais compacte pour la topologie forte. Cependant, nous allons voir qu'elle est compacte pour la topologie faible $\sigma(E, E')$ à condition que E vérifie une certaine propriété.

Un ensemble fermé pour la topologie faible $\sigma(E, E')$ est toujours fermé pour la topologie forte, mais la réciproque est fausse en dimension infinie. Néanmoins, la proposition suivante montre que la réciproque est vraie pour les ensembles convexes.

Proposition 1.8.5. *Soit $\mathcal{C}$ un convexe de E. Alors $\mathcal{C}$ est faiblement fermé si et seulement s'il est fortement fermé.*

Corollaire 1.8.1. *Si $f : E \to]-\infty, +\infty]$ est une fonction convexe, semi-continue inférieurement pour la topologie forte, alors f est semi-continue inférieurement pour la topologie faible $\sigma(E, E')$.*

Soit E'' l'espace dual de E'. On a une injection canonique

$$\begin{aligned} J : E &\to E'' \\ x &\mapsto (L \mapsto L(x)). \end{aligned}$$

Définition 1.8.15 (Espace réflexif)**.** *Lorsque J est surjective, E est dit réflexif.*

Théorème 1.8.5 (Kakutani)**.** *L'espace de Banach E est réflexif si, et seulement si, sa boule unité fermée est compacte pour la topologie faible $\sigma(E, E')$.*

Corollaire 1.8.2. *Si E est réflexif, tout convexe fermé borné de E est compact pour la topologie faible.*

Il en résulte le corollaire suivant, qui établit une existence de minimum.

Corollaire 1.8.3. *Soient $\mathcal{C}$ un convexe fermé non vide de l'espace de Banach réflexif E et $f : \mathcal{C} \to]-\infty, +\infty]$ une fonction faiblement semi-continue inférieurement. Supposons vérifiée l'une des deux conditions :*

- *$\mathcal{C}$ est borné.*
- $\lim_{\|x\| \to +\infty, x \in \mathcal{C}} f(x) = +\infty$.

Alors f atteint son minimum sur $\mathcal{C}$.

D'après le Corollaire 1.8.1, ce résultat s'applique dès que f est convexe semi-continue inférieurement (pour la topologie forte). Enfin, voici une propriété intéressante des espaces réflexifs relative aux suites :

Théorème 1.8.6. *Si E est réflexif, une suite bornée d'éléments de E admet une sous-suite extraite qui converge pour la topologie faible $\sigma(E, E')$.*

1.8.4. Lien entre divergences de Bregman et familles exponentielles

Plusieurs articles dans la littérature sont consacrés aux familles exponentielles (y compris en dimension infinie), en lien avec la dualité des fonctions convexes et les divergences de Bregman (voir par exemple Gamboa et Gassiat [91], Csiszár, Gamboa et Gassiat [57] et Csiszár et Matús [58, 59]). Nous présentons ici la relation entre divergences de Bregman et familles exponentielles exposée dans Banerjee, Merugu, Dhillon et Ghosh [17].

Les familles exponentielles (voir par exemple Barndorff-Nielsen [19], Azoury et Warmuth [14]) constituent une classe de modèles statistiques paramétriques de la forme $\{P_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$.

Définition 1.8.16 (Paramétrisation naturelle d'une famille exponentielle)**.** *Une famille exponentielle $\mathcal{E}_\psi$ est une famille de lois de probabilité dont la densité par rapport à une mesure σ-finie ν s'écrit sous la forme*

$$f_\theta(x) = \exp(\langle \theta, t(x) \rangle - \psi(\theta)),\ \theta \in \Theta, \tag{1.16}$$

où $\Theta = \{\theta \in \mathbb{R}^d, \psi(\theta) < +\infty\}$. Ici, θ est le paramètre naturel, $t(x)$ une statistique exhaustive et ψ la fonction génératrice des cumulants.

Comme $\ln \int f_\theta(x)\mathrm{d}\nu(x) = \ln 1 = 0$, on a

$$\psi(\theta) = \ln \int \exp(\langle \theta, t(x) \rangle)\mathrm{d}\nu(x).$$

La fonction ψ caractérise la famille exponentielle.

La représentation (1.16) est dite minimale lorsque les composantes de la statistique exhaustive sont affinement indépendantes, et d est alors appelé l'ordre de la famille exponentielle.

Définition 1.8.17 (Famille exponentielle régulière)**.** *Si la représentation (1.16) est minimale et Θ est ouvert, la famille exponentielle est dite régulière.*

Lorsque la famille exponentielle est régulière, la fonction ψ vérifie une propriété importante, comme l'indique la proposition suivante.

Proposition 1.8.6. *Si $\mathcal{E}_\psi$ est régulière, (Θ, ψ) est de Legendre.*

Explicitons à présent le raisonnement qui conduit à associer une divergence de Bregman à une famille exponentielle.

Soient $\mathcal{E}_\psi$ une famille exponentielle régulière et ϕ la fonction convexe conjuguée de ψ. Comme ψ est strictement convexe et différentiable sur le convexe ouvert Θ, son gradient $\nabla\psi$ et l'inverse du gradient $(\nabla\psi)^{-1}$ sont bien définis. Or, la fonction $x \mapsto \psi(x) - \langle y, x \rangle$ de Θ dans $\mathbb{R}$ est convexe, et son gradient s'annule en $x = (\nabla\psi)^{-1}(y)$. Elle admet donc un minimum global sur Θ en ce point (voir par exemple Rouvière [164]). Il en résulte d'après la définition de ϕ que

$$\phi(y) = \langle y, (\nabla\psi)^{-1}(y) \rangle - \psi((\nabla\psi)^{-1})$$

ou encore

$$\phi(\nabla\psi(x)) = \langle \nabla\psi(x), x \rangle - \psi(x). \tag{1.17}$$

Si X est une variable aléatoire de densité f_θ, notons

$$m = \mathbb{E}[t(X)] = \int t(x) \exp(\langle \theta, t(x) \rangle - \psi(\theta))\mathrm{d}\nu(x).$$

En dérivant par rapport à θ l'égalité

$$\int \exp(\langle \theta, t(x) \rangle - \psi(\theta))\mathrm{d}\nu(x) = 1,$$

il vient

$$\int (t(x) - \nabla\psi(\theta)) \exp(\langle \theta, t(x) \rangle - \psi(\theta))\mathrm{d}\nu(x) = 0.$$

Comme $\int f_\theta(x)\mathrm{d}\nu(x) = 1$, ceci entraîne

$$m = \nabla\psi(\theta). \tag{1.18}$$

En combinant les égalités (1.17) et (1.18), la fonction ϕ conjuguée de ψ vérifie

$$\phi(m) = \langle m, \theta\rangle - \psi(\theta), \tag{1.19}$$

et en prenant le gradient, on obtient

$$\theta = \nabla\phi(m). \tag{1.20}$$

Les égalités (1.19) et (1.20) impliquent alors

$$\begin{aligned}\langle\theta, t(x)\rangle - \psi(\theta) &= \langle m, \theta\rangle - \psi(\theta) + \langle t(x) - m, \theta\rangle \\ &= \phi(m) + \langle t(x) - m, \nabla\phi(m)\rangle \\ &= -d_\phi(t(x), m) + \phi(t(x)).\end{aligned}$$

Finalement,

$$\begin{aligned}f_\theta(x) &= \exp(\langle\theta, t(x)\rangle - \psi(\theta)) \\ &= \exp(-d_\phi(t(x), m) + \phi(t(x))).\end{aligned}$$

Définition 1.8.18 (Divergence de Bregman régulière)**.** *Soit d_ϕ une divergence de Bregman. Si ϕ est la duale de Legendre d'une fonction strictement convexe, définie sur un domaine ouvert, qui soit le logarithme de la transformée de Laplace d'une mesure finie, d_ϕ est appelée une divergence de Bregman régulière.*

Il est alors possible de démontrer la relation suivante entre familles exponentielles régulières et divergences de Bregman régulières.

Théorème 1.8.7 (Banerjee, Merugu, Dhillon et Ghosh [17])**.** *Chaque famille exponentielle régulière correspond à une unique divergence de Bregman régulière.*

Remarque 1.8.3. En fait, Banerjee *et al.* [17] prennent pour t l'identité et considèrent donc des familles exponentielles de la forme

$$f_\theta(x) = \exp(\langle\theta, x\rangle - \psi(\theta)).$$

Avant d'illustrer cette relation sur deux exemples, notons qu'elle est en rapport avec le résultat suivant, liant divergences de Bregman et distance de Kullback-Leibler (voir par exemple Nielsen, Boissonnat et Nock [149]).

Proposition 1.8.7. *Si $\mathcal{E}_\psi$ est une famille exponentielle et $\mathrm{KL}(f_{\theta_1}, f_{\theta_2})$ désigne la distance de Kullback-Leibler entre deux éléments de $\mathcal{E}_\psi$, correspondant aux paramètres naturels θ_1 et θ_2, alors*

$$\mathrm{KL}(f_{\theta_1}, f_{\theta_2}) = d_\psi(\theta_2, \theta_1) = d_\phi(m_1, m_2).$$

Exemple 1.8.1 Nous proposons de développer les exemples de la loi binomiale, associée à la perte logistique, et de la loi multinomiale, associée à la distance de Kullback-Leibler.

1. Loi binomiale. Soit X une variable aléatoire de loi binomiale $\mathcal{B}(N, p)$, c'est-à-dire
$$\mathbb{P}\{X = k\} = \binom{N}{k} p^k (1-p)^{N-k}, \quad k = 1, \ldots, N.$$
Comme
$$\begin{aligned}\ln(\mathbb{P}\{X = k\}) &= \ln\binom{N}{k} + k \ln p + (N-k)\ln(1-p) \\ &= \ln\binom{N}{k} + k \ln \frac{p}{1-p} + N \ln(1-p),\end{aligned}$$
on a
$$\mathbb{P}\{X = k\} = \binom{N}{k} \exp\left(k \ln \frac{p}{1-p} + N \ln(1-p)\right).$$
Posons $t(k) = k$ et $\theta = \ln \dfrac{p}{1-p}$. Ainsi, $\psi(\theta) = -N \ln(1-p)$. Pour exprimer p en fonction de θ, remarquons que
$$e^\theta = \frac{p}{1-p}$$
équivaut à
$$e^\theta(1-p) - p = 0,$$
c'est-à-dire
$$p(1 + e^\theta) = e^\theta.$$
Finalement,
$$p = \frac{e^\theta}{1+e^\theta} = \frac{1}{1+e^{-\theta}}.$$
La fonction ψ vérifie
$$\begin{aligned}\psi(\theta) &= -N \ln(1-p) \\ &= -N \ln\left(1 - \frac{1}{1+e^{-\theta}}\right) \\ &= -N \ln\left(\frac{e^{-\theta}}{1+e^{-\theta}}\right).\end{aligned}$$

et

$$m = \psi'(\theta) = \frac{Ne^{\theta}}{1+e^{\theta}} = Np.$$

Si ϕ désigne la fonction conjuguée de ψ, d'après l'égalité (1.19),

$$\begin{aligned}
\phi(m) &= m\theta - \psi(\theta) \\
&= m \ln\left(\frac{p}{1-p}\right) + N\ln(1-p) \\
&= m \ln\left(\frac{m}{N-m}\right) + N\ln\left(\frac{N-m}{N}\right) \\
&= (N-m)\ln\left(\frac{N-m}{N}\right) + m\ln\left(\frac{m}{N}\right).
\end{aligned}$$

Pour $(x,y) \in [0,N]\times]0,N[$, la divergence de Bregman $d_\phi(\cdot,\cdot)$ correspondante est alors définie par

$$\begin{aligned}
d_\phi(x,y) &= \phi(x) - \phi(y) - (x-y)\phi'(y) \\
&= (N-x)\ln\left(\frac{N-x}{N}\right) + x\ln\frac{x}{N} - (N-y)\ln\left(\frac{N-y}{n}\right) - y\ln\frac{y}{N} \\
&\quad - (x-y)\ln\left(\frac{y}{N-y}\right) \\
&= \boxed{x\ln\frac{x}{y} + (N-x)\ln\left(\frac{N-x}{N-y}\right)}.
\end{aligned}$$

On retrouve ainsi une perte de type logistique.

2. Loi multinomiale.

Soit X une variable aléatoire de loi multinomiale $\mathcal{M}(N, p_1, \ldots, p_d)$, avec $\sum_{j=1}^d p_j = 1$, c'est-à-dire

$$\mathbb{P}\{N_1 = n_1, \ldots, N_d = n_d\} = \frac{N!}{n_1! \ldots n_d!} \prod_{j=1}^d p_j^{n_j}, \quad \sum_{j=1}^d n_j = N, n_j \in \mathbb{N}.$$

Comme

$$\ln(\mathbb{P}\{N_1 = n_1, \ldots, N_d = n_d\}) = \ln\frac{N!}{n_1! \ldots n_d!} + \sum_{j=1}^d n_j \ln p_j,$$

on a

$$\begin{aligned}
\mathbb{P}\left\{N_1 = n_1, \ldots, N_d = n_d\right\} &= \frac{N!}{n_1! \ldots n_d!} \exp\left(\sum_{j=1}^{d} n_j \ln p_j\right) \\
&= \frac{N!}{n_1! \ldots n_d!} \exp\left(\sum_{j=1}^{d-1} n_j \ln p_j + n_d \ln p_d\right) \\
&= \frac{N!}{n_1! \ldots n_d!} \exp\left(\sum_{j=1}^{d-1} n_j \ln p_j + \Big(N - \sum_{j=1}^{d-1} n_j\Big) \ln p_d\right) \\
&= \frac{N!}{n_1! \ldots n_d!} \exp\left(\sum_{j=1}^{d-1} n_j \ln \frac{p_j}{p_d} + N \ln p_d\right) \\
&= \frac{N!}{n_1! \ldots n_d!} \exp\left(\sum_{j=1}^{d-1} n_j \ln \frac{p_j}{p_d} - N \ln \frac{1}{p_d}\right) \\
&= \frac{N!}{n_1! \ldots n_d!} \exp\left(\sum_{j=1}^{d-1} n_j \ln \frac{p_j}{p_d} - N \ln \Big(\sum_{j=1}^{d} \frac{p_j}{p_d}\Big)\right).
\end{aligned}$$

Donc

$$\mathbb{P}\left\{N_1 = n_1, \ldots, N_d = n_d\right\} = \frac{N!}{n_1! \ldots n_d!} \exp\left(\sum_{j=1}^{d-1} n_j \ln \frac{p_j}{p_d} - N \ln \Big(1 + \sum_{j=1}^{d-1} \frac{p_j}{p_d}\Big)\right).$$

Posons $t(k) = k$ et $\theta_j = \ln \frac{p_j}{p_d}$, $j = 1, \ldots, d-1$. Ainsi,

$$\psi(\theta) = N \ln \Big(1 + \sum_{j=1}^{d-1} e^{\theta_j}\Big),$$

et $m = \nabla\psi(\theta)$ est le vecteur composé des $\frac{Ne^{\theta_j}}{1+\sum_{j=1}^{d-1} e^{\theta_j}} = Np_j, \quad j = 1, \ldots d-1$.
Si ϕ désigne la fonction conjuguée de ψ,

$$\begin{aligned}
\phi(m) &= \langle m, \theta\rangle - \psi(\theta) \\
&= \sum_{j=1}^{d-1} Np_j \ln \frac{p_j}{p_d} + N \ln p_d \\
&= \sum_{j=1}^{d-1} Np_j \ln p_j + \Big(1 - \sum_{j=1}^{d-1} p_j\Big) N \ln p_d \\
&= \sum_{j=1}^{d} Np_j \ln p_j \\
&= N \sum_{j=1}^{d} \frac{m_j}{N} \ln \frac{m_j}{N}.
\end{aligned}$$

Pour $(x, y) \in (\mathbb{R}^+)^d \times (\mathbb{R}^{+*})^d$ ayant des composantes de somme égale à N, la divergence de Bregman $d_\phi(\cdot, \cdot)$ associée est donnée par

$$\begin{aligned} d_\phi(x, y) &= \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle \\ &= N \sum_{j=1}^{d} \frac{x_j}{N} \ln \frac{x_j}{N} - N \sum_{j=1}^{d} \frac{y_j}{N} \ln \frac{y_j}{N} - \sum_{j=1}^{d} (x_j - y_j)\Big(1 + \ln \frac{y_j}{N}\Big) \\ &= \boxed{N \sum_{j=1}^{d} \frac{x_j}{N} \ln \frac{x_j/N}{y_j/N}}. \end{aligned}$$

Il s'agit d'un multiple de la distance de Kullback-Leibler entre les vecteurs $\frac{x}{N}$ et $\frac{y}{N}$.

2. Projection-based curve clustering*

Sommaire

*Article écrit en collaboration avec Benjamin Auder. A paraître dans *Journal of Statistical Computation and Simulation*.

Abstract

This paper focuses on unsupervised curve classification in the context of nuclear industry. At the Commissariat à l'Energie Atomique (CEA), Cadarache (France), the thermal-hydraulic computer code CATHARE is used to study the reliability of reactor vessels. The code inputs are physical parameters and the outputs are time evolution curves of a few other physical quantities. As the CATHARE code is quite complex and CPU-time consuming, it has to be approximated by a regression model. This regression process involves a clustering step. In the present paper, CATHARE output curves are clustered using a k-means scheme, with a projection onto a lower dimensional space. We study the properties of the empirically optimal cluster centers found by the clustering method based on projections. The choice of the projection basis is discussed, and an algorithm is implemented to select the best projection basis among a library of orthonormal bases. The approach is illustrated on a simulated example and then applied to the industrial problem.

2.1. Introduction

2.1.1. The CATHARE code

A major concern in nuclear industry is the life span of reactor vessels. To go on using the current nuclear reactors, their reliability has to be proved. For this purpose, complex computer codes are developed to simulate the behavior of the vessel under different sequences of accidents. At the Commissariat à l'Energie Atomique (CEA), Cadarache (France), one of the main types of accident under study is the pressurized thermal shock. This is a problem due to the combined stresses from a rapid temperature and pressure change. More specifically, as a reactor vessel gets older, the potential for failure by cracking when it is cooled rapidly at high pressure increases greatly. The analysis of pressurized thermal shock is made of two main steps. First, a thermal-hydraulic analysis is done to determine the temporal evolutions of temperature, pressure and thermal exchange coefficient in the vessel annular space, since these features have an influence on the mechanical and thermal charge on the vessel inner surface. Some evolution curves $x_1(t), \ldots, x_n(t)$ corresponding to the thermal exchange coefficient are depicted in Figure 2.1 ($n = 66$). Each curve $x_i(t)$ is obtained as the simulation result for a certain vector of input physical parameters. The curves of temperature, pressure and thermal exchange coefficient obtained during this first step are then used as limit conditions in the

second part of the analysis, which is a mechanical investigation aiming at checking if some defects on the vessel annular space could propagate and gain importance to such an extent that this would cause a break of the vessel inner surface. For further details on the reliability of reactor vessels, we refer the reader to Auder, De Crecy, Iooss, and Marquès [13].

The simulation step relies on a computer code called CATHARE (Code Avancé de THermohydraulique pour les Accidents des Réacteurs à Eau, in English Code for Analysis of THermalhylaudrics during an Accident of Reactor and safety Evaluation). The CATHARE code is a system code for pressurized water reactors safety analysis, accident management, definition of plant operating procedures and for research and development. The project is a result of a joint effort of the reactor vendor AREVA, the CEA, EDF (Electricité de France) and the IRSN (Institut de Radioprotection et de Sûreté Nucléaire). The first delivered version V1.3L was available in 1997. The CEA team CATHARE located in Grenoble (France) is in charge of the development, the assessment and the maintenance of the code. (See `http://www-cathare.cea.fr`.)

The CATHARE code allows to simulate the evolution of temperature, pressure and thermal exchange coefficient, given the physical parameters as inputs. However, this code is so slow (about 6 to 10 hours for one run) that it cannot be used directly for reliability calculations. To bypass this obstacle, the strategy drawn up by the CEA is to build a so-called metamodel which is a fast approximation of the original code, precise enough to carry out statistical computations. The term "metamodel" indicates that a computer code approximating a physical process has already been developed, and now this code is modeled in turn. Here, as summarized in a flowchart in Figure 2.2, the purpose is the construction of a regression model based on a few hundreds CATHARE code outputs, obtained during one week of computation on a supercomputer in 2007. . The inputs were sampled randomly by latin hypercube methods (see, e.g., McKay, Conover, and Beckman [144] and Loh [133]), so that we have no control over the inputs in the learning sample. As different kinds of behavior for temperature, pressure and thermal exchange coefficient may be observed depending on the physical parameters, a preliminary unsupervised classification of CATHARE code output curves is essential. Once the curves have been clustered in meaningful classes, the regression model can be adjusted for each group of outputs separately. The clustering step is the object of the present paper.

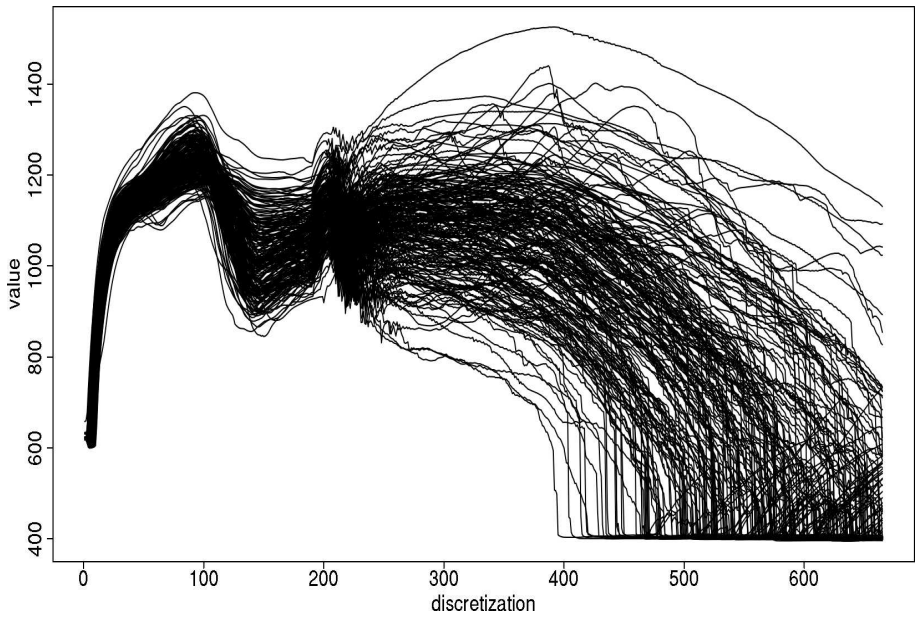


FIGURE 2.1.: 66 evolution curves of thermal exchange coefficient in a nuclear vessel.

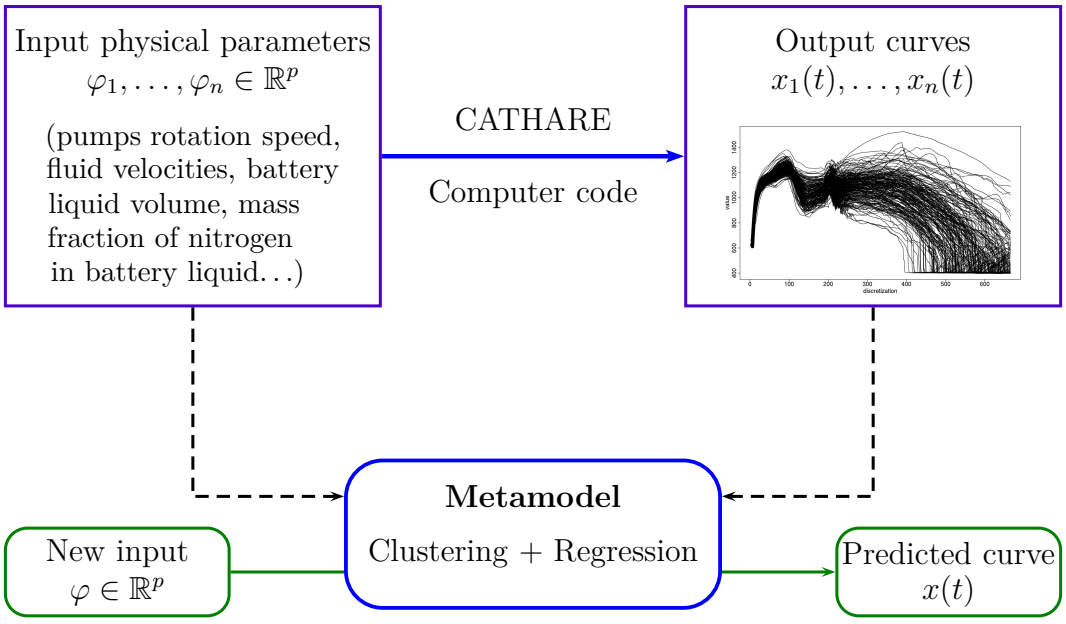


FIGURE 2.2.: Flowchart of the CATHARE code.

2.1.2. Clustering

Clustering is the problem of partitioning data into a finite number of groups (denoted hereafter by k), or clusters, so that the data items inside each of them are very similar among themselves and as different as possible from the elements of the other clusters (Duda, Hart, and Stork [76, Chapter 10]). In our industrial context, the data is made of evolution curves of temperature, pressure or thermal exchange coefficient. Using a probabilistic point of view, these curves can be seen as independent draws $X_1(t), \dots, X_n(t)$ with the same distribution as a generic random variable $X(t)$ taking values in a functional space $(E, \|\cdot\|)$ — typically, the Hilbert space of square integrable functions.

A widely used clustering method is the so-called k-means clustering, which consists in partitioning the random observations $X_1, \dots, X_n \in E$ into k classes by minimizing the empirical distortion

$$W_{\infty,n}(\mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{\ell=1,\dots,k} \|X_i - c_\ell\|^2,$$

over all possible cluster centers $\mathbf{c} = (c_1, \dots, c_k) \in E^k$. Here, μ_n denotes the empirical measure associated with the sample $X_1, \dots, X_n$, i.e.,

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

for every Borel subset A of E. In other words, we look for a Voronoi partition of E. The Voronoi partition $C_1, \dots, C_k$ associated with $\mathbf{c} = (c_1, \dots, c_k) \in E^k$ is defined by letting an element $x \in E$ belong to C_ℓ if it is closer (with respect to the norm $\|\cdot\|$) to c_ℓ than to any other c_j (ties are broken arbitrarily). The ℓ-th cluster is made of the observations X_i assigned to c_ℓ, or equivalently, falling in the Voronoi cell C_ℓ. In this framework, the accuracy of the clustering scheme is assessed by the distortion or mean squared error

$$W_\infty(\mathbf{c}) = \mathbb{E}\Big[\min_{\ell=1,\dots,k} \|X - c_\ell\|^2 \Big],$$

where $\mathbb{E}$ stands for expectation with respect to the distribution of X. This clustering method is in line with the more general theory of quantization. More specifically, it corresponds to the empirical version of nearest neighbor quantization (Linder [131], Gersho and Gray [94], Graf and Luschgy [98]). However, the problem of finding a minimizer of the criterion $W_{\infty,n}(\mathbf{c})$ is in general NP-hard, and there is no efficient algorithm to find the optimal solution in reasonable time. That is why several iterative algorithms have been developed to give approximate solutions. One of the first to be described historically is Lloyd's algorithm (Lloyd [132]).

The challenge is to adapt the k-means clustering method to our setting. The main difficulty here is the high dimensionality of the data, which casts the problem into the general class of functional statistics. For a comprehensive introduction to this topic, see Ramsay and Silverman [159] and Ferraty and Vieu [84]. A possible approach to reduce the infinite dimension of the observations $X_1, \ldots, X_n$ consists in projecting them onto a lower-dimensional subspace. In this context, Abraham, Cornillon, Matzner-Løber, and Molinari [1] project the curves on a B-spline basis, and get clusters with the k-means algorithm applied to the coefficients. These authors argue that projecting onto a smooth spline basis plays the role of a denoising procedure, given that the observed curves could contain measurement errors, and also allows to deal with curves which were not measured at the same time. James and Sugar [110] use a B-spline basis to model the centers of the clusters and write each curve in cluster ℓ as a main effect defined by spline coefficients plus an error term. This allows for some deviations around a model curve specific to cluster ℓ. The authors add a Gaussian error term to model the individual variations among one cluster. This way, the main effect is enriched, and the model can take into account more complex behaviors. The method in Gaffney [90] is similar, also with a B-spline basis. Another option is to use a Self-Organizing Map algorithm on the coefficients (Rossi, Conan-Guez, and El Golli [162]), again obtained by projecting the functions onto a B-spline basis. These bases are often used because they are easy to implement, and require a relatively minimal number of parametric assumptions. Besides, Biau, Devroye, and Lugosi [32] examine the theoretical performance of clustering with random projections based on the Johnson-Lindenstrauss Lemma, which represent a sound alternative to orthonormal projections thanks to their distance-preserving properties. Chiou and Li [51] propose a method which generalizes the k-means algorithm to some extent, by considering covariance structures via functional principal component analysis. In the approach of these authors, each curve is decomposed on an adaptive local basis (valid for the elements in the cluster), and the clusters are determined according to the full approximation onto each basis. In the wavelet-based method for functional data clustering developed in Antoniadis, Brossat, Cugliari, and Poggi [8], a smooth curve is reduced to a finite number of representative features, by considering the contribution of each wavelet coefficient to the global energy of the curve. Recently, Bruna and Mallat [42] have discussed a dimensionality reduction method which consists in choosing the dimension of the approximation space by model selection via penalization.

In the present contribution, we propose to investigate the problem of clustering output curves $X_1, \ldots, X_n$ of the CATHARE code, assuming that they arise from a random variable X taking its values in some subset of the space of square integrable functions. As a general strategy, we reduce the infinite dimension of

X by considering only the first d coefficients of the expansion on a Hilbert basis, and then perform clustering in $\mathbb{R}^d$. We study the theoretical properties of this clustering method with projection. A bound expressing what is lost when replacing the empirically optimal cluster centers by the centers obtained by projection is offered (Section 2.2). Since the result may depend on the basis choice, several projection bases are used in practice, and we look for the best one minimizing a given criterion. To this end, an algorithm based on Coifman and Wickerhauser [54] is implemented, searching for an optimal basis among a library of wavelet packet bases available in the **R** package *wmtsa*, and this "optimal basis" is compared with the Fourier basis, the Haar basis, and the functional principal component basis (Section 2.3). Finally, this algorithm is applied to a simulated example and to our industrial problem (Section 2.4). Proofs are postponed to Section 2.6.

2.2. Finite-dimensional projection for clustering

As mentioned earlier, we are concerned with square integrable functions. Since all results can be adapted to $L^2([a,b])$ by an appropriate rescaling, we consider for the sake of simplicity the space $L^2([0,1])$. As an infinite-dimensional separable Hilbert space, $L^2([0,1])$ is isomorphic via the choice of a Hilbert basis to the space ℓ^2 of square-summable sequences. We focus more particularly on functions in $L^2([0,1])$ whose coefficients in the expansion on a given Hilbert basis belong to the subset $\mathcal{S}$ of ℓ^2 given by

$$\mathcal{S} = \left\{ \mathbf{x} = (x_j)_{j\geq 1} \in \ell^2 : \sum_{j=1}^{+\infty} \varphi_j x_j^2 \leq R^2 \right\}, \tag{2.1}$$

where $R > 0$ and $(\varphi_j)_{j\geq 1}$ is a nonnegative increasing sequence such that

$$\lim_{j\to+\infty} \varphi_j = +\infty.$$

It is worth pointing out that $\mathcal{S}$ is closely linked with the basis choice, even if the basis does not appear explicitly in the definition. To illustrate this important fact, three examples are discussed below.

Example 2.2.1 (Sobolev ellipsoids). For $\beta \in \mathbb{N}^*$ and $L > 0$, the periodic Sobolev class $W^{per}(\beta, L)$ is the space of all functions $f \in [0,1] \to \mathbb{R}$ such that $f^{(\beta-1)}$ is absolutely continuous, $\int_0^1 (f^{(\beta)}(t))^2 \mathrm{d}t \leq L^2$ and $f^{(\ell)}(0) = f^{(\ell)}(1)$ for $\ell = 0, \ldots, \beta-1$. Let $(\psi_j)_{j\geq 1}$ denote the trigonometric basis. Then a function $f = \sum_{j=1}^{+\infty} x_j \psi_j$ is in $W^{per}(\beta, L)$ if and only if the sequence $\mathbf{x} = (x_j)_{j\geq 1}$ of its Fourier coefficients belongs to

$$\mathcal{S} = \left\{ \mathbf{x} \in \ell^2 : \sum_{j=1}^{+\infty} \varphi_j x_j^2 \leq R^2 \right\},$$

where

$$\varphi_j = \begin{cases} j^{2\beta} & \text{for even } j \\ (j-1)^{2\beta} & \text{for odd } j \end{cases}$$

and $R = \dfrac{L}{\pi^\beta}$.

For the proof of this result and further details about Sobolev classes, we refer the reader to the book of Tsybakov [183]. Note that the set $\mathcal{S}$ could also be defined by $\varphi_j = j^r e^{\alpha j}$ with $\alpha > 0$ and $r \geq -\alpha$ (Tsybakov [182]).

Example 2.2.2 (Reproducing Kernel Hilbert Spaces). Let $K : [0,1] \times [0,1] \to \mathbb{R}$ be a Mercer kernel, i.e., K is continuous, symmetric and positive definite. Recall that a kernel K is said to be positive definite if for all finite sets $\{x_1, \ldots, x_m\}$, the matrix A defined by $a_{ij} = K(x_i, x_j)$ for $1 \leq i, j \leq m$ is positive definite. For example, the Gaussian kernel $K(x,y) = \exp(-\frac{(x-y)^2}{\sigma^2})$ and the kernel $K(x,y) = (c^2 + (x-y)^2)^{-a}$ with $a > 0$ are Mercer kernels. For $x \in [0,1]$, let $K_x : y \mapsto K(x,y)$. Then, Moore-Aronszajn's Theorem (Aronszajn [11]) states that there exists a unique Hilbert space $(\mathcal{H}_K, \langle \cdot, \cdot \rangle)$ of functions on $[0,1]$ such that:

1. For all $x \in [0,1]$, $K_x \in \mathcal{H}_K$.
2. The span of the set $\{K_x, x \in [0,1]\}$ is dense in $\mathcal{H}_K$.
3. For all $f \in \mathcal{H}_K$ and $x \in [0,1]$, $f(x) = \langle K_x, f \rangle$.

The Hilbert space $\mathcal{H}_K$ is said to be the reproducing kernel Hilbert space (for short, RKHS) associated with the kernel K. Next, the operator $\mathcal{K}$ defined by

$$\mathcal{K}f : y \mapsto \int_0^1 K(x,y) f(x) \mathrm{d}x$$

is self-adjoint, positive and compact. Consequently, there exists a complete orthonormal system $(\psi_j)_{j \geq 1}$ of $L^2([0,1])$ such that $\mathcal{K}\psi_j = \lambda_j \psi_j$, where the set of eigenvalues $\{\lambda_j, j \geq 1\}$ is either finite or a sequence tending to 0 at infinity. Moreover, the λ_j are nonnegative. Suppose that $\mathcal{K}$ is not of finite rank — so that $\{\lambda_j, j \geq 1\}$ is infinite — and that the eigenvalues are sorted in decreasing order, that is $\lambda_j \geq \lambda_j + 1$ for all $j \geq 1$. Clearly, there is no loss of generality in assuming that $\lambda_j > 0$ for all $j \geq 1$. Indeed, if not, $L^2([0,1])$ is replaced by the linear subspace spanned by the eigenvectors corresponding to non-zero eigenvalues.

According to Mercer's theorem, K has the representation

$$K(x,y) = \sum_{j=1}^{+\infty} \lambda_j \psi_j(x) \psi_j(y),$$

where the convergence is absolute and uniform (Cucker and Smale [60, Chapter III, Theorem 1]). Moreover, $\mathcal{H}_K$ may be characterized through the eigenvalues of the operator $\mathcal{K}$ by

$$\mathcal{H}_{\mathcal{K}} = \Big\{ f \in L^2([0,1]) : f = \sum_{j=1}^{+\infty} x_j \psi_j, \sum_{j=1}^{+\infty} \frac{x_j^2}{\lambda_j} < \infty \Big\},$$

with the inner product

$$\Big\langle \sum_{j=1}^{+\infty} x_j \psi_j, \sum_{j=1}^{+\infty} y_j \psi_j \Big\rangle = \sum_{j=1}^{+\infty} \frac{x_j y_j}{\lambda_j}$$

(Cucker and Smale [60, Chapter III, Theorem 4]). Then, letting

$$\mathcal{S} = \Big\{ \mathbf{x} \in \ell^2, \sum_{j=1}^{+\infty} \frac{x_j^2}{\lambda_j} \leq R^2 \Big\},$$

the set $\mathcal{S}$ is of the desired form (2.1), with $\varphi_j = 1/\lambda_j$.

Example 2.2.3 (Besov ellipsoids and wavelets). Let $\alpha > 0$. For $f \in L^2([0,1])$, the Besov semi-norm $|f|_{B_2^\alpha}(L^2)$ is defined by

$$|f|_{B_2^\alpha}(L^2) = \Big(\sum_{j=0}^{+\infty} [2^{j\alpha} \omega_r(f, 2^{-j}, [0,1])_2]^2 \Big)^{1/2}$$

where $\omega_r(f,t,[0,1])_2$ denotes the modulus of smoothness of f, as defined for instance in DeVore and Lorentz [70], and $r = \lfloor \alpha \rfloor + 1$. Let $\Lambda(j)$ be an index set at resolution level j and $(x_{j,\ell})_{j \geq 0, \ell \in \Lambda(j)}$ the coefficients of the expansion of f in a suitable wavelet basis. Then, for f such that $|f|_{B_2^\alpha}(L^2) \leq \rho$, the coefficients $x_{j,\ell}$ satisfy

$$\sum_{j=0}^{+\infty} \sum_{\ell \in \Lambda(j)} 2^{2j\alpha} x_{j,\ell}^2 \leq \rho^2 C^2,$$

where $C > 0$ depends only on the basis. We refer to Donoho and Johnstone [74] for more details.

Let us now come back to the general setting

$$\mathcal{S} = \Big\{ \mathbf{x} = (x_j)_{j \geq 1} \in \ell^2 : \sum_{j=1}^{+\infty} \varphi_j x_j^2 \leq R^2 \Big\},$$

and consider the problem of clustering the sample $X_1, \ldots, X_n$ with values in $\mathcal{S}$. Some notation and assumptions are in order. First, we will suppose that

$$\mathbb{P}\{\|X\| \leq R\} = 1.$$

Notice that the fact that X takes its values in $\mathcal{S}$ is in general not enough to imply $\mathbb{P}\{\|X\| \leq R\} = 1$. Secondly, let j_0 be the smallest integer j such that $\varphi_j > 0$. To avoid technical difficulties, we require in the sequel $d \geq j_0$. For all $d \geq 1$, we will denote by Π_d the orthogonal projection on $\mathbb{R}^d$ and let $\mathcal{S}_d = \Pi_d(\mathcal{S})$. Lastly, observe that $\mathcal{S}_d$ identifies with the ellipsoid

$$\Big\{\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d : \sum_{j=1}^{d} \varphi_j x_j^2 \leq R^2\Big\}.$$

As explained in the introduction, the criterion to minimize is

$$W_{\infty,n}(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n} \min_{\ell=1,\ldots,k} \|X_i - c_\ell\|^2,$$

and the performance of the clustering obtained with the centers $\mathbf{c} = (c_1, \ldots, c_k) \in \mathcal{S}^k$ is measured by the distortion

$$W_\infty(\mathbf{c}) = \mathbb{E}\Big[\min_{\ell=1,\ldots,k} \|X - c_\ell\|^2\Big].$$

The quantity

$$W_\infty^\star = \inf_{\mathbf{c}\in\mathcal{S}^k} W_\infty(\mathbf{c})$$

represents the optimal risk we can achieve. With the intention of performing clustering in the projection space $\mathcal{S}^d$, we also introduce the "finite-dimensional" distortion

$$W_d(\mathbf{c}) = \mathbb{E}\Big[\min_{\ell=1,\ldots,k} \|\Pi_d(X) - \Pi_d(c_\ell)\|^2\Big]$$

and its empirical counterpart

$$W_{d,n}(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n} \min_{\ell=1,\ldots,k} \|\Pi_d(X_i) - \Pi_d(c_\ell)\|^2,$$

as well as

$$W_d^\star = \inf_{\mathbf{c}\in\mathcal{S}^k} W_d(\mathbf{c}).$$

Let us observe that, as the support of the empirical measure μ_n contains at most n points, there exists an element $\hat{\mathbf{c}}_{d,n}$ which is a minimizer of $W_{d,n}(\mathbf{c})$ on $\mathcal{S}^k$. Moreover, in view of its definition, $W_{d,n}(\mathbf{c})$ only depends on the centers projection $\Pi_d(\mathbf{c})$ (one has $W_{d,n}(\mathbf{c}) = W_{d,n}(\Pi_d(\mathbf{c}))$ for all $\mathbf{c}$) and we can thus assume that $\hat{\mathbf{c}}_{d,n} \in (\mathcal{S}_d)^k$. Notice also that for all $c \in \mathcal{S}$,

$$\|\Pi_d(X) - \Pi_d(c)\|^2 \leq \|X - c\|^2$$

(the projection Π_d is 1-Lipschitz), which implies, for all $\mathbf{c}$,

$$W_d(\mathbf{c}) \leq W_\infty(\mathbf{c}).$$

The following lemma provides an upper bound for the maximal deviation

$$\sup_{\mathbf{c}\in\mathcal{S}^k} [W_\infty(\mathbf{c}) - W_d(\mathbf{c})].$$

Lemma 2.2.1. *We have*

$$\sup_{\mathbf{c}\in\mathcal{S}^k} [W_\infty(\mathbf{c}) - W_d(\mathbf{c})] \leq \frac{4R^2}{\varphi_d}.$$

We are now in a position to state the main result of this section.

Theorem 2.2.1. *Let $\hat{\mathbf{c}}_{d,n} \in (\mathcal{S}_d)^k$ be a minimizer of $W_{d,n}(\mathbf{c})$. Then,*

$$\mathbb{E}[W_\infty(\hat{\mathbf{c}}_{d,n})] - W_\infty^\star \leq \mathbb{E}[W_d(\hat{\mathbf{c}}_{d,n})] - W_d^\star + \frac{8R^2}{\varphi_d}. \tag{2.2}$$

Theorem 2.2.1 expresses the fact that the expected excess clustering risk in the infinite dimensional space is bounded by the corresponding "finite-dimensional risk" plus an additional term representing the price to pay when projecting onto $\mathcal{S}_d$. Yet, the first term in the right-hand side of inequality (2.2) above is known to tend to 0 when n goes to infinity. More precisely, as $\mathbb{P}\{\|\Pi_d(X)\| \leq R\} = 1$, we have

$$\mathbb{E}[W_d(\hat{\mathbf{c}}_{d,n})] - W_d^\star \leq \frac{Ck}{\sqrt{n}},$$

where $C = 12R^2$ (Biau, Devroye, and Lugosi [32]). In our setting, to keep the same rate of convergence $O(1/\sqrt{n})$ in spite of the extra term $8R^2/\varphi_d$, φ_d must be of the order $\sqrt{n}$. For Sobolev ellipsoids (Example 2.2.1), where $\varphi_j \geq (j-1)^{2\beta}$, this means a dimension d of the order $n^{1/4\beta}$. When $\varphi_j = j^r e^{\alpha j}$, the rate of convergence is $O(1/\sqrt{n})$ as long as d is chosen of the order $\ln n/(2\alpha)$. In the RKHS context (Example 2.2.2), consider the case of eigenvalues $\{\lambda_j, j \geq 1\}$ with polynomial or exponential-polynomial decay, which covers a broad range of kernels (Williamson, Smola, and Schölkopf [188]). If $\lambda_j = O(j^{-(\alpha+1)}), \alpha > 0$, then $1/\varphi_d = O(d^{-(\alpha+1)})$, and d must be of the order $n^{1/(2\alpha+2)}$, whereas $\lambda_j = O(e^{-\alpha j^p}), \alpha, p > 0$, leads to a projection dimension d of the order $(\ln n/(2\alpha))^{1/p}$. Obviously, the upper bound (2.2) is better for large φ_d, and consequently large d. Nevertheless, from a computational point of view, the projection dimension should not be chosen too large.

Remark 2.2.1. Throughout, we assumed that $\mathbb{P}\{\|X\| \leq R\} = 1$. This requirement, called the peak power constraint, is standard in the clustering and signal processing literature. We do not consider in this paper the case where this assumption is not satisfied, which is feasible but leads to technical complications (see Merhav and Ziv [145], Biau, Devroye, and Lugosi [32] for results in this direction). Besides, the number of clusters is assumed to be fixed throughout the paper. Several methods for estimating k have been proposed in the literature (see, e.g., Milligan and Cooper [147] and Gordon [97]).

As already mentioned, the subset of coefficients $\mathcal{S}$ is intimately connected to the underlining Hilbert basis. As a consequence, all the results presented strongly depend on the orthonormal system considered. Therefore, the choice of a proper basis is crucial and is discussed in the next section.

2.3. Basis selection

Wavelet packet best basis algorithm In this section, we describe an algorithm searching for the best projection basis among a "library". If $\{\psi_\alpha, \alpha \in I\} \subset L^2([0,1])$ is a collection of elements in $L^2([0,1])$ which span $L^2([0,1])$ and allow to build several different bases by choosing various subsets $\{\psi_\alpha, \alpha \in I_\beta\} \subset L^2([0,1])$, the collection of bases built this way is called a library of bases. Here, I is some index set, and β runs over some other index set.

More specifically, we focus on the best basis algorithm of Coifman and Wickerhauser [54] (see also Wickerhauser [187]), which yields an optimal basis among a library of wavelet packets. Wavelets are functions which cut up a signal into different frequency components to study each component with a resolution matched to its scale. Unlike the Fourier basis, wavelets are localized both in time and frequency. Hence, they have advantages over traditional Fourier methods when the signal contains discontinuities as well as noise. For detailed expositions of the mathematical aspects of wavelets, see the books of Daubechies [61], Mallat [138] and Meyer [146].

Let the sequence of functions $(\psi_\nu)_{\nu \geq 0}$ be defined by

$$\begin{aligned}
\psi_0(t) &= H\psi_0(t), \quad \int_{\mathbb{R}} \psi_0(t)\mathrm{d}t = 1, \\
\psi_{2\nu}(t) &= H\psi_\nu(t) = \sqrt{2}\sum_{p\in\mathbb{Z}} h(p)\psi_\nu(2t-p), \\
\psi_{2\nu+1}(t) &= G\psi_\nu(t) = \sqrt{2}\sum_{p\in\mathbb{Z}} g(p)\psi_\nu(2t-p),
\end{aligned}$$

where H and G are orthogonal quadrature filters, i.e., convolution-decimation operators satisfying some algebraic properties (see, e.g., Wickerhauser [187]). Let Λ_ν denote the closed linear span of the translates $\psi_\nu(\cdot - p), p \in \mathbb{Z}$, of ψ_ν, and

$$\sigma^s \Lambda_\nu = \{2^{-s/2} x(2^{-s}t), x \in \Lambda_\nu\}.$$

To every such subspace of $L^2(\mathbb{R})$ corresponds a dyadic interval

$$I_{s\nu} = \left[\frac{\nu}{2^s}, \frac{\nu+1}{2^s}\right[.$$

For all (s, ν), these subspaces give an orthogonal decomposition

$$\sigma^s \Lambda_\nu = \sigma_{s+1}\Lambda_{2\nu} \overset{\perp}{\oplus} \sigma_{s+1}\Lambda_{2\nu+1}.$$

Observe that for $\nu = 0, \ldots, 2^s - 1$, the $I_{s\nu}$ are dyadic subintervals of $[0, 1[$.

The next proposition provides a library of orthonormal bases built with functions of the form $\psi_{s\nu p} = 2^{-s/2}\psi_\nu(2^{-s}t - p)$, called wavelet packets of scale index s, frequency index ν and position index p.

Proposition 2.3.1 (Wickerhauser [187]). *If $s \leq L$ for some finite maximum L, H and G are orthogonal quadrature filters and $\mathcal{I}$ is a collection of disjoint dyadic intervals whose union is $\mathbb{R}^+$, then $\mathcal{B}_\mathcal{I} = \{\psi_{s\nu p}, p \in \mathbb{Z}, I_{s\nu} \in \mathcal{I}\}$ is an orthonormal basis for $L^2(\mathbb{R})$. Moreover, if $\mathcal{I}$ is a disjoint dyadic cover of $[0, 1[$, then $\mathcal{B}_\mathcal{I}$ is an orthonormal basis of Λ_0.*

This construction yields orthonormal bases of $L^2(\mathbb{R})$. Some changes must be made to obtain bases of $L^2([0, 1])$. Roughly, they consist in considering not all scales and shifts, and adapting the wavelets which overlap the boundary of $[0, 1]$ (see for instance Cohen, Daubechies, and Vial [53]).

The library can be seen as a binary tree whose nodes are the spaces $\sigma^s\Lambda_\nu$ (see Figures 2.3 and 2.4). An orthonormal basis is given by the leaves of some subtree. Figures 2.5 and 2.6 show two examples of bases which can be obtained in this way.

To define an optimal basis, a notion of information cost is needed. Coifman and Wickerhauser [54] propose to use Shannon entropy. In our context, the basis choice will be done with respect to some reference curve x_0 which has to be representative of the data. We compute, for each basis in the library, the Shannon entropy of the coefficients of x_0 in this basis, and select the basis minimizing this entropy. The construction of this best basis, relying on the binary tree structure of the library, is achieved by comparing, at each node, starting from the bottom of the tree,

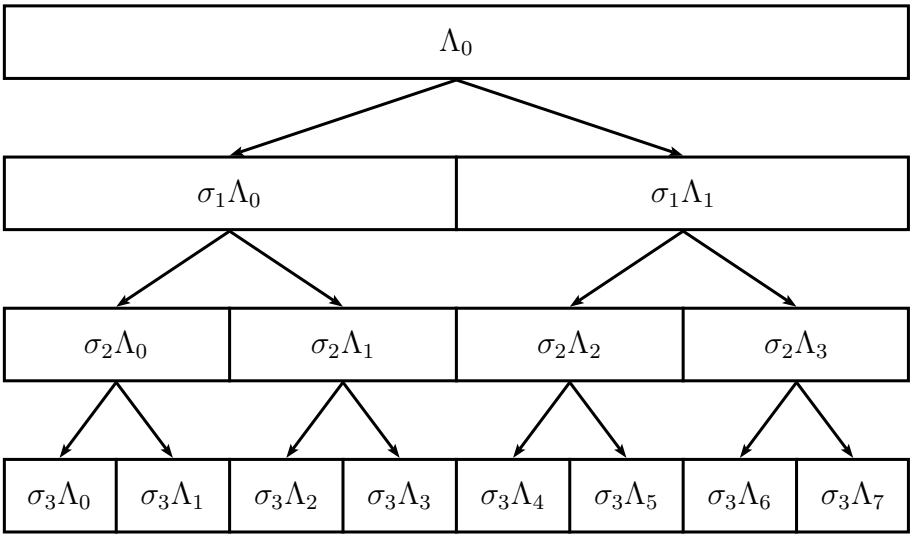


FIGURE 2.3.: Tree structure of wavelet packet bases.

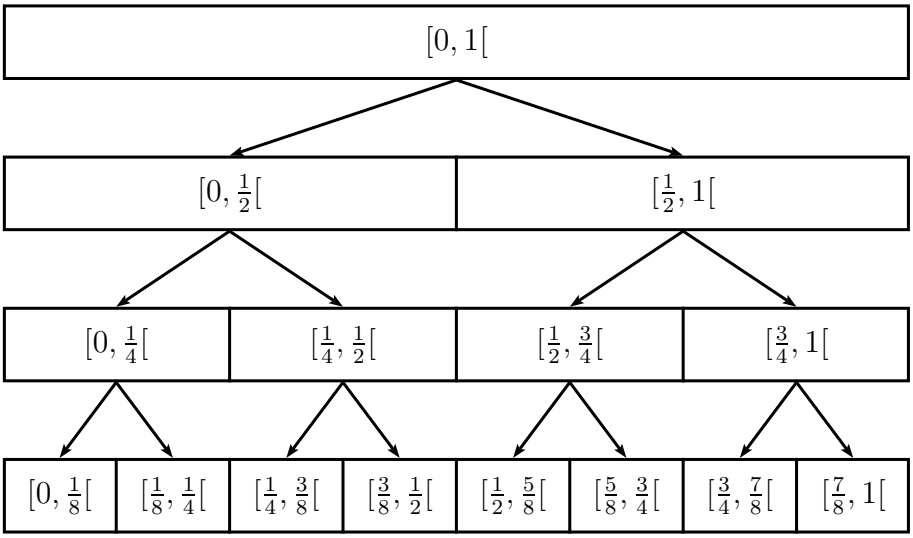


FIGURE 2.4.: Correspondence with dyadic covers of [0,1[.

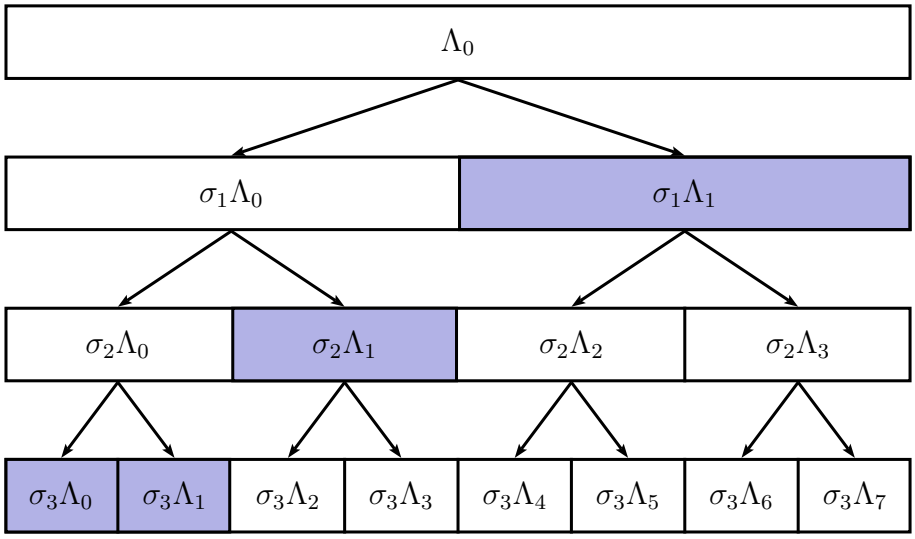


FIGURE 2.5.: The wavelet basis.

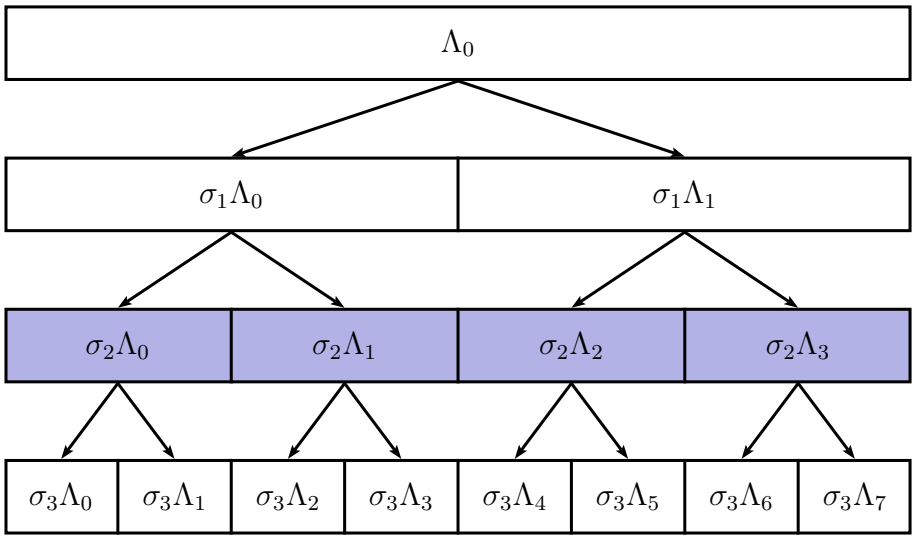


FIGURE 2.6.: An example of fixed level wavelet packet basis.

parents with their two children. We decided to take x_0 as the median curve in the sample in the L^2 norm sense. Indeed, it is more likely to present characteristic behaviors of the other curves than the mean, because the mean curve is a smooth average representative, which is probably too easy to approximate with a few basis functions. However, we observed that these two functions surprisingly give rise to almost the same basis. Both choices are thus possible. The mean curve is useful if we know that some noise has to be removed, whereas the median curve seems a better choice to reflect the small-scale irregularities.

We have implemented the algorithm in **R**, and it has been run through all filters available in the **R** package *wmtsa*. These filters belong to four families, extremal phase family (Daubechies wavelets), including Haar basis, least asymmetric family (Symmlets), "best localized" wavelets and Coiflets. For example, the least asymmetric family contains ten different filters "s2", "s4", "s6", "s8", "s10", "s12", "s14", "s16", "s18", "s20". Finally, we keep the clustering result obtained with the basis minimizing the distortion among the various filters. In the sequel, this basis will be called Best-Entropy basis.

In the applications, the performance of the Best-Entropy basis will be compared with the Haar wavelet basis, the Fourier basis and the functional principal component analysis basis. For the sake of completeness, we recall here the definition of these bases.

The Haar wavelet basis Let $\phi(t) = \mathbf{1}_{[0,1]}(t)$ and $\psi(t) = \mathbf{1}_{[0,1/2[}(t) - \mathbf{1}_{[1/2,1]}(t)$ (see Figure 2.7). Then, the family $\{\phi, \psi_{j,\ell}\}$, where

$$\psi_{j,\ell}(t) = 2^{j/2}\psi(2^j t - \ell), j \geq 0, 0 \leq \ell \leq 2^j - 1,$$

constitutes a Hilbert basis of $L^2([0,1])$, called Haar basis.

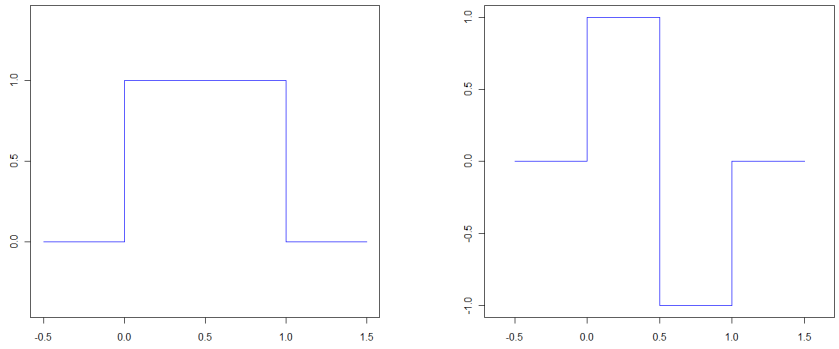


FIGURE 2.7.: Haar scaling function ϕ and mother wavelet function ψ.

The Fourier basis The Fourier basis on $[0,1]$ is the complete orthonormal system of $L^2([0,1])$ built with the trigonometric functions

$$\psi_1(t) = 1, \quad \psi_{2j}(t) = \sqrt{2}\cos(2\pi jt), \quad \psi_{2j+1}(t) = \sqrt{2}\sin(2\pi jt), \quad j \geq 1.$$

Functional principal component analysis Principal component analysis for functional data (for short, functional PCA) is the generalization of the usual principal component analysis for vector data. The Euclidean inner product is replaced by the inner product in $L^2([0,1])$. More precisely, functional PCA consists in writing $X_i(t)$ as

$$X_i(t) = \mathbb{E}[X(t)] + \sum_{j=1}^{+\infty} x_{ij}\psi_j(t),$$

where the $(x_{ij})_{j\geq 1}$ and the functions $(\psi_j)_{j\geq 1}$ are defined as follows. At the first step, the function ψ_1 is chosen to maximize the variance of

$$x_{i1} = \int \psi_1(t)X_i(t)\mathrm{d}t.$$

subject to

$$\int \psi_1(t)^2\mathrm{d}t = 1.$$

Then, each ψ_j is computed by maximizing the variance of

$$x_{ij} = \int \psi_j(t)X_i(t)\mathrm{d}t.$$

subject to

$$\int \psi_j(t)^2\mathrm{d}t = 1$$

and to the orthogonality constraints

$$\int \psi_\ell(t)\psi_j(t)\mathrm{d}t = 0,\ 1 \leq \ell \leq j-1.$$

Functional PCA can be characterized in terms of the eigenanalysis of covariance operators. If $(\lambda_j)_{j\geq 1}$ denotes the eigenvalues and $(\psi_j)_{j\geq 1}$ the eigenfunctions of the operator $\mathcal{C}$ defined by $\mathcal{C}(f)(s) = \int_0^1 C(s,t)f(t)\mathrm{d}t$, where $C(s,t) = \mathrm{cov}(X(s), X(t))$, then

$$X_i(t) = \mathbb{E}[X(t)] + \sum_{j=1}^{+\infty} x_{ij}\psi_j(t),$$

where the x_j are uncorrelated centered random variables with variance $\mathbb{E}[x_{ij}^2] = \lambda_j$. There are similarities with the context of Example 2.2.2, but here the kernel depends on X. In practice, the decomposition can easily be computed with discrete

matrix operations, replacing $C(s,t)$ by the covariance matrix of the X_i. This basis has some nice properties. In particular, considering a fixed number of coefficients, it minimizes among all orthogonal bases the average squared L^2 distance between the original curve and its linear representation (see, e.g., the book of Ghanem and Spanos [95]). For more details on functional PCA, we refer the reader to Ramsay and Silverman [159].

Observe that since the functional PCA basis is a stochastic basis and the Best-Entropy basis algorithm also uses the data, rigorously the sample should be divided into two subsamples, one to build the basis, and the other for clustering.

2.4. Experimental results and analysis

We evaluated the performance of the clustering method with projection, using the various bases described in the previous section, for two different kinds of curves. First, a simulated example where the right clusters are known is discussed, to illustrate the efficiency of the method. Then, we focus on our industrial problem and cluster output curves of a "black box" computer code.

Observe that, although the curves considered in Section 2.2 and Section 2.3 were true functions, in practice, we have to deal with curves sampled on a finite number of discretization points. Therefore, a preprocessing step based on spline interpolation is necessary.

2.4.1. Synthetic control chart time series

Control chart time series are used for monitoring process environments, to achieve appropriate control and to produce high quality products. Different types of series can be encountered, but only one, a kind of white noise, indicates a normal working. All the other types of series must be detected, because they correspond to abnormal behavior of the process.

The data set contains a few hundreds to a few thousands curves generated by the process described in Alcock and Manolopoulos [5], discretized on 128 time points. There are six types of curves: normal, cyclic, increasing trend, decreasing trend, upward shift and downward shift, which are represented in Figure 2.8. The equations which generated the data are indicated below.

(A) *Normal pattern*: $y(t) = m + rs$ where $m = 30$, $s = 2$ and $r \sim \mathcal{U}(-3, 3)$.

(B) *Cyclic pattern*: $y(t) = m + rs + a \sin \frac{2\pi t}{T}$ where $a, T \sim \mathcal{U}(10, 15)$.

(C) *Increasing shift*: $y(t) = m + rs + gt$ with $g \sim \mathcal{U}(0.2, 0.5)$.

(D) *Decreasing shift*: $y(t) = m + rs - gt$.

(E) *Upward shift*: $y(t) = m + rs + hx$ where $x \sim \mathcal{U}(7.5, 20)$, $h = \mathbf{1}_{[t_0,D]}$, $t_0 \sim \mathcal{U}\left(\frac{D}{3}, \frac{2D}{3}\right)$, and D is the number of discretization points.

(F) *Downward shift*: $y(t) = m + rs - hx$.

The two main advantages using this synthetic data set is that we can simulate as many curves as we wish and we know the right clusters.

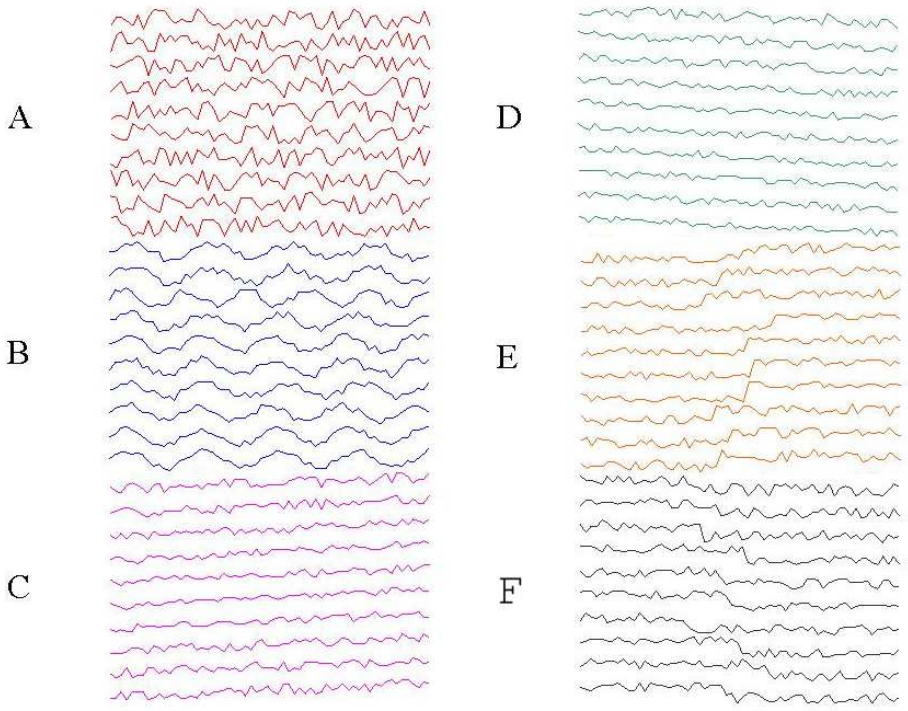


FIGURE 2.8.: 10 example curves for each of the 6 types of control chart.

The k-means algorithm on the projected coefficients has been run for all four bases (Best-Entropy, Haar, functional PCA and Fourier basis), with varying sample size n and projection dimension d. Since the result of a k-means algorithm may depend on the choice of the initial centers, the algorithm is restarted 100 times. The maximum number of iterations per run is set to 500. This program tries to globally minimize the projected empirical distortion $W_{d,n}(\mathbf{c})$. To evaluate its performance, we compute an approximation $W(d, n)$ of the distortion $W_\infty(\hat{\mathbf{c}}_{d,n})$ using a set of 18000 sample curves. This is possible in this simulated example, since we can generate as many curves as we want. The distortion $W(d, n)$ is computed for d varying from 2 to 50 and n ranging from 100 to 3100. We restricted ourselves to the case $d \leq 50$, since there are only 128 discretization points. Moreover, as

pointed out earlier, d must not be too large for computational complexity reasons. Indeed, a projection dimension $d = 50$ is already high for a practical use.

Figures 2.9 and 2.10 show the contours plots corresponding to the evolution of $W(d, n)$ as a function of d and n, for the functional PCA and the Haar basis. We remark that the norm of the gradient of $W(d, n)$ vanishes when d and n are close to their maximal values. Hence, as expected according to Theorem 2.2.1, $W(d, n)$ is decreasing in d and n. When d or n is too small (for instance $d = 2$ or $n = 100$), the clustering results are inaccurate. Besides, they are not stable with respect to the n observations chosen. However, for larger values of these parameters, the partitions obtained quickly become satisfactory. The choice $n \geq 300$ together with $d \geq 6$ generally provides good results and is reasonably low for many applications.

Figure 2.11 shows the curve corresponding to the evolution of $W(d, n)$ as a function of d for $n = 500$, for all four bases, whereas Figure 2.12 represents the evolution of $W(d, n)$ versus n for $d = 10$. According to Section 2, φ_d must be of the order $\sqrt{n}$. For $n = 500$, $\sqrt{n}$ is about 22. Considering that φ_d and d are approximately of the same order, a projection dimension close to 22 should thus be suitable. Indeed, we see via Figure 2.11 that $W(d, n)$ does not decrease much more after this value.

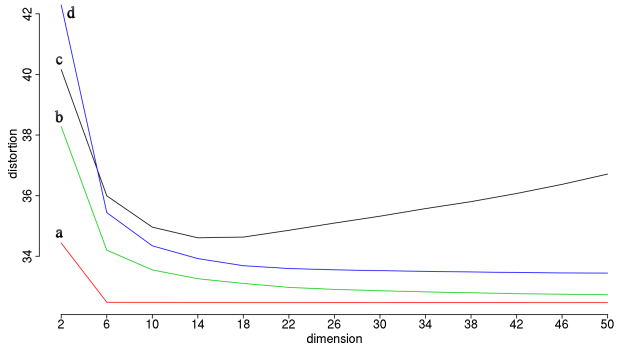


FIGURE 2.11.: Evolution of $W(d, n)$ for $n = 500$ and d ranging from 2 to 50, for (a) functional PCA basis, (b) Haar basis, (c) Fourier basis and (d) Best-Entropy basis.

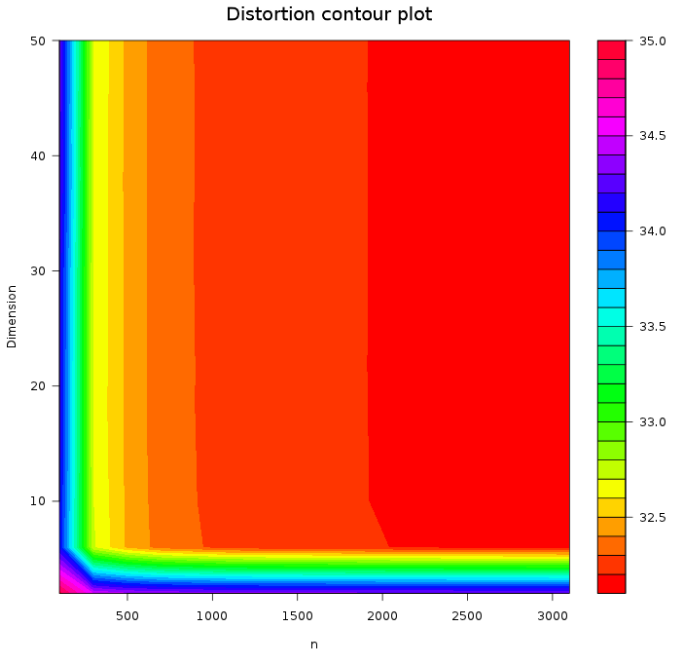


FIGURE 2.9.: Contour plot of $W(d, n)$ for the functional PCA basis.

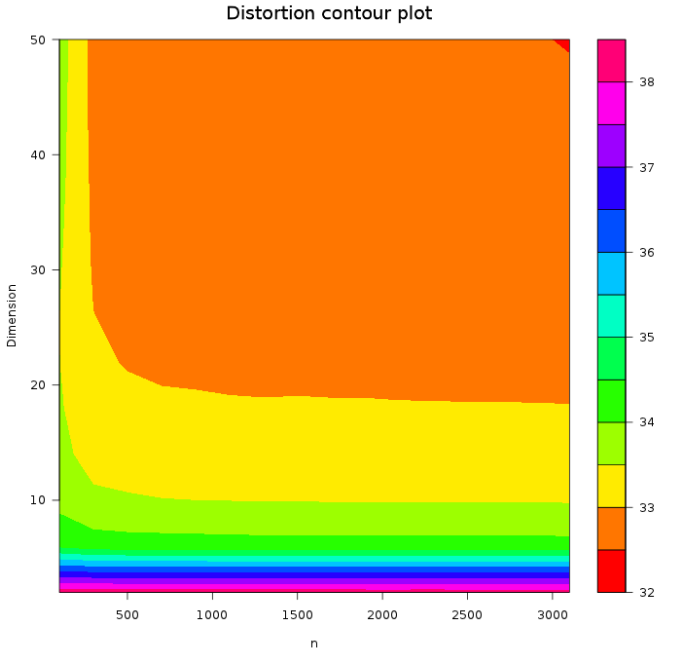


FIGURE 2.10.: Contour plot of $W(d, n)$ for the Haar basis.

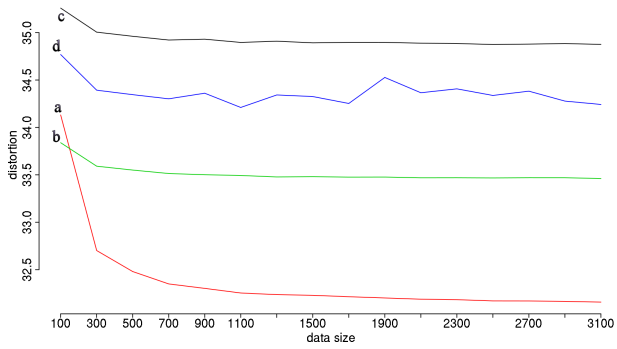


FIGURE 2.12.: Evolution of $W(d, n)$ for $d = 10$ and n ranging from 100 to 3100, for (a) functional PCA basis, (b) Haar basis, (c) Fourier basis and (d) Best-Entropy basis.

The evolution of the distortion for the Fourier basis looks quite odd: it shows a first decreasing step before increasing again. However, this increasing can be explained in the following way. The centers chosen first are wrong, but seem to give a better distortion than the "real" clustering. When the dimension grows large enough, these first wrong centers no longer represent a local minimum, and the k-means algorithm moves slowly toward the "right" clustering, although losing a bit in distortion. This interpretation is confirmed if we look at the clusters corresponding to each distortion. Furthermore, when the dimension is high relatively to the number of discretization points, some basis functions which oscillate a lot may not be sampled correctly. As a result, the coefficients estimated by approximating the inner products can become very inaccurate as d increases. For very high d, these computed coefficients confound with some noise. Consequently, data becoming more noisy without adding any information, the distortion will increase. The other bases tend to oscillate too, so that they would probably show the same behavior if d were increased above 50. Besides, the small fluctuations observed for the Best-Entropy basis indicate that this basis is not suitable for clustering of control chart time series. The functional PCA basis always gives the lowest distortion. However, the distortions obtained for the three other bases are quite similar, with a preference for the Haar basis over the Best-Entropy wavelet basis, the Fourier basis being the worst choice. As an example, Table 2.1 gives the values of $W(d, n)$ for $n = 1100$ and $d = 30$.

Basis	Fourier	Functional PCA	Haar	Best-Entropy
Distortion	35.3	32.3	32.8	33.4

TABLE 2.1.: Distortion $W(d, n)$ for $n = 1100$ and $d = 30$.

Figure 2.13 represents the 6 clusters for the Fourier basis, for $n = 300$ and $d = 10$, whereas Figure 2.14 shows them for $d = 30$. The classes obtained with the algorithm are shown in colors, and the real clusters are indicated in the caption. For relatively small values of d, the normal and cyclic patterns are merged into one big cluster, and one cluster corresponding to increasing (or decreasing) shift pattern is split in two. For large enough d, the normal and cyclic designs are well detected, and the overall clustering is correct despite some mixing increasing-upward shift or decreasing-downward shift. We also tested the algorithm for smaller values of the number k of classes. As expected, for the particular choice $k = 3$, clusters A and B are merged into one single group, and the same occurs for the cluster pairs $\{C, E\}$ and $\{D, F\}$.

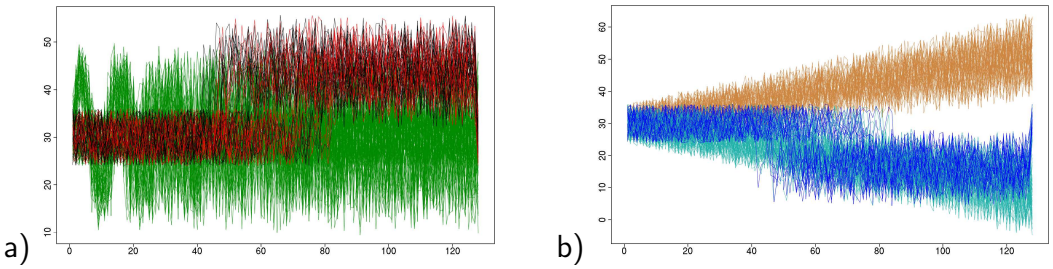


FIGURE 2.13.: (a) Clusters A and B in green, cluster E in red and black. (b) Clusters C, D and F in brown, light blue and blue respectively. (Fourier basis, $n = 300$, $d = 10$.)

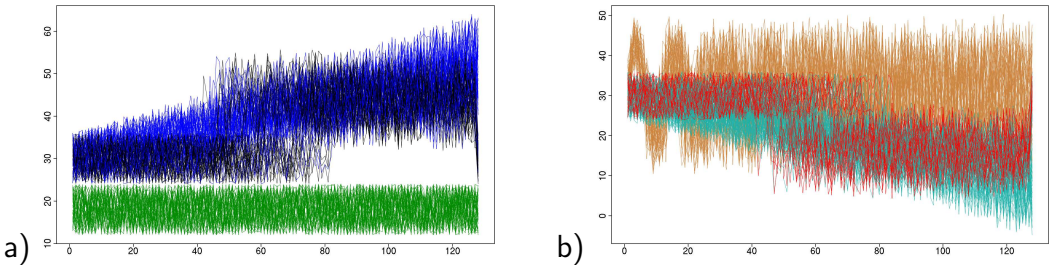


FIGURE 2.14.: (a) Clusters A, C and E in green, blue, black. (b) Clusters B, D and F in brown, light blue and red. (Fourier basis, $n = 300$, $d = 30$.)

2.4.2. Industrial code examples

Let us now turn to the industrial issue which motivated our study. As explained in the introduction, the computer codes used in nuclear engineering have become very complex and costly in CPU-time consumption. That is why we try to approximate them with a cheap function substituted to the code. In order to build a

regression model, a preliminary analysis of the different types of outputs is essential. This leads to data clustering, applied here to a computer code with functional outputs. Two different kinds of outputs are presented, the temperature evolution with a data set containing 100 curves, and the thermal exchange coefficient evolution with a data set of 200 curves.

Temperature curves The data is made of 100 CATHARE code outputs representing the evolution of the temperature in the vessel annular space (Figure 2.15). Here, the sample size is fixed to $n_0 = 100$. However, the discretization can be controlled to some extent with spline interpolation. In this case, 256 discretization points are used.

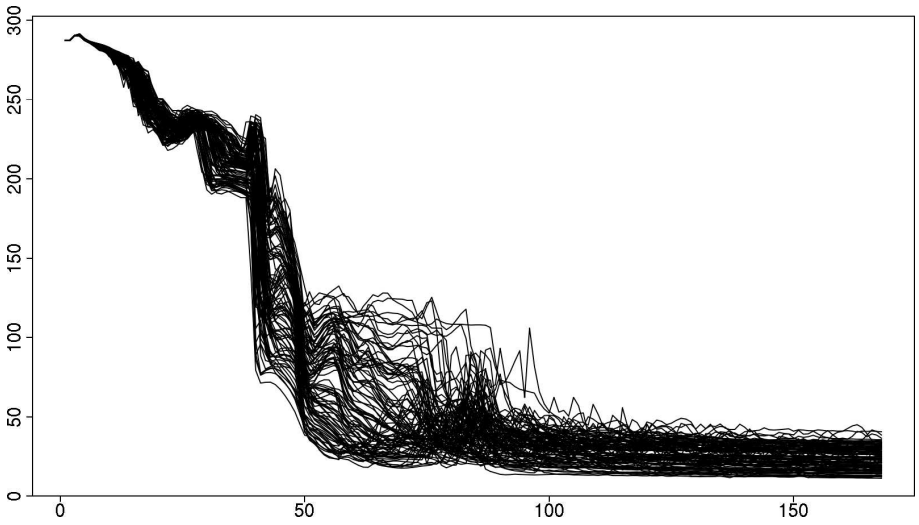


FIGURE 2.15.: The 100 temperature curves.

Observe that all curves converge in the long-time limit to the same value, corresponding to the temperature of the cold water injected. These curves have been clustered, for physical reasons pertaining to nuclear engineering, in two groups. More precisely, there is a critical set of physical parameters beyond which the thermal shock is more violent and the temperature changes more rapidly (see Auder, De Crecy, Iooss, and Marquès [13] for more details). The algorithm on the projected coefficients has been run for the Best-Entropy, Haar, functional PCA and Fourier bases, with varying dimension, with the same settings as before. Since we consider real-life data, it is not possible to compute an approximation of $W_\infty(\hat{\mathbf{c}}_{d,n})$ as in the simulated example. Hence, the distortion $W(d, n_0)$ is simply the output $W_{d,n_0}(\hat{\mathbf{c}}_{d,n_0})$ of the clustering algorithm, with fixed $n_0 = 100$. This distortion is

computed for d varying from 2 to 50. Figure 2.16 shows the curve corresponding to the evolution of $W(d, n)$ as a function of d. As expected, it is decreasing in d.

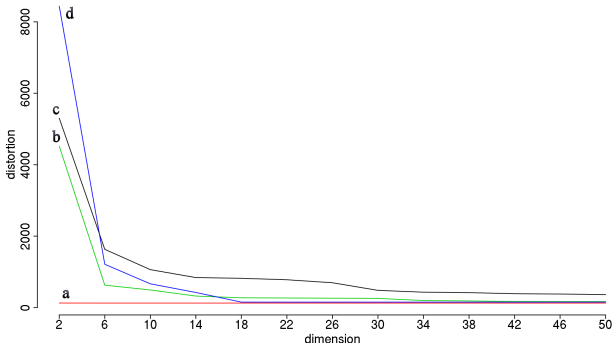


FIGURE 2.16.: Evolution of $W(d, n_0)$ for the 100 temperature curves, d ranging from 2 to 50, for (a) functional PCA basis, (b) Haar basis, (c) Fourier basis and (d) Best-Entropy basis.

Basis	Fourier	Functional PCA	Haar	Best-Entropy
Distortion	480.8	125.3	254.9	151.5

TABLE 2.2.: Distortion values for $d = 30$.

We note that until $d = 16$, the Haar basis provides lower distortion, but for larger values of d, the Best-Entropy basis is better. As before, the Fourier basis is the worst and the functional PCA basis is the best. This can also be checked from Table 2.2, which presents the distortion obtained for each basis. Although the functional PCA basis gives the best result in terms of distortion, we see that using any of the other three bases is not that bad. Indeed, the same partitioning is found every time (Figure 2.17).

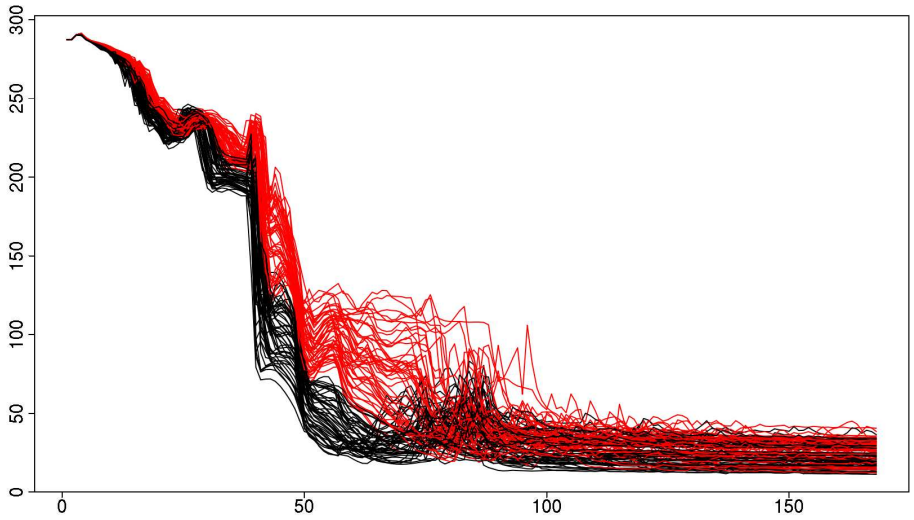


FIGURE 2.17.: The temperature curves divided in two groups.

Finally, Figure 2.18 shows the two centers representing the classes obtained for $d = 30$ with the Fourier basis, functional PCA basis, Haar basis and Best-Entropy basis. The two curves obtained with the functional PCA basis characterize with an especially good accuracy the shape of the data items in the corresponding clusters.

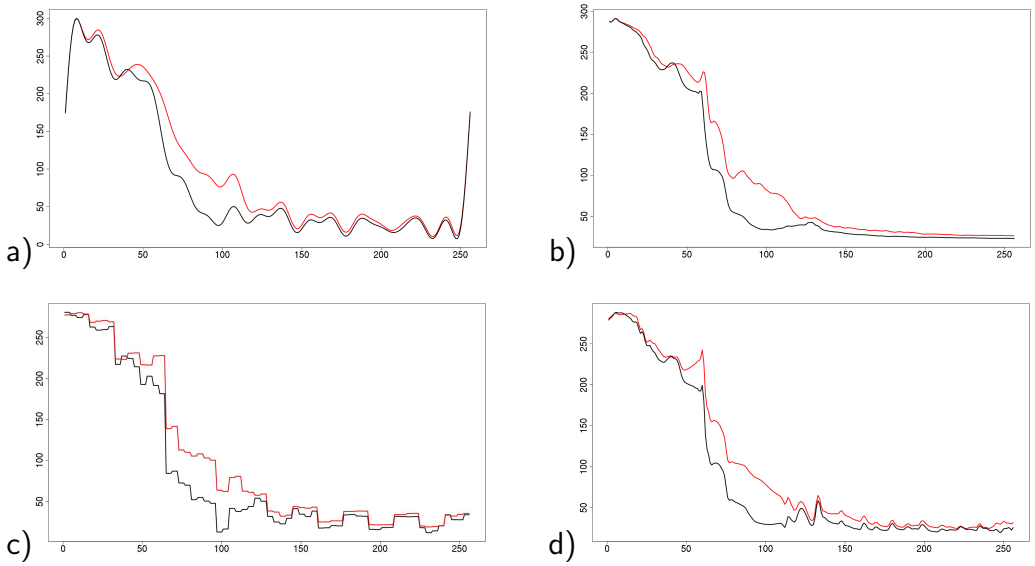


FIGURE 2.18.: The two centers for $d = 30$ for (a) Fourier, (b) functional PCA, (c) Haar and (d) Best-Entropy basis.

Thermal exchange coefficient curves Figure 2.19 shows all 200 CATHARE code outputs. Here, the number of discretization points is set to 1024. The data has been partitioned in three groups. As for the temperature curves, $W(d, n_0)$ is computed for d varying from 2 to 50 ($n_0 = 200$).

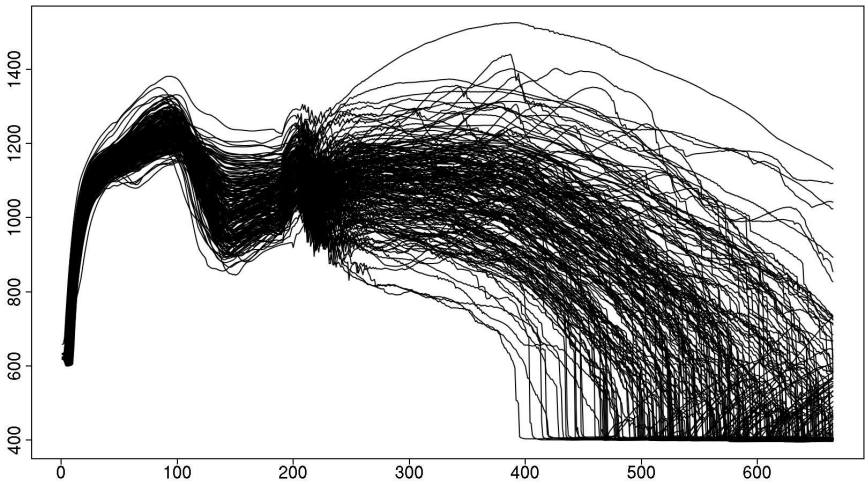


FIGURE 2.19.: The 200 thermal exchange coefficient curves.

We can see via Figure 2.20 that $W(d, n_0)$ is decreasing in d again. The functional PCA basis is still the best choice, with a fast convergence (stabilization from $d = 10$). Interestingly, the Fourier basis shows smaller distortion values than the two wavelets basis in this case, as Table 2.3 indicates.

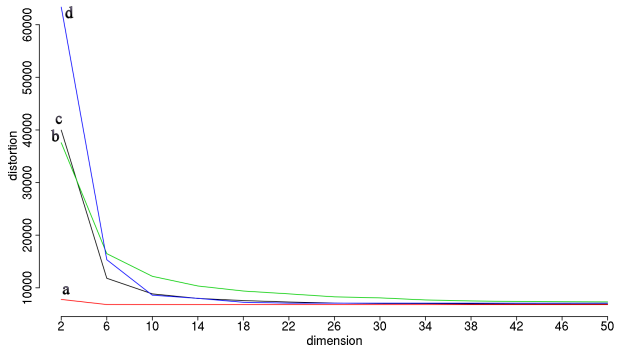


FIGURE 2.20.: Evolution of $W(d, n_0)$ for the 200 thermal exchange coefficient curves, d ranging from 2 to 50 for (a) functional PCA basis, (b) Haar basis, (c) Fourier basis and (d) Best-Entropy basis.

Basis	$d = 6$	$d = 10$	$d = 18$	$d = 26$	$d = 34$	$d = 42$	$d = 50$
Fourier	11810	8843.8	7570.0	7097.0	7013.2	6893.3	6881.2
Functional PCA	6815.8	6802.0	6801.4	6801.2	6801.1	6801.1	6801.0
Haar	16491	12185	9385.2	8279.3	7688.0	7390.6	7313.9
Best-Entropy	15338	8596.2	7244.5	7082.1	7082.1	7082.1	6801.0

TABLE 2.3.: Distortion values for the thermal exchange coefficient curves.

All the partitions obtained are very similar. A typical example is given by Figure 2.21. However, as for the temperature curves, it is interesting to look at the curves selected as centers. Figure 2.22 shows the three centers obtained with the Fourier basis, functional PCA basis, Haar basis and Best-Entropy basis, for $d = 30$.

2.5. Conclusion

These clusters allow to build accurate models for the industrial application. The partitioning method presented in this article has been integrated in our metamodel written in **R**. More specifically, given an array of n input vectors corresponding to n output curves, the purpose is to learn a function $\phi : z \mapsto x$ mapping an input vector to a continuous curve. In order to improve the accuracy of this task, we begin with a clustering step and then look for a regression model in each cluster separately. The metamodel lets the user choose between several clustering techniques, either assuming some clusters shapes (like this projected k-means) or trying to discover them in data (like the ascendant hierarchical clustering). The latter are attractive as they do not make assumptions about the results, but they generally need a relatively good sampling of the data. Consequently, the k-means-like techniques are useful in many of our industrial applications, where only a few samples are available. Moreover, these methods provide easily interpretable clusters. In each cluster, after a dimension reduction step, which can either be achieved through the decomposition on an orthonormal basis (linear), or any manifold learning algorithm (nonlinear, with the assumption that the outputs lie on a functional manifold), a statistical learning method is applied to predict representation within this cluster. The mainly used method at this stage is the Projection Pursuit Regression algorithm (see Friedman, Jacobson, and Stuetzle [86]). Finally, a simple k-nearest neighbors classifier gives the most probable cluster for a new input, the corresponding regression function is applied, and the curve can be reconstruct from its predicted representation.

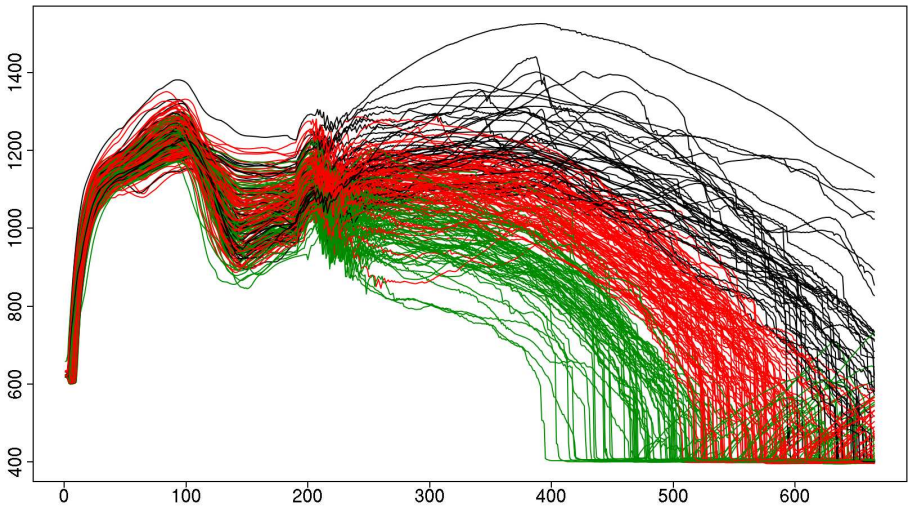


FIGURE 2.21.: Three clusters obtained with $d = 14$, functional PCA basis.

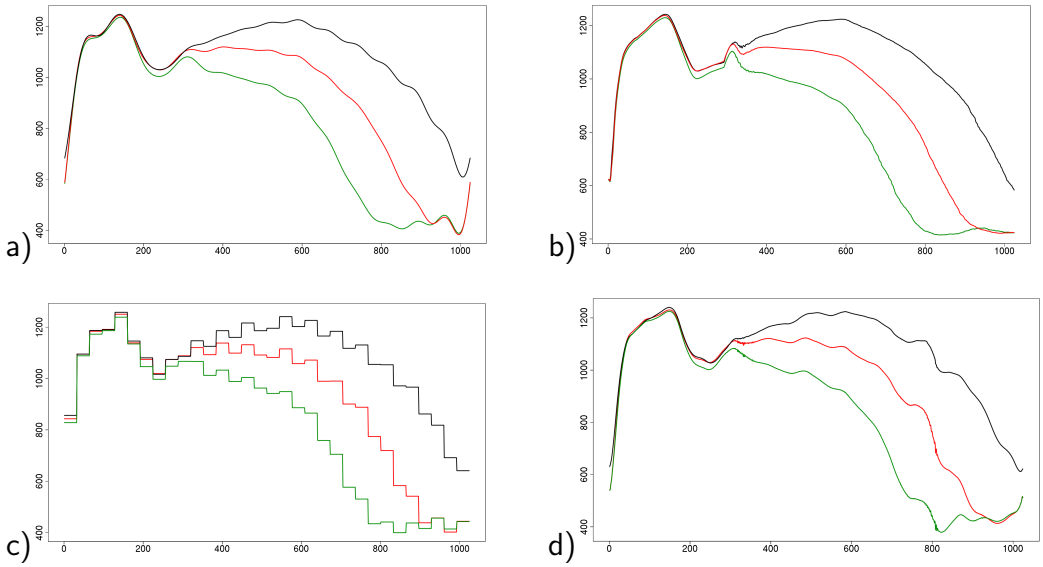


FIGURE 2.22.: The three centers with (a) Fourier, (b) functional PCA, (c) Haar and (d) Best-Entropy basis, for $d = 30$.

For the moment, our metamodel with the clustering method presented have successfully been used on two different scenarios involving the CATHARE code (minor or major break, for each we get temperature, pressure and thermal exchange coefficient curves).

Another research track could consist in considering other types of distances between curves. Distances involving derivatives might be hard to estimate on the thermal exchange coefficient dataset, because several curves are varying rapidly over short period of time, contrasting for instance with the Tecator dataset (`http://lib.stat.cmu.edu/datasets/tecator`), on which such distances proved successful (Ferraty and Vieu [84, Chapter 8], Rossi and Villa [163]). However, further investigations are needed to know if a smoothing step before clustering based on m-order derivatives would lead to improved results. As the "true" classes are unknown, such a procedure can only be validated within a cross validation framework involving the full metamodel. Experiments with the L^1 distance or some mixed distances related to functions shapes (Heckman and Zamar [106]) could also be studied in future research.

2.6. Proofs

2.6.1. Proof of Lemma 2.2.1

If we define the remainder R_d by $R_d(\mathbf{x}) = \mathbf{x} - \Pi_d(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{S}$, then for $(\mathbf{x}, \mathbf{y}) \in \mathcal{S}^2$,

$$\begin{aligned}
\|R_d(\mathbf{x}-\mathbf{y})\|^2 &\leq 2\|R_d(\mathbf{x})\|^2 + 2\|R_d(\mathbf{y})\|^2 \\
&= 2\sum_{j=d}^{+\infty} x_j^2 + 2\sum_{j=d}^{+\infty} y_j^2 \\
&= 2\sum_{j=d}^{+\infty} \frac{\varphi_j x_j^2}{\varphi_j} + 2\sum_{j=d}^{+\infty} \frac{\varphi_j y_j^2}{\varphi_j} \\
&\qquad (\varphi_j > 0 \text{ for all } j \geq d, \text{ since } d > j_0) \\
&\leq 2\sum_{j=d}^{+\infty} \frac{\varphi_j x_j^2}{\varphi_d} + 2\sum_{j=d}^{+\infty} \frac{\varphi_j y_j^2}{\varphi_d} \\
&\leq \frac{4R^2}{\varphi_d}.
\end{aligned}$$

Thus, for $\mathbf{c} \in \mathcal{S}^k$,

$$\begin{aligned}
W_\infty(\mathbf{c}) - W_d(\mathbf{c}) &= \mathbb{E}\Big[\min_{\ell=1,\dots,k} \|X - c_\ell\|^2 - \min_{\ell=1,\dots,k} \|\Pi_d(X) - \Pi_d(c_\ell)\|^2\Big] \\
&= \mathbb{E}\Big[\min_{\ell=1,\dots,k} \|\Pi_d(X) + R_d(X) - \Pi_d(c_\ell) - R_d(c_\ell)\|^2 \\
&\quad - \min_{\ell=1,\dots,k} \|\Pi_d(X) - \Pi_d(c_\ell)\|^2\Big] \\
&= \mathbb{E}\Big[\min_{\ell=1,\dots,k} \Big(\|\Pi_d(X) - \Pi_d(c_\ell)\|^2 + \|R_d(X) - R_d(c_\ell)\|^2\Big) \\
&\quad - \min_{\ell=1,\dots,k} \|\Pi_d(X) - \Pi_d(c_\ell)\|^2\Big] \\
&\qquad \text{(since } \Pi_d \text{ is the orthogonal projection on } \mathbb{R}^d\text{)} \\
&\le \mathbb{E}\Big[\max_{\ell=1,\dots,k} \|R_d(X) - R_d(c_\ell)\|^2\Big] \\
&\le \frac{4R^2}{\varphi_d}.
\end{aligned}$$

Hence,

$$\sup_{\mathbf{c}\in\mathcal{S}^k} [W_\infty(\mathbf{c}) - W_d(\mathbf{c})] \le \frac{4R^2}{\varphi_d},$$

as desired.

2.6.2. Proof of Theorem 2.2.1

We have

$$W_\infty(\hat{\mathbf{c}}_{d,n}) - W_\infty^\star = W_\infty(\hat{\mathbf{c}}_{d,n}) - W_d(\hat{\mathbf{c}}_{d,n}) + W_d(\hat{\mathbf{c}}_{d,n}) - W_d^\star + W_d^\star - W_\infty^\star.$$

According to Lemma 2.2.1, on the one hand,

$$\begin{aligned}
W_\infty(\hat{\mathbf{c}}_{d,n}) - W_d(\hat{\mathbf{c}}_{d,n}) &\le \sup_{\mathbf{c}\in\mathcal{S}^k} [W_\infty(\mathbf{c}) - W_d(\mathbf{c})] \\
&\le \frac{4R^2}{\varphi_d},
\end{aligned}$$

and on the other hand,

$$\begin{aligned}
W_d^\star - W_\infty^\star &= \inf_{\mathbf{c}\in\mathcal{S}^k} W_d(\mathbf{c}) - \inf_{\mathbf{c}\in\mathcal{S}^k} W_\infty(\mathbf{c}) \\
&\le \sup_{\mathbf{c}\in\mathcal{S}^k} [W_\infty(\mathbf{c}) - W_d(\mathbf{c})] \\
&\le \frac{4R^2}{\varphi_d},
\end{aligned}$$

and the theorem is proved.

3. Choix du nombre de groupes*

Sommaire

Dans ce chapitre, nous nous intéressons, dans le contexte de clustering décrit précédemment, au problème essentiel du choix du nombre k de groupes. En effet, s'il peut arriver dans certaines situations que ce choix soit dicté par les applications, le nombre de classes n'est en général pas connu. Il s'avère donc nécessaire de l'estimer, pour éviter d'oublier des classes réellement présentes dans la structure des données ou, au contraire, de fabriquer artificiellement des groupes superflus. L'objet du chapitre est de proposer une procédure permettant de sélectionner automatiquement k au moyen d'un critère pénalisé basé sur la distorsion empirique. Notre approche repose sur la théorie de sélection de modèle par pénalisation développée par Birgé et Massart [35] et Barron, Birgé et Massart [21]. Quelques rappels de sélection de modèle sont rassemblé dans l'Annexe B.

3.1. Cadre du problème

Dans ce chapitre, nous nous plaçons dans $(\mathbb{R}^d, \|\cdot\|)$ euclidien pour la clarté de l'exposition. Soient $\mathbf{X}_1, \ldots, \mathbf{X}_n$ des vecteurs aléatoires indépendants, de même loi qu'un vecteur aléatoire $\mathbf{X}$ à valeurs dans $\mathbb{R}^d$. Supposons que

$$\mathbb{P}\{\|\mathbf{X}\| \leq R\} = 1, \tag{3.1}$$

hypothèse déjà rencontrée dans les Chapitres 1 et 2.

*Ce chapitre a donné lieu à un article repris dans l'Annexe E.

Dans le but de former à partir de $\mathbf{X}_1, \ldots, \mathbf{X}_n$ un nombre fini de groupes cohérents, nous utilisons la méthode de clustering des k-means ou version empirique de la quantification des plus proches voisins. Rappelons qu'il s'agit, pour classer $\mathbf{X}_1, \ldots, \mathbf{X}_n$ en k groupes, de minimiser en $\mathbf{c} = (c_1, \ldots, c_k) \in (\mathbb{R}^d)^k$ le critère empirique

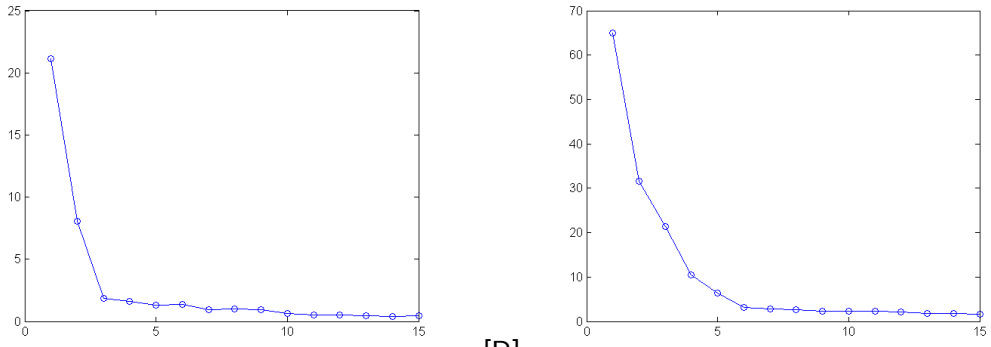
$$\frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|\mathbf{X}_i - c_j\|^2.$$

Notre objectif à présent est de sélectionner le nombre k de classes, supposé fixé dans les chapitres précédents.

Considérons la distorsion empirique

$$W(\mu_n, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{c \in \mathbf{c}} \|\mathbf{X}_i - c\|^2, \tag{3.2}$$

définie pour $\mathbf{c}$ ayant un nombre de composantes k quelconque, $1 \leq k \leq n$. Il semble naturel d'utiliser cette quantité dépendant des observations pour choisir la valeur de k appropriée. Or, la distorsion empirique est une fonction décroissante de k, comme illustré dans la Figure 3.1. En effet, plus il y a de groupes, plus chaque donnée est proche du centre du groupe auquel elle est affectée. Minimiser directement ce critère conduit à choisir k le plus grand possible, ce qui évidemment ne présente pas un grand intérêt. Il suffit de penser à la situation où $k = n$, chaque observation formant un groupe à elle seule ! Hastie, Tibshirani et Friedman [105] notent qu'évaluer la distorsion sur un ensemble test de données indépendant ne permet pas de trouver k, car si les centres sont très nombreux, ils rempliront de manière dense tout l'espace des données, et chaque observation sera très proche de l'un d'eux. Ainsi, il n'est pas possible de procéder par validation croisée comme en apprentissage supervisé. Cependant, le critère (3.2) a tendance à décroître plus fortement lorsqu'une augmentation de la valeur de k conduit à séparer deux vraies classes présentes dans la structure des données, qu'en cas d'éclatement artificiel d'un groupe.



[A] [B]

FIGURE 3.1.: Tracé de la distorsion en fonction de k pour deux exemples. [A] 3 classes. [B] 6 classes.

La distorsion empirique intervient dans la plupart des moyens proposés dans la littérature pour choisir k. On trouve une présentation de différentes méthodes dans Milligan et Cooper [147] ainsi que Hardy [101], et Gordon [97] compare les performances des cinq meilleures règles exposées dans [147]. Il faut distinguer les procédures globales, consistant à effectuer un clustering pour plusieurs valeurs de k afin de déterminer $\hat{k}$ d'après une certaine fonction de k, des procédures locales, dans lesquelles on se demande à chaque étape si un groupe doit être divisé (ou deux groupes fusionnés en un seul). Certaines méthodes globales ne sont pas définies pour $k = 1$ et ne permettent donc pas de décider s'il est effectivement pertinent de former des groupes.

Calinski et Harabasz [47] proposent de choisir la valeur de k maximisant un indice basé sur le quotient de l'inertie inter-classes $B(k)$ — somme des distances au carré entre les centres des groupes et le centre de gravité de l'ensemble des observations — et de la distorsion empirique ou inertie intra-classes $W(k)$:

$$\frac{B(k)/(k-1)}{W(k)/(n-k)}.$$

La méthode de Krzanowski et Lai [124] consiste à maximiser $W(k)k^{2/d}$, ou plus précisément la quantité équivalente

$$\left|\frac{\mathrm{DIFF}(k)}{\mathrm{DIFF}(k+1)}\right|,$$

où

$$\mathrm{DIFF}(k) = W(k-1)(k-1)^{2/d} - W(k)k^{2/d},$$

tandis que dans la règle d'Hartigan [102], un nouveau groupe est ajouté tant que la quantité

$$H(k) = \left(\frac{W(k)}{W(k+1)} - 1 \right) (n - k - 1)$$

dépasse un certain seuil. La statistique *Silhouette* de Kaufman et Rousseeuw [113] est donnée par

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))},$$

où $a(i)$ est la moyenne des distances entre $\mathbf{X}_i$ et les observations qui se trouvent dans la même classe que $\mathbf{X}_i$, et $b(i)$ est la moyenne des distances entre $\mathbf{X}_i$ et les observations du groupe le plus proche (c'est-à-dire le groupe tel que $b(i)$ soit minimal). Une observation $\mathbf{X}_i$ est bien classée lorsque $s(i)$ est grand. Kaufman et Rousseeuw [113] suggèrent donc de choisir le $\hat{k}$ maximisant la moyenne des $s(i)$ pour $i = 1, \ldots, n$. La méthode de la *Gap Statistic* de Tibshirani, Walther et Hastie [181], présentée et utilisée dans les simulations dans la Section 1.6 du Chapitre 1, compare la variation du logarithme de la distorsion empirique pour le problème de clustering considéré avec celle obtenue pour des données uniformément distribuées. Kim, Park et Park [119] développent un indice donnant $\hat{k}$ en combinant deux fonctions de monotonie opposée qui présentent un saut autour de la valeur de k optimale, alors que Sugar et James [177] proposent d'appliquer à la distorsion empirique une transformation $w \mapsto w^{-p}$ avec $p > 0$. Notons qu'il existe également des méthodes basées sur la stabilité des partitions, dans lesquelles $\hat{k}$ est sélectionné d'après les résultats obtenus en classant plusieurs sous-échantillons de l'ensemble des observations (voir par exemple Levine et Domany [128] et Ben-Hur, Elisseeff et Guyon [30]). La relation entre le nombre k et la stabilité des groupes est analysée d'un point de vue théorique dans Shamir et Tishby [168, 169], Ben-David, Luxburg et Pál [29], Ben-David, Pál et Simon [27] et Ben-David et Luxburg [28].

3.2. Le choix de k

La méthode proposée dans ce chapitre pour évaluer k est basée sur la distorsion empirique et repose sur la théorie de sélection de modèle introduite par Birgé et Massart [35] et Barron, Birgé et Massart [21]. Pour chaque k, la minimisation de la distorsion empirique

$$W(\mu_n, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{c \in \mathbf{c}} \|\mathbf{X}_i - c\|^2$$

conduit à une certaine table de codage à k composantes, et l'objectif est de sélectionner la meilleure sur toutes les valeurs possibles de k.

Plus formellement, pour tout k, $1 \leq k \leq n$, le modèle S_k est défini comme l'ensemble (dénombrable) des éléments $\mathbf{c} = (c_1, \ldots, c_k)$ de $\mathcal{Q}^k$, où $\mathcal{Q}$ est une grille suffisamment fine de $\mathbb{R}^d$. Observons qu'en pratique, un algorithme ne peut effectivement fournir que des centres appartenant à une telle grille. Pour tout k, soit $\hat{\mathbf{c}}_k$ un minimiseur de $W(\mu_n, \mathbf{c})$ sur S_k, c'est-à-dire

$$\hat{\mathbf{c}}_k \in \arg\min_{\mathbf{c} \in S_k} W(\mu_n, \mathbf{c}).$$

Afin de déterminer la meilleure table de codage dans la collection $\{\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_n\}$, nous cherchons à minimiser en k un critère de la forme

$$\mathrm{crit}(k) = W(\mu_n, \hat{\mathbf{c}}_k) + \mathrm{pen}(k),$$

où $\mathrm{pen} : \{1, \ldots, n\} \to \mathbb{R}^+$ est une fonction de pénalité dont le rôle est d'éviter le choix d'un k trop grand. Etant donné une fonction $\mathrm{pen}(k)$, $\hat{k}$ désignera un minimiseur de $\mathrm{crit}(k)$ et $\tilde{\mathbf{c}} = \hat{\mathbf{c}}_{\hat{k}}$ la table de codage associée.

La particularité de ce problème est qu'aucune cible $\mathbf{c}^\star$ pertinente ne peut être définie. Notons que contrairement au cas des modèles de mélange, nous ne disposons pas d'information sur la forme de la loi de $\mathbf{X}$. Comme mentionné plus haut, si nous posons $\mathbf{c}^\star \in \arg\min W(\mu, \mathbf{c})$, où la minimisation porte sur tous les vecteurs $\mathbf{c}$ ayant un nombre de composantes inférieur ou égal à n, le nombre de composantes de $\mathbf{c}^\star$ sera quoiqu'il arrive maximal, c'est-à-dire égal à n. Exprimé différemment, il n'y a pas de terme de variance dans cette situation, de sorte que minimiser le risque d'un estimateur revient à minimiser son biais. Dans ce contexte, la pénalité ajoutée à la distorsion empirique $W(\mu_n, \mathbf{c})$ peut s'interpréter comme un terme de « variance fictive ».

Une pénalité peut être construite en adaptant le Théorème 8.1 de Massart [141], rappelé dans l'Annexe B.2 (Théorème B.2.2). La démonstration du Théorème 3.2.1 ci-dessous, donnée dans la Section 3.4, repose sur la majoration suivante, établie par Linder [130] :

$$\mathbb{E}\left[\sup_{\mathbf{c} \in S_k} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c}))\right] \leq aR^2\sqrt{\frac{kd}{n}},$$

où $a > 0$ est une constante numérique.

Théorème 3.2.1. *On considère une famille de poids positifs $\{x_k\}_{1 \leq k \leq n}$ tels que*

$$\sum_{k=1}^{n} e^{-x_k} = \Sigma$$

et une fonction de pénalité $\text{pen} : \{1, \ldots, n\} \to \mathbb{R}^+$. *Soit* $\tilde{\mathbf{c}} = \hat{\mathbf{c}}_{\hat{k}}$. *Il existe une constante* $a > 0$ *telle que, si pour tout* $1 \leq k \leq n$,

$$\text{pen}(k) \geq R^2 \left[a\sqrt{\frac{kd}{n}} + 4\sqrt{\frac{x_k}{2n}} \right], \tag{3.3}$$

alors

$$\mathbb{E}\left[W(\mu, \tilde{\mathbf{c}})\right] \leq \inf_{1 \leq k \leq n} \left(W(\mu, S_k) + \text{pen}(k)\right) + R^2 \Sigma \sqrt{\frac{2\pi}{n}}, \tag{3.4}$$

où $W(\mu, S_k) = \inf_{\mathbf{c} \in S_k} W(\mu, \mathbf{c})$.

Avant d'illustrer en pratique la mise en œuvre de la méthode de choix de k, terminons la section en commentant ce résultat.

Le Théorème 3.2.1 propose une fonction de pénalité $\text{pen}(k)$ qui décroît vers 0 à la vitesse $1/\sqrt{n}$ et donne une majoration de l'espérance de la distorsion prise en $\tilde{\mathbf{c}}$, table de codage obtenue en minimisant le critère pénalisé $\text{crit}(k) = W(\mu_n, \hat{\mathbf{c}}_k) + \text{pen}(k)$. L'inégalité (3.4) assure que pour une pénalité assez grande, l'espérance de la distorsion prise en $\tilde{\mathbf{c}}$ reste relativement faible, proche du minimum sur k de la distorsion à un reste tendant vers 0 près.

Puisqu'un modèle est d'autant plus complexe que k est grand, le terme en $\sqrt{k/n}$ dans la pénalité traduit la complexité des modèles. Mentionnons que la démonstration du Théorème 3 de Linder [130] montre qu'il est possible de choisir $a = 96$. Cependant, cette valeur, qui résulte de majorations, n'a pas directement un intérêt pratique, d'autant plus que l'expression (3.3) fait intervenir le rayon R. En fait, le Théorème 3.2.1 donne une forme de pénalité plutôt qu'une fonction exacte.

Considérons à présent les poids $\{x_k\}_{1 \leq k \leq n}$. Plus ces poids sont importants, plus Σ est petit. Néanmoins, ils ne doivent pas être trop grands puisqu'ils interviennent également dans la pénalité. Dans le cadre de la sélection de modèle pour des modèles linéaires gaussiens, où chaque modèle S_m, $m \in \mathcal{M}$, a pour dimension D_m, une solution proposée s'il n'y a pas de redondance dans la dimension des modèles est de prendre x_m proportionel à D_m (Massart [141], Section 4.2.1). Par analogie, les poids peuvent ici être choisis proportionnels à k. Puisque dans notre situation, le cardinal de la collection de modèles est au plus n, une autre alternative consiste à poser $x_k = \ln n$ pour tout k, ce qui ne modifie pas la forme de la pénalité et implique $\Sigma = 1$. Ainsi, en première approximation, la pénalité est proportionnelle à $\sqrt{k/n}$.

L'inégalité (3.4) du Théorème 3.2.1 donne en fait une vitesse de convergence. En effet, une extension du Lemme de Pierce [155], due à Luschgy et Pagès[135], assure que, pour une grille $\mathcal{Q}$ convenable, $W(\mu, S_k)$ peut être majoré par

$$\frac{A(d)R^2}{k^{2/d}},$$

où $A(d)$ est une constante positive ne dépendant que de la dimension d. En optimisant alors les deux termes $\text{pen}(k)$ et $W(\mu, S_k)$, on obtient k de l'ordre de $n^{\frac{d}{d+4}}$. Il en résulte que l'espérance de la distorsion prise en $\tilde{\mathbf{c}}$ tend vers 0 à la vitesse $\mathcal{O}(n^{-\frac{2}{d+4}})$, comme énoncé dans le corollaire suivant.

Corollaire 3.2.1. *Si la fonction de pénalité* $\text{pen}(k)$ *est de l'ordre* $\sqrt{k/n}$ *et* $\tilde{\mathbf{c}} = \hat{\mathbf{c}}_{\hat{k}}$ *minimise le critère pénalisé*

$$\text{crit}(k) = W(\mu_n, \hat{\mathbf{c}}_k) + \text{pen}(k),$$

alors

$$\mathbb{E}\left[W(\mu, \tilde{\mathbf{c}})\right] \leq C(d, R)n^{-\frac{2}{d+4}},$$

où la constante $C(d, R)$ *ne dépend que de* d *et* R.

Remarque 3.2.1. Soit $\rho \geq 0$ (par exemple, $\rho = n^{-2}$). Si pour tout k, $\hat{\mathbf{c}}_k$ est seulement un minimiseur approché du risque empirique $W(\mu_n, \mathbf{c})$, au sens où pour tout $\mathbf{c} \in S_k$,

$$W(\mu_n, \hat{\mathbf{c}}_k) \leq W(\mu_n, \mathbf{c}) + \rho,$$

le Théorème 3.2.1 reste vrai, à condition d'ajouter ρ dans le terme de droite de l'inégalité.

Remarquons enfin que le Théorème 3.2.1 s'adapte au cas d'observations à valeurs dans un espace de Hilbert séparable $(\mathcal{H}, \|\cdot\|)$, en remplaçant le premier terme du membre de droite dans l'inégalité (3.3) par un terme en $k/\sqrt{n}$. En effet, il existe une constante $a(R)$ telle que

$$\mathbb{E}\left[\sup_{\mathbf{c}\in S_k}(W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c}))\right] \leq a(R)\frac{k}{\sqrt{n}}$$

(Biau, Devroye et Lugosi [32]). En outre, la norme peut être remplacée par d'autres mesures de distorsion, par exemple les divergence de Bregman étudiées dans le Chapitre 1, dès lors que l'on dispose d'une borne supérieure pour l'espérance de la déviation maximale

$$\mathbb{E}\left[\sup_{\mathbf{c}\in S_k}(W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c}))\right].$$

3.3. Quelques illustrations en pratique

Dans cette section, nous proposons de tester la méthode de choix de k suggérée par le Théorème 3.2.1 sur quelques exemples simulées ainsi que sur données réelles. Nous avons vu qu'en première approximation, la pénalité donnée par le théorème est de la forme $c\sqrt{k/n}$, où la constante c doit être déterminée en pratique. Pour ce faire, nous emploierons l'heuristique de pente, introduite par Birgé et Massart [37] et développée par Arlot et Massart [10]. Cette méthode permet précisément de calibrer une pénalité connue à une constante multiplicative près. Le raisonnement menant à cette heuristique est présenté dans l'Annexe B.3. En bref, elle repose sur l'existence d'une pénalité minimale en-deçà de laquelle la procédure sélectionne invariablement les modèles les plus complexes. La pénalité optimale est alors proche de deux fois cette pénalité minimale. La condition essentielle pour appliquer la méthode est que le contraste empirique considéré soit une fonction décroissante de la complexité des modèles, et la forme de la pénalité une fonction croissante. Cette hypothèse est vérifiée dans le présent contexte. Cependant, en raison du défaut de variance et de cible évoqué plus haut, notre problème ne semble pas tout à fait adapté à l'utilisation de l'heuristique de pente, qui vise à sélectionner un oracle (voir l'Annexe B). Il s'avère toutefois que cette méthode conduit dans nos exemples à des résultats plutôt satisfaisants. Deux techniques basées sur l'heuristique de pente peuvent être utilisées pour la calibration des pénalités. La méthode du saut de dimension consiste à évaluer la pénalité minimale en repérant un saut abrupt dans la complexité des modèles sélectionnés. L'autre possibilité consiste à observer que la distorsion empirique est proportionnelle à la forme de la pénalité pour les modèles de grande complexité et à estimer directement la pente de cette droite. Ces deux méthodes ont été implémentées dans MATLAB par Baudry, Maugis et Michel [25], sous la forme d'une interface baptisée CAPUSHE (CAlibrating Penalty Using Slope HEuristics).

Sur le plan de l'implémentation, l'algorithme des k-means utilisé dans l'ensemble des exemples que nous allons présenter est initialisé en prenant l'unique centre pour $k = 1$ égal à la moyenne des observations. Ensuite, à chaque incrémentation de k, un nouveau centre choisi uniformément au hasard parmi les observations est ajouté aux $k-1$ centres résultant de l'étape précédente, ce qui rend l'algorithme aléatoire. Cette procédure d'initialisation est répétée 50 fois et les centres donnant la plus faible distorsion sont conservés. Notons qu'il existe une abondante littérature sur l'intéressant problème de l'initialisation de l'algorithme des k-means et qu'il serait possible de remplacer la stratégie retenue par une autre, plus robuste (voir par exemple Pena, Lozano et Larranaga [152], Su et Dy [176], Khan et Ahmad [118], Perim, Wandekokem et Varejão [153], Al-Shboul et Myaeng [3]).

3.3.1. Données simulées

Différents nombres de groupes et dimensions

Dans une première série de simulations, nous essayons de retrouver le nombre de groupes pour 5 ensembles d'observations, qui diffèrent selon la dimension d et le nombre de groupes k sous-jacent. Sur ces exemples, le nombre de classes correct est en général trouvé. Nous observons que la méthode du saut de dimension et l'estimation directe de la pente se comportent de manière similaire.

G1 Un seul groupe. Commençons par une situation dans laquelle il n'est pas pertinent de former des groupes. Nous considérons 200 points distribués uniformément dans l'hypercube unité en dimension 10.

G2 3 groupes en dimension 2. Les observations ont été simulées suivant une loi normale bivariée de variance la matrice identité et se répartissent en 3 groupes, centrés en $(0, 0)$, $(0, 6)$ et $(5, -3)$ respectivement, contenant chacun 30 observations (voir Figure 3.2). La Figure 3.3 montre un exemple de sortie de CAPUSHE pour ce jeu de données.

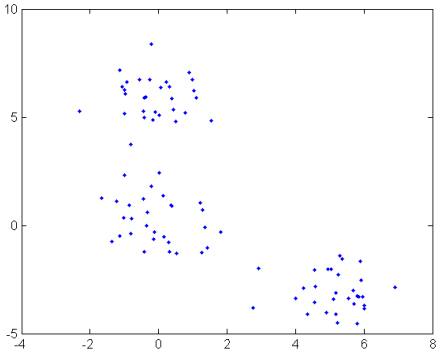
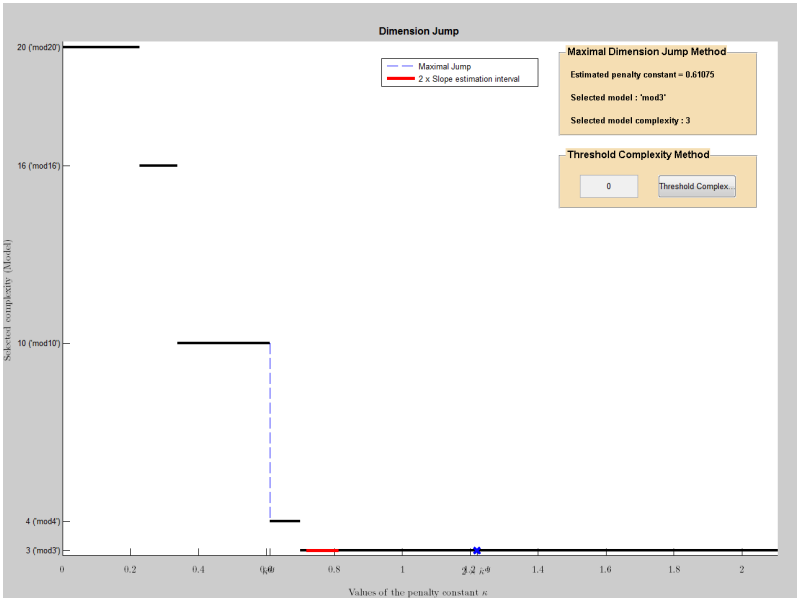


FIGURE 3.2.: Groupes de 30 observations suivant une loi normale centrée en $(0, 0)$, $(0, 6)$ et $(5, -3)$.

G3 4 groupes en dimension 3. Ensuite, nous utilisons 4 groupes d'observations suivant une loi normale en dimension 3 de variance la matrice identité. Ces groupes, visibles sur la Figure 3.4, sont centrés en $(0, 0, 0)$, $(3, 5, -1)$, $(-5, 0, 0)$, et $(6, 6, 6)$.

[A]



[B]

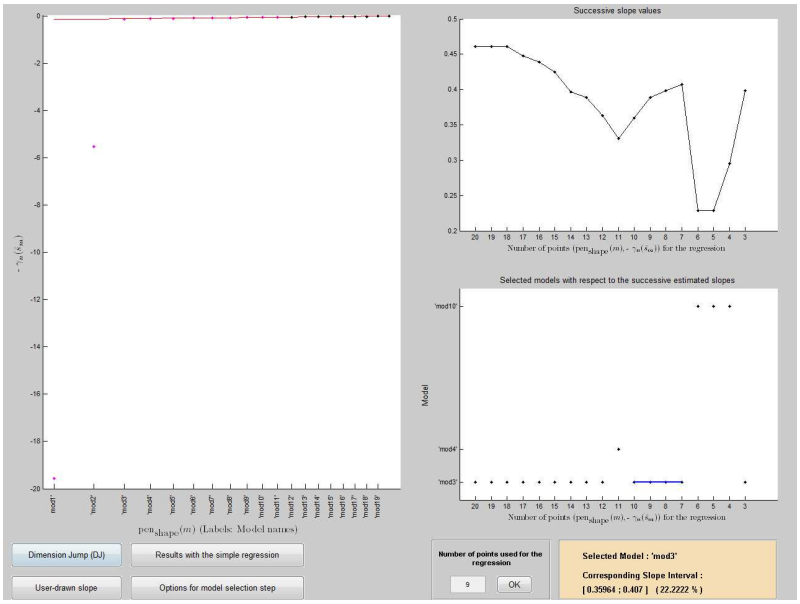


FIGURE 3.3.: Sortie de CAPUSHE (n=90, d=2, k=3). [A] Saut de dimension. Valeur $\hat{k}$ sélectionnée en fonction de la constante de pénalité. [B] Estimation de la pente. **A gauche** : Graphe du critère $-W(\mu_n, \hat{\mathbf{c}}_k)$ en fonction de $\sqrt{k/n}$. **En haut à droite** : Valeurs successives de la pente en fonction du nombre de points utilisés pour l'estimation. **En bas à droite** : Valeur $\hat{k}$ sélectionnée en fonction du nombre de points utilisés pour l'estimation de la pente.

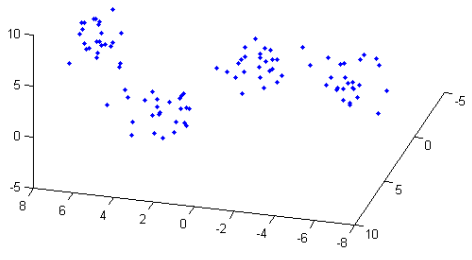


FIGURE 3.4.: Groupes de 25 observations suivant une loi normale centrée en $(0,0,0)$, $(3,5,-1)$, $(-5,0,0)$ et $(6,6,6)$.

G4 5 groupes en dimension 4. Ce jeu de données est formé de 5 groupes gaussiens en dimension 4. Les 5 groupes sont centrés respectivement en $(0,0,0,0)$, $(3,5,-1,0)$, $(-5,0,0,0)$, $(1,1,6,-2)$ et $(1,-3,-2,5)$.

G5 4 groupes en dimension 10. Enfin, nous avons simulé, toujours à l'aide de la loi normale, 4 groupes de données en dimension 10. Pour chacun des groupes, les 10 composantes du vecteur moyenne ont été tirées uniformément au hasard entre 0 et 10.

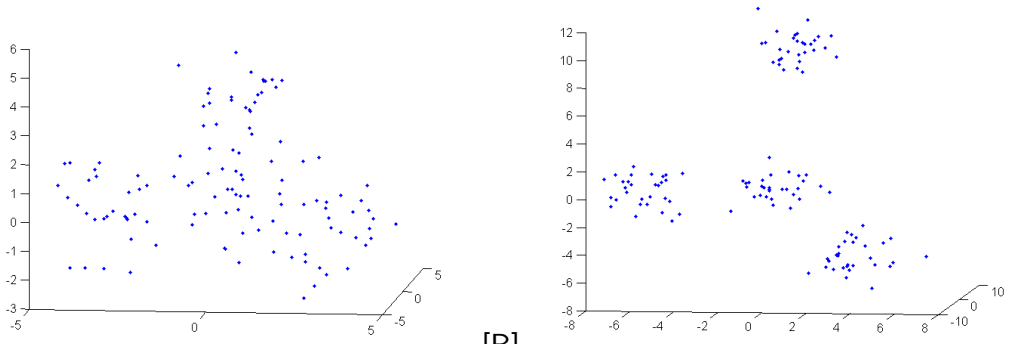
Le Tableau 3.1 montre, pour les 5 jeux de données simulés, le nombre moyen $\hat{k}$ de classes sur 20 essais obtenu par la méthode d'estimation de la pente.

Jeu de données	**G1**	**G2**	**G3**	**G4**	**G5**
Nombre de groupes $\hat{k}$	1.05	3.2	4.1	5	4.05

TABLEAU 3.1.: Nombre de classes donné par l'algorithme basé sur l'estimation directe de la pente (moyenne sur 20 essais).

Des groupes plus ou moins bien séparés

Dans ce paragraphe, nous considérons deux situations dans lesquelles les observations forment 4 groupes gaussiens en dimension 3. Dans le premier exemple (Figure 3.5 [A]), les classes sont moins bien séparés que dans le second (Figure 3.5 [B]). Nous avons comparé les résultats de l'algorithme basé sur la méthode de la pente et de la *Gap Statistic* de Tibshirani *et al.* [181] dans ces deux configurations.



[A] [B]

FIGURE 3.5.: Des groupes plus ou moins séparés : 4 groupes de 30 observations suivant une loi normale. [A] Groupes centrés en $(0,0,0)$, $(0,2,3)$, $(3,0,-1)$ et $(-3,-1,0)$. [B] Groupes centrés en $(0,0,0)$, $(0,6,10)$, $(3,0,-5)$ et $(-6,-3,0)$.

Dans le cas des groupes les moins bien séparés, centrés en $(0,0,0)$, $(0,2,3)$, $(3,0,-1)$ et $(-3,-1,0)$, la méthode de choix de k utilisant l'heuristique de pente trouve $\hat{k} = 4$ un peu plus d'une fois sur deux. Les autres valeurs produites par l'algorithme sont 3, 5, 6 et 7. Le chiffre 3 est obtenu rarement, alors que la *Gap Statistic*, quant à elle, donne presque toujours $\hat{k} = 3$. Pour 10 réalisations de ces groupes peu séparés, le Tableau 3.2 montre la moyenne de $\hat{k}$ sur 20 essais pour les méthodes de l'estimation directe de la pente et de la *Gap Statistic*. Le fait que la performance de cette dernière soit ici peu satisfaisante suggère que les groupes sont trop proches les uns de autres pour estimer précisément $\hat{k}$. Observons que les résultats des deux méthodes sont opposés, dans le sens où la *Gap Statistic* trouve un groupe de moins que la valeur attendue, tandis que la méthode de la pente a plutôt tendance à sous-pénaliser.

Pente	4.25	5.2	5.2	4.6	4.6	4.5	4.55	4.55	4.6	4.15
Gap Statistic	3	4	3	3	3	3.1	3	3	3	3.2

TABLEAU 3.2.: Nombre de classes donné par l'estimation directe de la pente et la méthode de la *Gap Statistic* pour 10 réalisations de groupes suivant une loi normale centrée en $(0,0,0)$, $(0,2,3)$, $(3,0,-1)$ et $(-3,-1,0)$ (moyenne sur 20 essais).

En revanche, lorsque les groupes sont bien séparés, l'algorithme reposant sur l'heuristique de pente semble fonctionner très bien, et les résultats sont identiques à ceux de la *Gap Statistic* : les deux méthodes donnent presque toujours le résultat attendu $\hat{k} = 4$. Pour les 4 groupes gaussiens bien séparés, centrés en

$(0,0,0)$, $(0,6,10)$, $(3,0,-5)$ et $(-6,-3,0)$, représentés dans la Figure 3.5 [B], nous obtenons par exemple les résultats de CAPUSHE visibles dans la Figure 3.6.

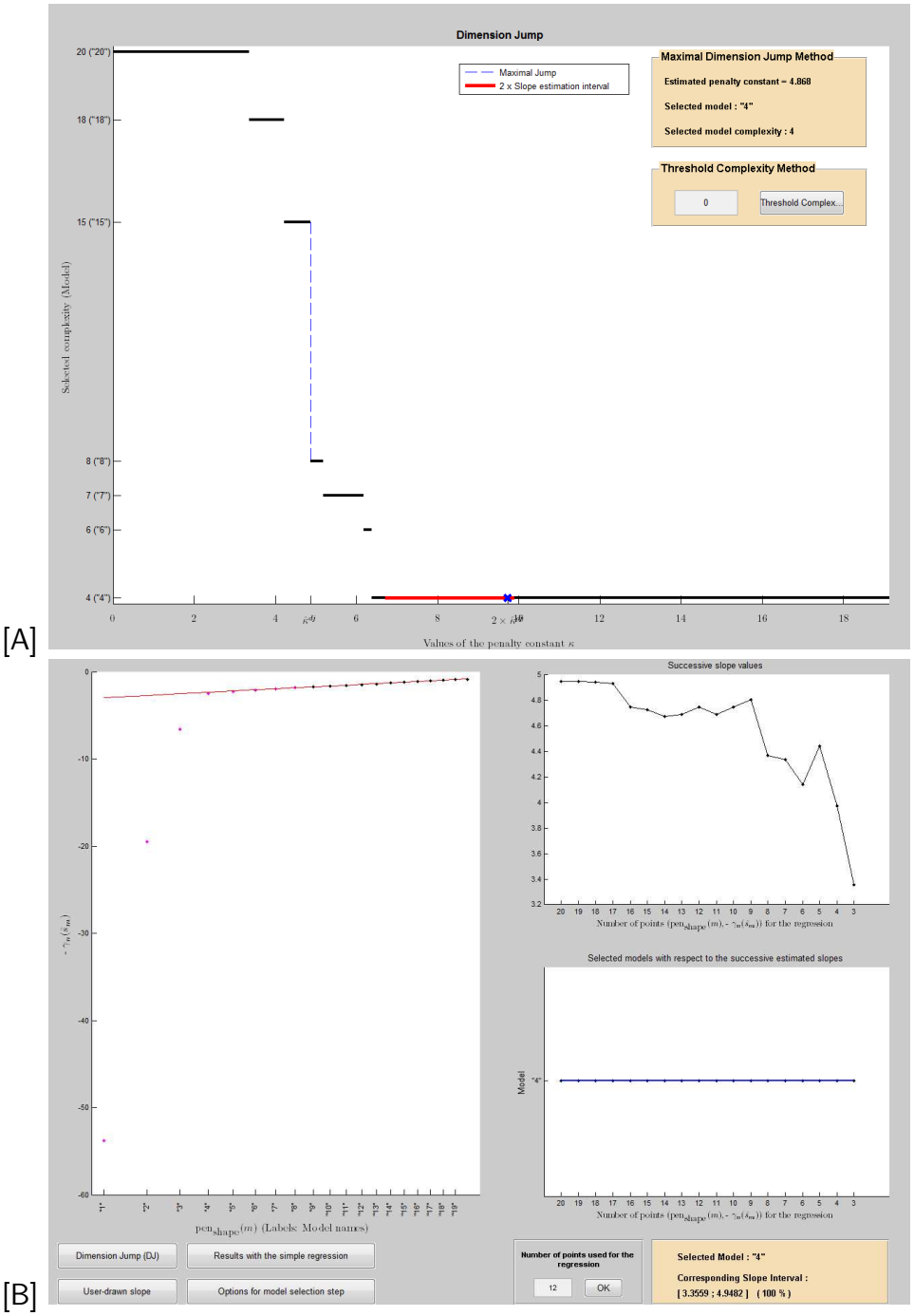


FIGURE 3.6.: Sortie de CAPUSHE. [A] Saut de dimension. [B] Estimation de la pente.

3.3.2. Données réelles

Zoo

Dans cet exemple, les observations, disponibles sur la base de données UCI Machine Learning Repository [85], rassemblent un certain nombre d'informations concernant différentes espèces animales (comme le loup, le hareng, le poulet, le crabe...). Pour chaque animal, 16 caractéristiques ont été collectées : poils, plumes, œufs, lait, qui vole, aquatique, prédateur, avec des dents, vertébré, venimeux, qui respire, queue, domestique, taille (booléens), nombre de pattes (entier appartenant à $\{0, 2, 4, 5, 6, 8\}$). Nous considérons un ensemble de 92 observations formant 5 groupes sous-jacents : mammifère, poisson, invertébré, oiseau, insecte. La plupart du temps, la sortie de l'algorithme est effectivement $\hat{k} = 5$, les autres valeurs observées étant 4 et 6. Le nombre de groupes moyen sur 20 essais était 5.05.

Dyslexie

Les données utilisées ici proviennent d'une étude sur la dyslexie effectuée dans le Laboratoire de Sciences Cognitives et Psycholinguistique (LSCP), unité de recherche mixte de l'Ecole des Hautes Etudes en Sciences Sociales (EHESS), de l'Ecole Normale Supérieure (ENS) et du Centre National de la Recherche Scientifique (CNRS), qui est située à Paris dans le Département d'Etudes Cognitives (DEC) de l'ENS. Pour mieux comprendre ce trouble affectant la fluidité et la précision d'une personne lorsqu'elle lit, parle et écrit, différentes hypothèses doivent être testées en comparant les performances d'adultes dyslexiques avec celles d'adultes témoins (`http://www.ehess.fr/lscp/persons/ramus/fr/phonodysfr.html`).

Des individus dyslexiques et non-dyslexiques, âgés de 18 à 31 ans, ont pris part à des expériences principalement basées sur des tests RAN (*Rapid Automatized Naming*, voir par exemple Denkla et Rudel [67]), qui consistent à nommer rapidement des chiffres, des couleurs et des objets, et sur l'écoute d'une liste de mots et de « non-mots ». Un « non-mot » est un ensemble de syllabes qui ne forment pas un mot existant. Par exemple, « nedo », « malani », « sonper » sont des non-mots. Pour chacune des 57 personnes considérées, nous disposons de résultats de temps de réponse, nombre d'erreurs et taux de précision des réponses.

La valeur $\hat{k}$ sélectionnée le plus souvent par l'algorithme dans ce problème est 4, la moyenne sur 20 essais étant 3.9. Nous constatons que le choix $\hat{k} = 2$, qui correspondrait aux personnes dyslexiques d'une part et au groupe contrôle d'autre part, est peu fréquent. Cependant, ce résultat peut être expliqué par la présence dans cette étude de quelques « faux positifs » et « faux négatifs ». En effet, certains individus dyslexiques ont répondu plutôt précisément et rapidement par rapport

aux autres personnes atteintes par le trouble, tandis que quelques individus du groupe témoin ont été plus lents et ont fait plus d'erreurs qu'attendu.

Ormeaux de Tasmanie

L'ormeau, également appelé oreille de mer, est un escargot de mer. La coquille de ce mollusque gastéropode marin appartenant à la famille des Haliotidae présente plusieurs couches ou « anneaux », qui peuvent être utilisées pour déterminer son âge, tâche importante dans l'étude de la biologie et de l'écologie d'une espèce. L'âge d'un ormeau est égal au nombre d'anneaux plus 1.5. Plus précisément, pour connaître l'âge d'un individu, le biologiste coupe le coquillage en deux et, après avoir utilisé une coloration, compte le nombre d'anneaux sous un microscope. Pour éviter cette activité fastidieuse, il peut être intéressant de prédire l'âge des ormeaux à partir d'autres mesures physiques, plus simples à obtenir.

Nous utilisons ici un jeu de données disponible sur UCI Machine Learning Repository [85] et provenant de la Division des Ressources Marines des Marine Research Laboratories, Taroona, Department of Primary Industry and Fisheries, Tasmanie (Nash, Sellers, Talbot, Cawthorn et Ford [148]). Les données contiennent des informations relatives à 4177 ormeaux, étiquetés « femelle », « mâle » ou « enfant ». Pour chacun d'entre eux, 7 caractéristiques ont été mesurées : longueur (plus grande dimension de la coquille), diamètre (dimension perpendiculaire à la longueur), hauteur (avec l'animal dans sa coquille), masse totale, masse sans la coquille, masse des viscères (après saignée), masse de la coquille (après séchage).

Considérant un sous-ensemble de 1303 ormeaux femelles, avec un nombre d'anneaux entre 5 et 23, nous avons essayé de retrouver d'après les mesures physiques le nombre de classes d'âge de coquillages. L'algorithme donne une valeur de $\hat{k}$ entre 16 et 22, et la moyenne sur 20 essais est 18.

3.4. Démonstration du Théorème 3.2.1

Le Théorème 3.2.1 est une adaptation du Théorème 8.1 de Massart [141] (Théorème B.2.2 dans l'Annexe B). Pour le démontrer, nous utiliserons le lemme suivant, qui est une conséquence de l'inégalité de McDiarmid [143] (voir Massart [141, Théorème 5.3]).

Lemme 3.4.1. *Si $\mathbf{X}_1, \ldots, \mathbf{X}_n$ sont des variables aléatoires indépendantes et $\mathcal{G}$ est une classe de fonctions à valeurs réelles au plus dénombrable, telle que $a \leq g \leq b$ pour toute fonction $g \in \mathcal{G}$, alors en posant $Z = \sup_{g\in\mathcal{G}} \sum_{i=1}^{n} (g(\mathbf{X}_i) - \mathbb{E}[g(\mathbf{X}_i)])$,*

on a, pour tout $\varepsilon \geq 0$,

$$\mathbb{P}\left\{Z - \mathbb{E}[Z] \geq \varepsilon\right\} \leq \exp\left(-\frac{2\varepsilon^2}{n(b-a)^2}\right).$$

Démonstration du Théorème. Remarquons que, par définition de $\tilde{\mathbf{c}}$,

$$W(\mu_n, \tilde{\mathbf{c}}) + \text{pen}(\hat{k}) \leq W(\mu_n, \mathbf{c}_k) + \text{pen}(k)$$

pour tout $k, 1 \leq k \leq n$ et $\mathbf{c}_k \in S_k$. Donc

$$W(\mu_n, \tilde{\mathbf{c}}) - W(\mu_n, \mathbf{c}_k) \leq \text{pen}(k) - \text{pen}(\hat{k}), \tag{3.5}$$

ce qui entraîne

$$W(\mu, \tilde{\mathbf{c}}) \leq W(\mu_n, \mathbf{c}_k) + W(\mu, \tilde{\mathbf{c}}) - W(\mu_n, \tilde{\mathbf{c}}) + \text{pen}(k) - \text{pen}(\hat{k}). \tag{3.6}$$

Considérons des poids positifs $\{x_k\}_{1 \leq k \leq n}$ tels que

$$\sum_{k=1}^{n} e^{-x_k} = \Sigma,$$

et fixons $z > 0$. Par le Lemme 3.4.1, on a, pour tout $k', 1 \leq k' \leq n$ et tout $\varepsilon \geq 0$,

$$\mathbb{P}\left\{\sup_{\mathbf{c} \in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c})) \geq \mathbb{E}\left[\sup_{\mathbf{c} \in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c}))\right] + \varepsilon\right\} \leq \exp\left(-\frac{n\varepsilon^2}{8R^4}\right).$$

Il en résulte que pour tout $k', 1 \leq k' \leq n$,

$$\mathbb{P}\left\{\sup_{\mathbf{c} \in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c})) \geq \mathbb{E}\left[\sup_{\mathbf{c} \in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c}))\right] + 4R^2\sqrt{\frac{x_{k'} + z}{2n}}\right\} \leq e^{-x_{k'} - z}.$$

Posant $E_{k'} = \mathbb{E}\left[\sup_{\mathbf{c} \in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c}))\right]$, on a, pour tout $k', 1 \leq k' \leq n$,

$$\sup_{\mathbf{c} \in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c})) \leq E_{k'} + 4R^2\sqrt{\frac{x_{k'} + z}{2n}},$$

sauf sur un ensemble de probabilité au plus Σe^{-z}. D'après l'inégalité (3.6), nous obtenons donc

$$\begin{aligned} W(\mu, \tilde{\mathbf{c}}) &\leq W(\mu_n, \mathbf{c}_k) + E_{\hat{k}} + 4R^2\sqrt{\frac{x_{\hat{k}} + z}{2n}} - \text{pen}(\hat{k}) + \text{pen}(k) \\ &\leq W(\mu_n, \mathbf{c}_k) + E_{\hat{k}} + 4R^2\sqrt{\frac{x_{\hat{k}}}{2n}} - \text{pen}(\hat{k}) + \text{pen}(k) + 4R^2\sqrt{\frac{z}{2n}}, \end{aligned}$$

sauf sur un ensemble de probabilité au plus Σe^{-z}. Ensuite, d'après Linder [130, Théorème 3], il existe une constante $a > 0$ telle que

$$E_{k'} \leq aR^2\sqrt{\frac{k'd}{n}}.$$

Donc, si pour tout $k', 1 \leq k' \leq n$,

$$\text{pen}(k') \geq R^2\left[a\sqrt{\frac{k'd}{n}} + 4\sqrt{\frac{x_{k'}}{n}}\right],$$

alors

$$W(\mu, \tilde{\mathbf{c}}) \leq W(\mu_n, \mathbf{c}_k) + \text{pen}(k) + 4R^2\sqrt{\frac{z}{2n}},$$

sauf sur un ensemble de probabilité au plus Σe^{-z}. Ceci se récrit

$$\mathbb{P}\left\{(4R^2)^{-1}\sqrt{2n}\Big[W(\mu, \tilde{\mathbf{c}}) - W(\mu_n, \mathbf{c}_k) + \text{pen}(k)\Big] \geq \sqrt{z}\right\} \leq \Sigma e^{-z},$$

ou encore, en posant $z = u^2$,

$$\mathbb{P}\left\{(4R^2)^{-1}\sqrt{2n}\Big[W(\mu, \tilde{\mathbf{c}}) - W(\mu_n, \mathbf{c}_k) + \text{pen}(k)\Big] \geq u\right\} \leq \Sigma e^{-u^2}.$$

Si $g_+ = \max(g, 0)$ désigne la partie positive de g, comme $\int_0^{+\infty} e^{-u^2}\mathrm{d}u = \frac{\sqrt{\pi}}{2}$,

$$\mathbb{E}\left[(W(\mu, \tilde{\mathbf{c}}) - W(\mu_n, \mathbf{c}_k) + \text{pen}(k))_+\right] \leq R^2\Sigma\sqrt{\frac{2\pi}{n}}.$$

Puisque $\mathbb{E}\left[W(\mu_n, \mathbf{c}_k)\right] = W(\mu, \mathbf{c}_k)$,

$$\mathbb{E}\left[W(\mu, \tilde{\mathbf{c}})\right] \leq W(\mu, \mathbf{c}_k) + \text{pen}(k) + R^2\Sigma\sqrt{\frac{2\pi}{n}}.$$

Comme cette inégalité est vraie pour tout k,

$$\mathbb{E}\left[W(\mu, \tilde{\mathbf{c}})\right] \leq \inf_{1\leq k\leq n}(W(\mu, S_k) + \text{pen}(k)) + R^2\Sigma\sqrt{\frac{2\pi}{n}},$$

où $W(\mu, S_k) = \inf_{\mathbf{c}\in S_k} W(\mu, \mathbf{c})$.

Deuxième partie .

Courbes principales

1. Un point sur les courbes principales

Sommaire

Les statisticiens utilisent différents moyens pour résumer de l'information et représenter les données par certaines grandeurs « simplifiées ». Parmi ces méthodes, l'analyse en composantes principales vise à déterminer les axes de variance maximale d'un nuage de points, afin de représenter les observations de manière compacte tout en rendant compte autant que possible de leur variabilité (voir par exemple Mardia, Kent et Bibby [140]). Cette technique, initiée au début du siècle dernier par les travaux de Pearson [151] et Spearman [172], puis développée par Hotelling [109], est certainement l'une des méthodes les plus célèbres et les plus fréquemment utilisées en analyse multivariée. Que ce soit dans le cadre de la réduction de dimension ou de l'extraction de caractéristiques, elle fournit souvent un important premier aperçu de la structure des données.

Cependant, dans certaines situations, plutôt que de représenter les observations à partir de droites, il est intéressant de résumer l'information de manière non

linéaire. Cette approche conduit à la notion de courbe principale, qui peut être vue comme une généralisation de la première composante principale. En bref, il s'agit de rechercher une courbe passant « au milieu » des données (voir Figure 1.1). Il existe plusieurs moyens de donner un sens mathématique à cette idée. La définition dépend par exemple de la propriété des composantes principales que l'on choisit de généraliser. La plupart du temps, une courbe principale est d'abord définie pour un vecteur aléatoire $\mathbf{X}$ de loi connue, puis adaptée à la situation pratique où l'on observe un échantillon $\mathbf{X}_1, \ldots, \mathbf{X}_n$ de $\mathbf{X}$.

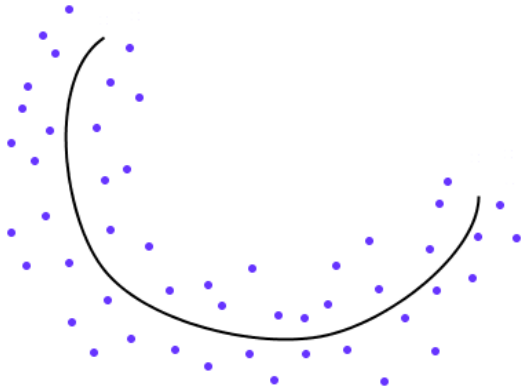


FIGURE 1.1.: Une courbe passant « au milieu » d'un nuage de points.

La définition originelle est due à Hastie et Stuetzle [104]. Elle est basée sur la propriété d'auto-consistance des composantes principales. Dans ce contexte, une courbe paramétrée $\mathbf{f} = (f_1, \ldots, f_d)$ de classe C^∞ est une courbe principale pour $\mathbf{X}$ si elle vérifie

$$\mathbf{f}(t) = \mathbb{E}[\mathbf{X} | t_{\mathbf{f}}(\mathbf{X}) = t],$$

où $t_{\mathbf{f}}(\mathbf{x})$ minimise la distance euclidienne entre $\mathbf{x}$ et $\mathbf{f}(t)$.

Etant donné la diversité des points de vue sur les courbes principales, il nous a semblé intéressant de présenter différentes définitions, en tâchant d'expliciter au mieux leurs liens. Cette synthèse est complétée par quelques propriétés des courbes principales d'ordre supérieur ainsi qu'un tour d'horizon des applications.

Dans tout le chapitre, l'espace $\mathbb{R}^d$ est muni de la norme euclidienne standard $\|\cdot\|$, $\mathbf{X}$ désigne un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, vérifiant $\mathbb{E}\|\mathbf{X}\|^2 < +\infty$, et $\mathbf{f} = (f_1, \ldots, f_d)$ est une courbe paramétrée définie sur un intervalle fermé $I = [a, b]$. Si rien n'est précisé, $\mathbf{f}$ est supposée paramétrée par l'arc. Les définitions et résultats utiles concernant les courbes paramétrées sont rassemblés dans l'Annexe C.

1.1. La première définition, basée sur la propriété d'auto-consistance

Dans cette section, nous présentons plus en détail la définition originelle des courbes principales, donnée par Hastie et Stuetzle dans les années 1980 (Hastie [103], Hastie et Stuetzle [104]).

Tout d'abord, il nous faut introduire la notion d'indice de projection.

Définition 1.1.1 (Indice de projection)**.** *Pour une courbe paramétrée* $\mathbf{f} : I \to \mathbb{R}^d$*, l'indice de projection* $t_{\mathbf{f}} : \mathbb{R}^d \to \mathbb{R}$ *est défini par*

$$t_{\mathbf{f}}(\mathbf{x}) = \sup\{t \in I, \|\mathbf{x} - \mathbf{f}(t)\| = \inf_{t'} \|\mathbf{x} - \mathbf{f}(t')\|\}. \tag{1.1}$$

Un argument de compacité permet de montrer que l'indice de projection est bien défini, c'est-à-dire qu'il existe au moins une valeur de t réalisant le minimum de $\|\mathbf{x} - \mathbf{f}(t)\|$ ([104, Proposition 5]). Pour $\mathbf{x} \in \mathbb{R}^d$, l'indice de projection $t_{\mathbf{f}}(\mathbf{x})$ est donc la plus grande valeur de t réalisant le minimum de $\|\mathbf{x}-\mathbf{f}(t)\|$, comme l'illustre la Figure 1.2.

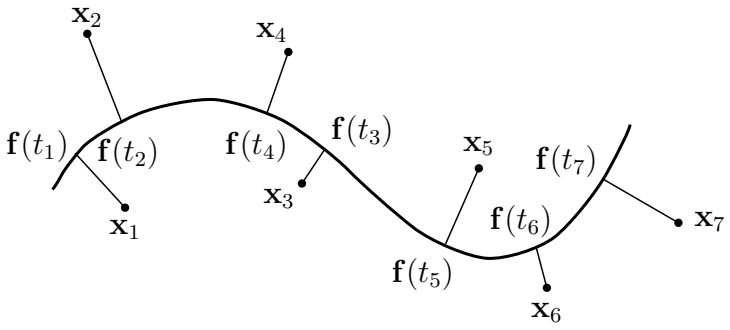


FIGURE 1.2.: Indice de projection. Pour tout i, t_i désigne $t_{\mathbf{f}}(\mathbf{x}_i)$.

A présent, il est possible de définir la propriété d'auto-consistance. Pour un exposé détaillé consacré à cette notion, le lecteur pourra consulter Tarpey et Flury [178].

Définition 1.1.2 (Courbe auto-consistante)**.** *La courbe paramétrée* $\mathbf{f} : I \to \mathbb{R}^d$ *est dite auto-consistante pour* $\mathbf{X}$ *si, pour presque tout* t*,*

$$\mathbb{E}[\mathbf{X}|t_{\mathbf{f}}(\mathbf{X}) = t] = \mathbf{f}(t).$$

Pour un ensemble de données, l'auto-consistance peut s'interpréter en disant que chaque point de la courbe $\mathbf{f}$ est la moyenne des observations qui se projettent au voisinage de ce point.

Finalement, une courbe principale est définie par Hastie et Stuetzle de la manière suivante :

Définition 1.1.3. *Une courbe paramétrée* $\mathbf{f}$ *de classe* C^∞ *est une courbe principale pour* $\mathbf{X}$ *si elle est sans point double, de longueur finie à l'intérieur de toute boule de* $\mathbb{R}^d$ *et auto-consistante.*

Remarque 1.1.1. Dire qu'une courbe paramétrée $\mathbf{f}$ est sans point double signifie que $\mathbf{f}(t_1) = \mathbf{f}(t_2)$ entraîne $t_1 = t_2$. Par ailleurs, une courbe ayant la forme d'une « spirale infinie » constitue un exemple de courbe paramétrée qui n'est pas de longueur finie dans toute boule.

Les courbes principales de Hastie et Stuetzle [104] apparaissent à plusieurs égards comme une généralisation non linéaire des composantes principales. En premier lieu, ces auteurs constatent que si une droite donnée par $\mathbf{y}(t) = \mathbf{a}t + \mathbf{b}$ est auto-consistante, il s'agit d'une composante principale. D'autre part, les composantes principales sont des points critiques de la distance au carré entre les observations et leurs projections sur des droites. Formellement, si $\mathcal{G}$ est une classe de courbes paramétrées sur l'intervalle I et $\mathbf{f}_t = \mathbf{f} + t\mathbf{g}$, où $\mathbf{g} \in \mathcal{G}$, on dit que la courbe $\mathbf{f}$ est un point critique relativement à la classe $\mathcal{G}$ si, pour toute $\mathbf{g} \in \mathcal{G}$,

$$\left.\frac{\mathrm{d}\mathbb{E}\|\mathbf{X} - \mathbf{f}_t(t_{\mathbf{f}_t}(\mathbf{X}))\|^2}{\mathrm{d}t}\right|_{t=0} = 0.$$

Une droite $\mathbf{y}(t) = \mathbf{a}t + \mathbf{b}$ est alors un point critique relativement à la classe des droites si, et seulement si, $\mathbf{a}$ est un vecteur propre de la matrice de covariance de $\mathbf{X}$ et $\mathbf{b} = 0$. On peut se demander si les courbes principales vérifient une propriété analogue, ce qui fait l'objet de la proposition suivante, prouvée dans [104].

Proposition 1.1.1. *Une courbe paramétrée* $\mathbf{f} : I \to \mathbb{R}^d$ *de classe* C^∞ *est une courbe principale si, et seulement si,* $\mathbf{f}$ *est un point critique relativement à la classe des courbes* $\mathbf{g}$ *de classe* C^∞ *sur* I *telles que* $\|\mathbf{g}\| \leq 1$ *et* $\|\mathbf{g}'\| \leq 1$.

L'existence de courbes principales répondant à la définition de Hastie et Stuetzle [104] est un problème ouvert en général. Duchamp et Stuetzle [75] ont étudié les cas particuliers de la distribution sphérique, elliptique, ainsi que d'une loi uniforme sur un rectangle ou un anneau. Remarquons qu'une loi concentrée sur une courbe régulière admet cette dernière pour courbe principale.

Signalons enfin que Hastie et Stuetzle [104] généralisent cette notion de courbe principale aux dimensions supérieures, en recherchant des surfaces vérifiant la propriété d'auto-consistance.

1.1.1. Description de l'algorithme de Hastie et Stuetzle

Comme il n'existe en général pas de moyen direct pour déterminer une courbe principale, la solution consiste à recourir à un algorithme itératif qui en fournira une approximation. En outre, en pratique, on ne connaît pas la loi de $\mathbf{X}$, mais on observe $\mathbf{X}_1, \ldots, \mathbf{X}_n$, supposées indépendantes et distribuées selon cette loi. Dans ce contexte, une courbe $\mathbf{f}$ est représentée par la ligne polygonale obtenue en joignant dans l'ordre des t_i croissants les points correspondant à n couples $(t_i, \mathbf{f}(t_i))$. La courbe étant paramétrée par l'arc, notons que les indices t_i vérifient $t_i = t_{i-1} + \|\mathbf{f}(t_i) - \mathbf{f}(t_{i-1})\|$. L'algorithme de Hastie et Stuetzle [104], qui alterne entre une étape de projection et une étape de calcul d'espérance conditionnelle, se déroule alors ainsi :

1. Initialisation.
   Soit $\mathbf{f}^{(0)}$ la droite correspondant à la première composante principale de $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

2. Etape de projection.
   Pour tout $i = 1, \ldots, n$, on pose $t_i^{(j)} = t_{\mathbf{f}^{(j)}}(\mathbf{X}_i)$.

3. Etape espérance conditionnelle.
   Il s'agit d'estimer $\mathbf{f}^{(j+1)}(t) = \mathbb{E}[\mathbf{X} | t_{\mathbf{f}^{(j)}}(\mathbf{X}) = t]$ aux points $t_1^{(j)}, \ldots, t_n^{(j)}$.

4. Condition d'arrêt.
   L'algorithme se termine lorsque la variation de $\frac{1}{n}\sum_{i=1}^n \|\mathbf{X}_i - \mathbf{f}^{(j)}(t_i^{(j)})\|^2$ tombe sous un seuil donné.

Les étapes 2 et 3 méritent d'être précisées. Pour calculer $t_i^{(j)}$ à l'étape 2, comme $\mathbf{f}^{(j)}$ est représentée par la ligne polygonale de sommets $\mathbf{f}^{(j)}(t_1^{(j-1)}), \ldots, \mathbf{f}^{(j)}(t_n^{(j-1)})$, on cherche m tel que la projection de $\mathbf{X}_i$ sur le segment $[\mathbf{f}^{(j)}(t_m^{(j-1)}), \mathbf{f}^{(j)}(t_{m+1}^{(j-1)})]$ soit minimale, et il suffit de prendre pour $t_i^{(j)}$ l'indice correspondant à cette projection (voir Figure 1.3). Avant de passer à l'étape suivante, les $t_i^{(j)}$ sont à nouveau rangés dans l'ordre croissant.

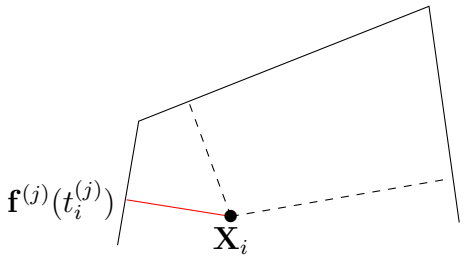


FIGURE 1.3.: Etape de projection. Pour déterminer $t_i^{(j)}$, il faut rechercher le segment minimisant la projection de $\mathbf{X}_i$ sur la ligne polygonale.

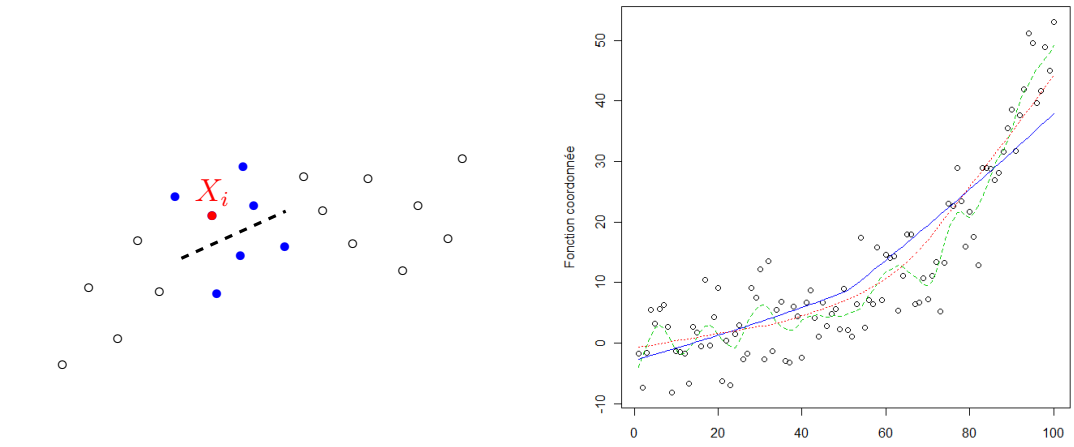
Pour l'étape 3, remarquons qu'en général, pour un point donné $(t_i^{(j)}, \mathbf{f}^{(j)}(t_i^{(j)}))$ de la courbe $\mathbf{f}^{(j)}$, une seule observation, $\mathbf{X}_i$, se projette sur ce point. La moyenne des données se projetant sur le point considéré est alors simplement $\mathbf{X}_i$. Il n'est donc pas pertinent d'utiliser directement cette moyenne pour évaluer l'espérance conditionnelle $\mathbf{f}^{(j+1)}(t) = \mathbb{E}[\mathbf{X}|t_{\mathbf{f}^{(j)}}(\mathbf{X}) = t]$, puisque la courbe principale obtenue passerait par toutes les observations. En revanche, cette étape peut être effectuée à l'aide d'une méthode de *scatterplot smoothing*. Hastie et Stuetzle [104] proposent d'appliquer à chaque fonction coordonnée un lissage LOWESS (*Locally Weighted Scatterplot Smoothing*) (Cleveland [52]) ou d'utiliser des splines cubiques de lissage (Silverman [171]).

Lissage LOWESS

Dans le premier cas, une fonction coordonnée $\mathbb{E}[X|t_{f^{(j)}}(X) = t]$ est estimée à l'aide de l'échantillon $(t_1^{(j)}, X_1), \ldots, (t_n^{(j)}, X_n)$ (avec les t_i rangés par ordre croissant) en utilisant la moyenne des observations X_m pour lesquelles t_m fait partie des « voisins » de t_i. Plus formellement, soit np le nombre de voisins à considérer, où $p \in [0, 1]$ est une certaine proportion. Chaque t_m appartenant au voisinage de t_i reçoit un poids

$$w_{im} = \left(1 - \left|\frac{t_m - t_i}{d_i}\right|^3\right)^3,$$

où d_i est la distance de t_i à son voisin le plus éloigné. A l'aide d'une régression linéaire des moindres carrés pondérés, on fait passer une droite au milieu des np observations considérées comme les voisins de X_i, comme l'illustre la Figure 1.4 [A]. L'espérance conditionnelle est alors estimée par le point de cette droite obtenu pour $t = t_i^{(j)}$.



[A] [B]

FIGURE 1.4.: [A] Droite de régression des moindre carrés pondérés (en pointillés) pour les voisins de X_i (points •). [B] Exemple de lissage LOWESS pour 3 voisinages différents.

Le choix du voisinage constitue une question importante faisant intervenir un compromis biais-variance : si le voisinage considéré est trop grand, la courbe ne reflète pas correctement la forme des données, mais s'il est trop petit, elle est trop irrégulière, voire interpole les données (voir Figure 1.4 [B]). En fait, la taille du voisinage dépend de l'objectif pratique, et dans certains cas, il existe un choix naturel de p dicté par l'application. Un moyen envisagé par Hastie et Stuetzle [104] pour choisir la proportion automatiquement est de diminuer progressivement p et de sélectionner la bonne valeur par validation croisée.

Avec des splines cubiques

La construction d'une courbe principale pour $\mathbf{X}_1, \ldots, \mathbf{X}_n$ peut également être considérée sous l'angle de la régression spline. Plus précisément, l'objectif dans ce contexte est de trouver $\mathbf{f}$, une spline cubique paramétrée sur $[0,1]$, et $t_1, \ldots, t_n \in [0,1]$, minimisant

$$\sum_{i=1}^{n} \|\mathbf{X}_i - \mathbf{f}(t_i)\|^2 + \lambda \int_0^1 \|\mathbf{f}''(t)\|^2 \mathrm{d}t, \tag{1.2}$$

où λ est un coefficient de pénalité. La paramétrisation par l'arc est ici remplacée par une paramétrisation sur l'intervalle fixé $[0,1]$ afin de pouvoir utiliser une telle pénalité de régularité sur la dérivée seconde de $\mathbf{f}$. Notons que le problème (1.2) diffère de la régression par deux aspects : non seulement $\mathbf{f}$ est une courbe paramétrée de $\mathbb{R}^d$, mais de plus les $t_1, \ldots, t_n \in [0,1]$ sont inconnus. L'algorithme associé

se déroule alors comme suit. Une courbe $\mathbf{f}^{(j)}$ étant donnée, l'étape de projection consiste à minimiser la quantité $\sum_{i=1}^{n} \|\mathbf{X}_i - \mathbf{f}^{(j)}(t_i^{(j)})\|^2$ en les $t_i^{(j)}$. Les éléments résultants sont ensuite renormalisés pour appartenir à $[0,1]$. Connaissant $t_1^{(j)}, \ldots, t_n^{(j)}$, $\mathbf{f}^{(j+1)} = (f_1^{(j+1)}, \ldots, f_d^{(j+1)})$ est obtenue en prenant pour fonctions coordonnées les splines cubiques f_m minimisant

$$\sum_{i=1}^{n}(X_{im} - f_m^{(j+1)}(t_i^{(j)}))^2 + \lambda \int (f_m^{(j+1)''}(t))^2 \mathrm{d}t, \quad m = 1, \ldots, d.$$

1.1.2. Biais d'estimation et biais de modèle

La procédure d'Hastie et Stuetzle [104] présente deux types de biais, ayant des effets opposés.

Biais d'estimation

Comme nous l'avons mentionné plus haut, le résultat de l'étape de lissage LOWESS dépend fortement du choix de la proportion p. De même, dans la méthode utilisant des splines, la courbe principale obtenue dépend du coefficient de pénalité λ. Le biais d'estimation trouve son origine dans la procédure par moyennes locales, qui a tendance à aplatir la courbe. Plus les paramètres λ ou p sont choisis grands, plus ce biais est important.

Notons que Banfield et Raftery [18], qui modélisent les contours de morceaux de banquise sur images satellite par des courbes principales fermées, développent une méthode permettant de réduire le biais dans la procédure d'estimation, tandis que Chang et Ghosh [50] remarquent qu'un meilleur résultat peut être obtenu en combinant l'algorithme de Banfield et Raftery [18] avec celui de Hastie et Stuetzle [104].

Un biais de modèle lié à la courbure

Supposons que $\mathbf{X} = (X_1, \ldots, X_d)$ s'écrive sous la forme

$$X_j = f_j(S) + \varepsilon_j, \quad j = 1, \ldots, d, \tag{1.3}$$

où S et les ε_j sont des variables aléatoires indépendantes et les ε_j sont centrées. Alors, la courbe $\mathbf{f} = (f_1, \ldots, f_d)$ n'est pas en général une courbe principale pour $\mathbf{X}$ au sens de Hastie et Stuetzle [104]. Un biais se produit lorsque la courbure est importante, comme l'illustre la Figure 1.5, qui montre des données distribuées selon une loi normale bivariée autour de la courbe en trait plein. La courbe principale, en pointillés, est décalée. En effet, si l'on considère les observations se projetant sur une zone de forte courbure, on constate qu'une plus grande partie d'entre elles est

située à l'extérieur de la courbure. On peut néanmoins noter que ce biais disparaît lorsque le produit de la variance du bruit et de la courbure tend vers 0 (Hastie et Stuetzle [104]).

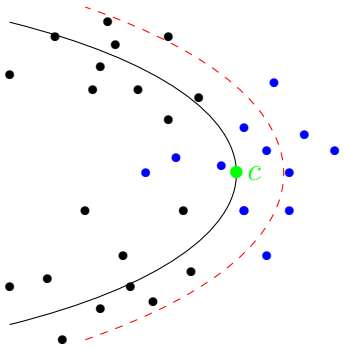


FIGURE 1.5.: Un biais lié à la courbure. Les observations se projetant au voisinage du point c étant plus nombreuses à l'extérieur de la courbure, la courbe principale ne peut être la courbe générative (en trait plein), elle est décalée (en pointillés).

Signalons enfin qu'en théorie, les paramètres intervenant dans l'algorithme pourraient être choisis de sorte que les deux types de biais se compensent exactement, mais il faudrait pour cela connaître le rayon de courbure de $\mathbf{f}$ ainsi que la variance du bruit.

1.2. Définition par un modèle de mélange

L'existence de ce biais de modèle conduit Tibshirani [180], quelques années plus tard, à introduire une nouvelle définition des courbes principales, basée sur les modèles de mélange. Il s'agit de définir directement les courbes principales à partir du modèle (1.3).

Plus précisément, supposons que le vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_d)$ admet une densité $g_{\mathbf{X}}$ et que $\mathbf{X}$ est construit en deux étapes. Tout d'abord, une variable latente S est tirée selon la densité g_S, puis $\mathbf{X}$ est généré selon une densité conditionnelle $g_{\mathbf{X}|S}$ de moyenne $\mathbf{f}(S)$, les variables $X_1, \ldots, X_d$ étant conditionnellement indépendantes sachant S.

La définition des courbes principales de Tibshirani [180] s'écrit alors comme suit.

Définition 1.2.1. *Une courbe principale de la densité $g_{\mathbf{X}}$ est un triplet $(g_S, g_{\mathbf{X}|S}, \mathbf{f})$ vérifiant les propriétés suivantes :*

1. *Les densités g_S et $g_{\mathbf{X}|S}$ sont cohérentes avec $g_{\mathbf{X}}$, c'est-à-dire, pour tout $\mathbf{x} \in \mathbb{R}^d$,*

$$g_{\mathbf{X}}(\mathbf{x}) = \int g_{\mathbf{X}|S}(\mathbf{x}|s) g_S(s) \mathrm{d}s.$$

2. *Les variables aléatoires $X_1, \ldots, X_d$ sont conditionnellement indépendantes sachant S.*
3. *La courbe $\mathbf{f}$ est une courbe paramétrée sur un intervalle fermé telle que*

$$\mathbf{f}(s) = \mathbb{E}[\mathbf{X}|S = s].$$

Cette définition ne coïncide pas en général avec celle de Hastie et Stuetzle [104]. Cependant, dans certaines situations particulières, les deux définitions conduisent au même résultat. Par exemple, c'est le cas pour les composantes principales d'une loi normale multivariée, ce qui est cohérent avec la motivation initiale de généraliser de manière non linéaire l'analyse en composantes principales.

Plaçons-nous à présent dans le cas où nous ne connaissons pas la densité $g_{\mathbf{X}}$, mais observons des vecteurs aléatoires $\mathbf{X}_1, \ldots, \mathbf{X}_n$ indépendants et de même densité $g_{\mathbf{X}}$. Plus formellement, soient $S_1, \ldots, S_n$ des variables latentes de densité g_S. Pour tout $i = 1, \ldots, n$, $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})$ est alors généré selon la densité $g_{\mathbf{X}|S_i}$. Les $X_{i1}, \ldots, X_{id}$ sont supposées conditionnellement indépendantes sachant S_i et $\mathbf{f}(s) = \mathbb{E}[\mathbf{X}|S = s]$. Tibshirani [180] propose de rechercher une courbe principale pour les observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ sous l'hypothèse que la densité conditionnelle $g_{\mathbf{X}|S}$ appartient à une famille paramétrique. Il s'agit d'estimer pour chaque s le point $\mathbf{f}(s)$ de la courbe ainsi que l'ensemble $\Sigma(s)$ des paramètres du modèle. Cette estimation peut être effectuée par maximum de vraisemblance. La log-vraisemblance observée s'écrit

$$\begin{aligned}
\mathcal{L}(\theta; \mathbf{x}) &= \ln \prod_{i=1}^{n} g_{\mathbf{X}}(\mathbf{x}_i) \\
&= \sum_{i=1}^{n} \ln g_{\mathbf{X}}(\mathbf{x}_i) \\
&= \sum_{i=1}^{n} \ln \int g_{\mathbf{X}|S}(\mathbf{x}_i|\theta(s)) g_S(s) \mathrm{d}s,
\end{aligned}$$

où $\theta = \theta(s) = (\mathbf{f}(s), \Sigma(s))$. En pratique, la maximisation de la log-vraisemblance est effectuée à l'aide d'un algorithme de type EM (*Expectation-Maximization*, Dempster, Laird et Rubin [66], Xu et Jordan [193]). Cet algorithme utilise la

log-vraisemblance des données complétées

$$\begin{aligned}\mathcal{L}(\theta; \mathbf{x}, s) = &= \ln \prod_{i=1}^{n} g_{(\mathbf{X},S)}(\mathbf{x}_i, s_i) \\ &= \sum_{i=1}^{n} \ln g_{(\mathbf{X},S)}(\mathbf{x}_i, s_i) \\ &= \sum_{i=1}^{n} \ln g_{\mathbf{X}|S}(\mathbf{x}_i|\theta(s_i)) g_S(s_i) \\ &= \sum_{i=1}^{n} \ln g_{\mathbf{X}|S}(\mathbf{x}_i|\theta(s_i)) + \sum_{i=1}^{n} \ln g_S(s_i).\end{aligned}$$

Considérons pour fixer les idées le cas où la densité conditionnelle $g_{\mathbf{X}|S}$ est gaussienne, c'est-à-dire

$$g_{\mathbf{X}|S}(\mathbf{x}_i|\theta(s_i)) = \prod_{j=1}^{d} \phi(x_{ij}|f_j(s_i), \sigma_j(s_i)),$$

où

$$\phi(x|f_j(s), \sigma_j(s)) = \frac{1}{\sigma_j(s)\sqrt{2\pi}} \exp\left(-\frac{(x - f_j(s))^2}{2\sigma_j^2(s)}\right).$$

Remarquons que la log-vraisemblance est maximale et peut valoir $+\infty$ pour une courbe $\mathbf{f}$ passant par toutes les observations. Pour éviter de sélectionner une telle courbe, une solution consiste, comme dans l'algorithme de Hastie et Stuetzle [104] avec des splines (Section 1.1.1), à mettre une pénalité sur la dérivée seconde. Plus précisément, la log-vraisemblance est pénalisée par la quantité

$$(b' - a') \sum_{j=1}^{d} \lambda_j \int_{a'}^{b'} f_j''(s)^2 \mathrm{d}s,$$

où a' et b' sont les bornes du plus petit intervalle contenant le support de la densité g_S. La log-vraisemblance des données complétées est alors donnée par

$$\sum_{i=1}^{n} \ln g_{\mathbf{X}|S}(\mathbf{x}_i|\theta(s_i)) + \sum_{i=1}^{n} \ln g_S(s_i) - (b' - a') \sum_{j=1}^{d} \lambda_j \int_{a'}^{b'} f_j''(s)^2 \mathrm{d}s. \tag{1.4}$$

1.3. Un problème de minimisation de moindres carrés

Le fait que l'existence de courbes principales au sens de Hastie et Stuetzle [104] ne soit pas assurée en général a motivé des définitions alternatives reposant sur la

minimisation d'un critère de moindres carrés pour des classes de courbes soumises à une contrainte de longueur ou de courbure. Dans cette section, nous présentons ces deux points de vue. Le critère considéré

$$\Delta(\mathbf{f}) = \mathbb{E}\left[\inf_{t\in I} \|\mathbf{X} - \mathbf{f}(t)\|^2\right] = \mathbb{E}\left[\|\mathbf{X} - \mathbf{f}(t_{\mathbf{f}}(\mathbf{X}))\|^2\right], \tag{1.5}$$

où $t_{\mathbf{f}}$ est l'indice de projection défini plus haut (1.1), est étroitement lié à la propriété d'auto-consistance caractérisant la définition de Hastie et Stuetzle [104].

1.3.1. Courbes principales de longueur bornée

La première des deux définitions de courbes principales sous forme de problème de moindres carrés est celle de Kégl, Krzyżak, Linder et Zeger [117], qui considèrent des courbes principales de longueur bornée.

Définition 1.3.1. *Une courbe* $\mathbf{f}$ *est une courbe principale de longueur (au plus)* $L > 0$ *pour* $\mathbf{X}$ *si* $\mathbf{f}$ *minimise* $\Delta(\mathbf{f})$ *sur toutes les courbes paramétrées de longueur inférieure ou égale à* L*.*

Observons qu'une courbe principale n'est ici pas supposée différentiable, comme dans le cas de Hastie et Stuetzle [104], mais seulement continue. La définition englobe ainsi les lignes polygonales. Ces dernières jouent un rôle important dans le point de vue de Kégl *et al.* [117], en particulier en ce qui concerne le côté algorithmique. La définition de la longueur d'une courbe non supposée différentiable est donnée dans l'Annexe C (Définition C.2.1).

Avec cette définition, le problème de l'existence d'une courbe principale est résolu, puisque, comme le montre la proposition suivante, la réponse est positive dans un cadre très général.

Proposition 1.3.1 (Kégl *et al.* [117])**.** *Dès que* $\mathbb{E}\|\mathbf{X}\|^2 < +\infty$*, l'existence d'une courbe principale pour* $\mathbf{X}$ *est assurée.*

Kégl *et al.* [117] remarquent que leur définition est liée à celle d'un quantificateur optimal. En effet, il s'agit dans les deux cas de minimiser un critère de type moindres carrés. En cherchant quelle est la relation entre la définition de Hastie et Stuetzle [104] et celle de Kégl *et al.* [117], nous sommes naturellement amenés à considérer cette analogie avec la quantification. Comme nous l'avons vu dans le Chapitre 1 de la première partie, si $k \geq 1$ est un entier, un k-quantificateur $q : \mathbb{R}^d \to \{c_1, \ldots, c_\ell\}$, $\ell \leq k$, est défini par les centres $c_1, \ldots, c_\ell$ et la partition $S_1, \ldots, S_\ell$ de $\mathbb{R}^d$ obtenue en posant $x \in S_j$ lorsque $q(x) = c_j$. Pour un ensemble de centres donné, le meilleur quantificateur (au sens de la distorsion) est celui associé

à la partition de Voronoi (Lemme 1.4.1), et à même partition, les centres optimaux sont donnés par $c_j = \mathbb{E}[\mathbf{X}|\mathbf{X} \in S_j]$, $j = 1, \ldots, \ell$ (Lemme 1.4.2). Dans le cas des courbes principales, la courbe $\mathbf{f}$ joue le rôle de la table de codage $\mathbf{c}$, et l'indice de projection celui de la partition. Nous savons qu'étant donné une courbe $\mathbf{f}$, nous pouvons calculer l'indice de projection $t_{\mathbf{f}}$ associé, défini par

$$t_{\mathbf{f}}(\mathbf{x}) = \sup\{t \in I, \|\mathbf{x} - \mathbf{f}(t)\| = \inf_{t'} \|\mathbf{x} - \mathbf{f}(t')\|\}.$$

D'autre part, pour une fonction $s : \mathbb{R}^d \to I$ donnée,

$$\mathbf{f}(t) = \mathbb{E}[\mathbf{X}|s(\mathbf{X}) = t]$$

minimise $\mathbb{E}[\|\mathbf{X} - \mathbf{y}\|^2|s(\mathbf{X}) = t]$ sur $\mathbb{R}^d$. En effet, si $t \in I$ et $\mathbf{y} \in \mathbb{R}^d$,

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{X} - \mathbf{y}\|^2|s(\mathbf{X}) = t] \\
&= \mathbb{E}[\|\mathbf{X} - \mathbf{f}(t) + \mathbf{f}(t) - \mathbf{y}\|^2|s(\mathbf{X}) = t] \\
&= \mathbb{E}[\|\mathbf{X} - \mathbf{f}(t)\|^2|s(\mathbf{X}) = t] + \mathbb{E}[\|\mathbf{f}(t) - \mathbf{y}\|^2|s(\mathbf{X}) = t],
\end{aligned}$$

car $\mathbb{E}[\langle \mathbf{X} - \mathbf{f}(t), \mathbf{f}(t) - \mathbf{y}\rangle|s(\mathbf{X}) = t] = 0$. Ainsi,

$$\mathbb{E}[\|\mathbf{X} - \mathbf{f}(t)\|^2|s(\mathbf{X}) = t] \leq \mathbb{E}[\|\mathbf{X} - \mathbf{y}\|^2|s(\mathbf{X}) = t],$$

avec égalité si, et seulement si, $\mathbf{y} = \mathbf{f}(t)$. Dans cette analogie, la définition de courbe principale de Hastie et Stuetzle [104] correspondrait en quantification à une définition implicite d'un quantificateur optimal

$$c_j = \mathbb{E}[\mathbf{X}|\mathbf{X} \in S_j(\mathbf{c})],\ j = 1, \ldots, \ell,$$

où la partition $S_1, \ldots, S_\ell$ n'est pas fixée, mais dépend elle-même de $c_1, \ldots, c_\ell$, tout comme $t_{\mathbf{f}}$ dépend de $\mathbf{f}$.

En pratique, la loi du vecteur aléatoire $\mathbf{X}$ est inconnue et nous disposons d'observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ supposées indépendantes et de même loi que $\mathbf{X}$. Le critère $\Delta(\mathbf{f})$ est alors remplacé par sa version empirique

$$\Delta_n(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n} \inf_{t \in I} \|\mathbf{X}_i - \mathbf{f}(t)\|^2 = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - \mathbf{f}(t_{\mathbf{f}}(\mathbf{X}_i))\|^2. \tag{1.6}$$

Dans le Chapitre 1 de la première partie, le risque d'un quantificateur empirique optimal est comparé avec le risque optimal (Section 1.5). Dans le présent contexte, la qualité d'une courbe principale obtenue en minimisant le critère empirique $\Delta_n(\mathbf{f})$ peut être évaluée de manière semblable. Considérant une ligne polygonale $\hat{\mathbf{f}}_{k,n}$ à k segments et de longueur au plus L, minimisant $\Delta_n(\mathbf{f})$, Kégl *et al.* [117] s'intéressent ainsi à la convergence du critère $\Delta(\mathbf{f})$ pris en $\hat{\mathbf{f}}_{k,n}$ vers le minimum de $\Delta(\mathbf{f})$ sur toutes les courbes paramétrées de longueur inférieure ou égale à L. Sous certaines hypothèses, ces auteurs obtiennent une vitesse de convergence en $n^{-1/3}$.

Théorème 1.3.1 (Kégl *et al.* [117]). *Supposons que* $\mathbb{P}\{\mathbf{X} \in \mathcal{C}\} = 1$, *où* $\mathcal{C}$ *est un convexe fermé borné de* $\mathbb{R}^d$. *Soit* $\mathcal{F}_L$ *l'ensemble des courbes paramétrées de longueur au plus* L, *dont l'image est incluse dans* $\mathcal{C}$. *Si* k *est proportionnel à* $n^{1/3}$ *et* $\hat{\mathbf{f}}_{k,n}$ *désigne une ligne brisée à* k *segments de longueur au plus* L *minimisant le critère* $\Delta_n(\mathbf{f})$, *alors*

$$\Delta(\hat{\mathbf{f}}_{k,n}) - \min_{\mathbf{f}\in\mathcal{F}_L} \Delta(\mathbf{f}) = \mathcal{O}(n^{-1/3}).$$

D'un point de vue pratique, Kégl *et al.* [117] proposent un algorithme itératif baptisé *Polygonal Line Algorithm* qui fournit une ligne brisée, approximation de courbe principale. L'algorithme est initialisé au moyen du plus petit segment correspondant à la première composante principale qui contienne toutes les projections des données. A chaque itération de l'algorithme, un sommet et donc un segment est ajouté à la ligne polygonale courante, puis les positions des sommets sont recalculées dans une boucle interne basée sur une étape de projection et une étape d'optimisation.

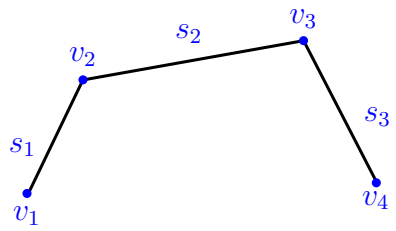


FIGURE 1.6.: Numérotation des segments et sommets pour $k = 3$.

A l'itération k, soit $\mathbf{f}_{k,n}$ une ligne polygonale de sommets $v_1, \ldots, v_{k+1}$ et de segments $s_1, \ldots, s_k$ comme dans l'exemple de la Figure 1.6.

Au cours de l'**étape de projection**, les données $\mathbf{X}_1, \ldots, \mathbf{X}_n$ sont réparties dans au plus $2k+1$ ensembles disjoints. Plus précisément, pour $\mathbf{x} \in \mathbb{R}^d$, notons $\Delta(\mathbf{x}, \mathbf{f}) = \inf_{t\in I} \|\mathbf{x} - \mathbf{f}(t)\|^2$, $\Delta(\mathbf{x}, s_j) = \inf_{\mathbf{y}\in s_j} \|\mathbf{x} - \mathbf{y}\|^2$, pour $j = 1, \ldots, k$, et $\Delta(\mathbf{x}, v_j) = \|\mathbf{x} - v_j\|^2$, pour $j = 1, \ldots, k+1$. Soient

$$V_j = \{\mathbf{x} \in \mathbb{R}^d, \Delta(\mathbf{x}, v_j) = \Delta(\mathbf{x}, \mathbf{f}), \Delta(\mathbf{x}, v_j) < \Delta(\mathbf{x}, v_\ell), \ell = 1, \ldots, j-1\},$$

pour $j = 1, \ldots, k+1$, et

$$S_j = \left\{\mathbf{x} \in \mathbb{R}^d \setminus \bigcup_{j=1}^{k+1} V_j, \Delta(\mathbf{x}, s_j) = \Delta(\mathbf{x}, \mathbf{f}), \Delta(\mathbf{x}, s_j) < \Delta(\mathbf{x}, s_\ell), \ell = 1, \ldots, j-1\right\},$$

pour $j = 1, \ldots, k$. Ces ensembles forment une partition de $\mathbb{R}^d$, comme l'illustre la Figure 1.7. En fonction du segment ou du sommet sur lequel se trouve sa projection sur la courbe, chaque observation est alors affectée à l'un des ensembles $V_1, \ldots, V_{k+1}, S_1, \ldots, S_k$.

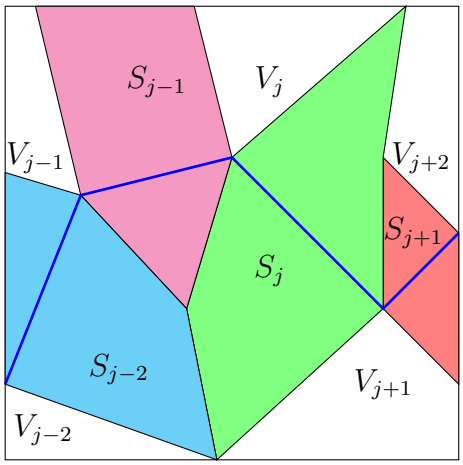


FIGURE 1.7.: Les ensembles $V_1, \ldots, V_{k+1}$ et $S_1, \ldots, S_k$ formant une partition de $\mathbb{R}^2$.

L'**étape d'optimisation** consiste ensuite à ajuster les positions des sommets. La nouvelle localisation du sommet v_j est déterminée en minimisant le critère

$$\frac{1}{n}\Bigg[\sum_{\mathbf{X}_i \in S_{j-1}} \Delta(\mathbf{X}_i, s_{j-1}) + \sum_{\mathbf{X}_i \in V_j} \Delta(\mathbf{X}_i, v_j) + \sum_{\mathbf{X}_i \in S_j} \Delta(\mathbf{X}_i, s_j)\Bigg], \tag{1.7}$$

auquel s'ajoute une pénalité sur les angles. Le sens du critère (1.7) est le suivant : pour optimiser la position du sommet v_j, on minimise une version locale de $\Delta_n(\mathbf{f})$, dans laquelle interviennent seulement les données qui se projettent sur ce sommet v_j ou sur l'un des deux segments contigus. Le terme de pénalité utilisé est proportionnel à la somme des cosinus des angles correspondant aux sommets v_{j-1}, v_j et v_{j+1}. Eviter les angles trop aigus permet en effet de contrôler la longueur de la courbe.

Avant l'ajout d'un nouveau sommet, les positions des sommets $v_1, \ldots, v_{k+1}$ sont ainsi recalculées de manière cyclique, avec un critère d'arrêt reposant sur la variation du critère Δ_n. Un **ajout de sommet** se fait en prenant le milieu du segment sur lequel se projettent le plus grand nombre de données. En cas d'égalité, le segment le plus long est choisi (voir Figure 1.8).

L'algorithme, résumé dans la Figure 1.9, se termine lorsque k dépasse un certain seuil, construit heuristiquement et réglé expérimentalement sur plusieurs essais. Cette condition d'arrêt fait intervenir le nombre d'observations n et le critère empirique Δ_n. Observons qu'il serait intéressant de disposer d'un choix automatique du nombre adéquat de segments avec une garantie théorique. Ce problème sera étudié dans le Chapitre 2.

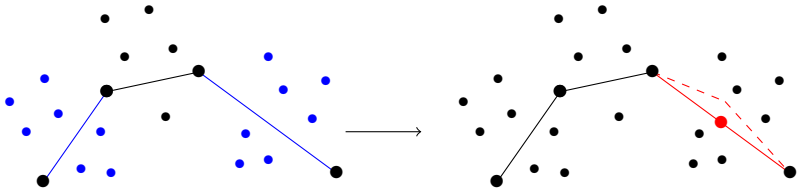


FIGURE 1.8.: Ajout d'un sommet. On cherche les segments sur lesquels se projettent le plus grand nombre de données : le nouveau sommet, qui sera ensuite ajusté, est le milieu du plus long d'entre eux.

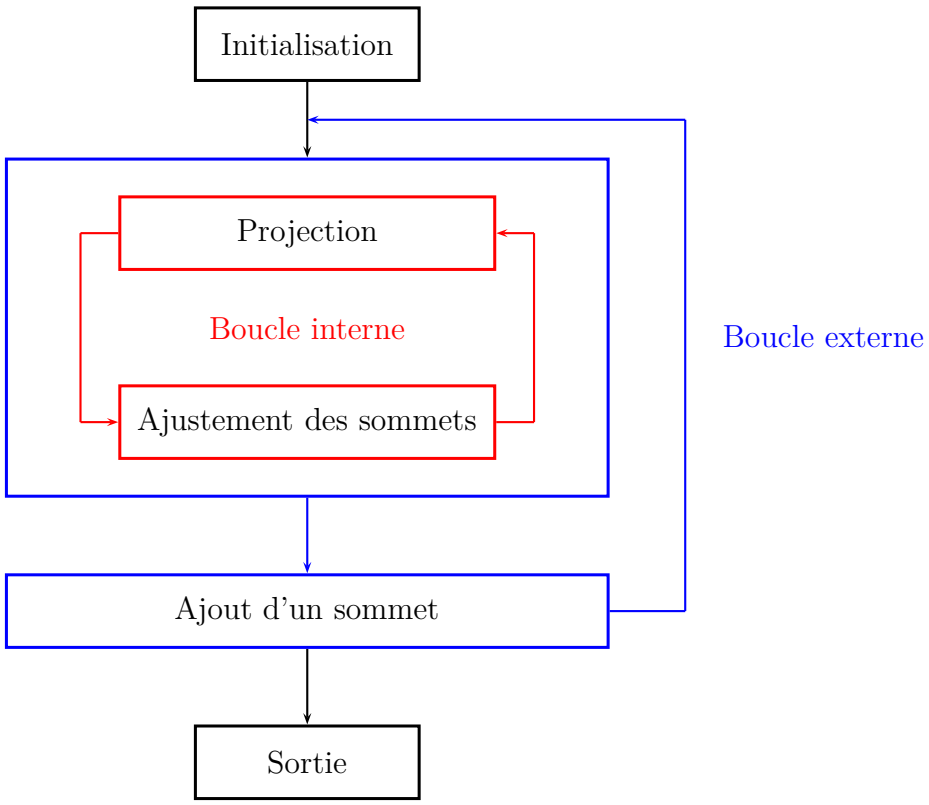


FIGURE 1.9.: Schéma résumant le *Polygonal Line Algorithm*.

1.3.2. Courbes principales de courbure intégrale bornée

Dans sa thèse [115], Kégl fait remarquer qu'il serait intéressant de remplacer la contrainte sur la longueur par une contrainte sur la courbure, qui serait en rapport plus direct avec le *Polygonal Line Algorithm*. C'est précisément ce que proposent Sandilya et Kulkarni [166], qui observent en outre que le point de vue de Kégl *et al.* [117] n'englobe pas tout à fait l'analyse en composantes principales classique, dans la mesure où une droite n'est pas de longueur finie. En conséquence, ces auteurs suggèrent d'utiliser les mêmes critères (1.5) et (1.6) que Kégl *et al.* [117], mais en plaçant la contrainte sur la courbure, et plus précisément sur la notion de courbure intégrale.

Donnons pour commencer la définition de la courbure intégrale dans le cas d'une ligne polygonale. Comme illustré dans la Figure 1.10, cette quantité est alors décrite simplement, de manière géométrique.

Définition 1.3.2 (Courbure intégrale d'une courbe linéaire par morceaux)**.** *Soit* $\mathbf{f}$ *une ligne polygonale, de sommets* $v_1, \dots, v_{k+1}$*. On appelle* $\overrightarrow{s_j}$ *le vecteur* $\overrightarrow{v_j v_{j+1}}$*, et* ϕ_{j+1} *l'angle (non-orienté) de vecteurs* $(\overrightarrow{s_j}, \overrightarrow{s_{j+1}})$*. La courbure intégrale de* $\mathbf{f}$ *est alors donnée par*

$$\mathscr{K}(\mathbf{f}) = \sum_{j=2}^{k} \phi_j.$$

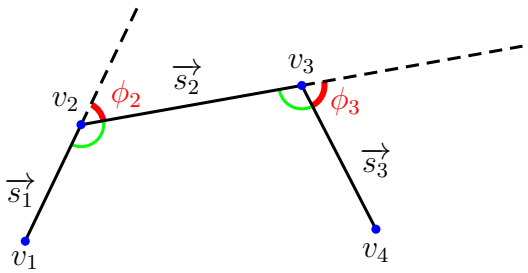


FIGURE 1.10.: En notant $\overrightarrow{s_j}$ le vecteur $\overrightarrow{v_j v_{j+1}}$ pour tout $j = 1, \dots, k$, la courbure intégrale de la ligne polygonale de sommets les v_j est la somme des angles $\phi_{j+1} = (\overrightarrow{s_j}, \overrightarrow{s_{j+1}})$.

La définition pour une courbe générale s'obtient en approximant cette courbe par des fonctions linéaires par morceaux (voir également l'Annexe C).

Définition 1.3.3 (Courbure intégrale)**.** *La courbure intégrale d'une courbe paramétrée* $\mathbf{f}$ *sur* $[\alpha, \beta]$ *est définie par*

$$\mathscr{K}(\mathbf{f}, \alpha, \beta) = \sup_{p} \sup_{\mathbf{g}} \mathscr{K}(\mathbf{g}),$$

où $\mathbf{g}$ *est une courbe linéaire par morceaux de sommets* $\mathbf{f}(t_0), \ldots, \mathbf{f}(t_p)$*, avec* $\alpha = t_0 < t_1 < \cdots < t_{p-1} < t_p = \beta$*. La courbure intégrale de la courbe* $\mathbf{f}$ *entière est alors donnée par*

$$\mathscr{K}(\mathbf{f}) = \sup_{\alpha,\beta} \mathscr{K}(\mathbf{f}, \alpha, \beta).$$

Signalons que pour une courbe régulière, la courbure intégrale mesure l'intégrale de la courbure par rapport à l'abscisse curviligne. Pour davantage de détails sur la notion de courbure intégrale, on pourra par exemple se reporter à Alexandrov et Reshetnyak [7].

Sandilya et Kulkarni [166] considèrent des courbes de courbure intégrale inférieure ou égale à $K \geq 0$. Cependant, imposer une courbure intégrale finie ne suffit pas à garantir l'existence d'une courbe principale minimisant le critère $\Delta(\mathbf{f})$. Pour illustrer ce problème, les auteurs donnent l'exemple de $\mathbf{X} = (X_1, X_2) \in \mathbb{R}^2$, où X_1 est une variable gaussienne et X_2 une variable de Bernoulli. Les observations sont distribuées selon la loi gaussienne unidimensionnelle sur deux droites parallèles, tombant sur chacune des droites avec probabilité $1/2$ (Figure 1.11). Dans ce cas, la borne inférieure de $\Delta(\mathbf{f})$ sur toutes les courbes paramétrées de courbure intégrale au plus π est 0, mais aucune courbe ne réalise le minimum. Ceci vient du fait que la limite de courbes dont la courbure intégrale s'accumule à l'infini n'est plus une courbe, mais une union de courbes.

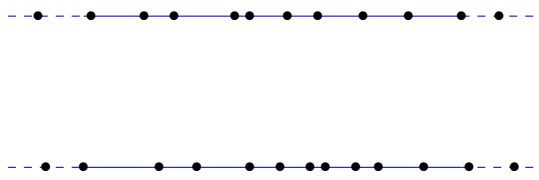


FIGURE 1.11.: Données sur deux droites parallèles, distribuées selon une loi gaussienne, le choix de la droite s'effectuant à l'aide d'une variable de Bernoulli.

Pour éviter ce problème, une contrainte additionnelle est introduite, assurant que la courbure intégrale de $\mathbf{f}$ à l'intérieur d'une boule fermée B_R de rayon R converge suffisamment vite vers la courbure intégrale totale.

Définition 1.3.4. *Une courbe* $\mathbf{f}$ *est une courbe principale de courbure intégrale (au plus)* $K \geq 0$ *pour* $\mathbf{X}$ *si* $\mathbf{f}$ *minimise* $\Delta(\mathbf{f})$ *sur toutes les courbes de la classe* $\mathcal{C}_{K,\tau}$ *définie par*

$$\mathcal{C}_{K,\tau} = \{\mathbf{f} : \mathscr{K}(\mathbf{f}) \leq K, \mathscr{K}(\mathbf{f}) - \mathscr{K}(\mathbf{f}|_{B_R}) \leq \tau(R)\},$$

où τ *est une fonction continue qui décroît vers 0.*

La classe $\mathcal{C}_{K,\tau}$ comprend les courbes $\mathbf{f}$ de courbure intégrale au plus K, telles que la différence entre la courbure intégrale totale de $\mathbf{f}$ et la courbure intégrale de la restriction de $\mathbf{f}$ à une boule tende vers 0 lorsque le rayon de la boule augmente.

Avec cette condition supplémentaire, Sandilya et Kulkarni [166] sont en mesure de démontrer l'existence des courbes principales de courbure intégrale bornée.

Proposition 1.3.2 (Sandilya et Kulkarni [166]). *Si* $\mathbb{E}\|\mathbf{X}\|^2 < +\infty$, *l'existence d'une courbe principale pour* $\mathbf{X}$ *est garantie.*

Soit $\hat{\mathbf{f}}_{k,n}$ une courbe minimisant le critère $\Delta_n(\mathbf{f})$ sur la classe des lignes polygonales à k segments appartenant à $\mathcal{C}_{k,\tau}$ et dont l'image est incluse dans la boule B_{R_k}. Ici, $(R_k)_{k\geq 1}$ désigne une suite croissante tendant vers l'infini. Le théorème suivant, qui établit une vitesse de convergence de $\Delta(\hat{\mathbf{f}}_{k,n})$ vers le risque optimal, correspond dans le présent cadre au résultat de Kégl *et al.* [117] énoncé dans le Théorème 1.3.1. L'hypothèse $\mathbb{P}\{\mathbf{X} \in \mathcal{C}\} = 1$ est remplacée par un contrôle en fonction de R de la quantité $\mathbb{E}\|\mathbf{X}\|^2$ en dehors des boules de rayon R.

Théorème 1.3.2 (Sandilya et Kulkarni [166]). *Supposons que pour tout* $R > 0$,

$$\mathbb{E}[\|\mathbf{X}\|^2 \mathbf{1}_{B_R^c}(\mathbf{X})] \leq R^{-\alpha}.$$

Si $k = n^{1/3}$ *et* $\hat{\mathbf{f}}_{n,k} \in \mathcal{C}_{k,\tau}$ *est une ligne brisée à* k *segments d'image incluse dans* B_{R_k}, *minimisant le critère empirique* $\Delta_n(\mathbf{f})$, *alors*

$$\Delta(\hat{\mathbf{f}}_{k,n}) - \min_{\mathbf{f}\in\mathcal{C}_{K,\tau}} \Delta(\mathbf{f}) = \mathcal{O}(n^{-\frac{\alpha}{6+3\alpha}}).$$

1.4. Définitions reposant sur une analyse locale

Les définitions de courbe principale auxquelles nous nous sommes intéressés jusqu'ici sont toutes liées à la propriété d'auto-consistance des composantes principales. Cette dernière intervient explicitement dans la définition originelle de Hastie et Stuetzle [104] ainsi que celle de Tibshirani [180], et nous avons vu que réaliser le minimum du critère de type moindres carrés de Kégl *et al.* [117] et Sandilya et Kulkarni [166] revient pour ainsi dire à vérifier cette propriété. La présente section est consacrée à des points de vue un peu différents, dans lesquels une courbe principale est construite à partir d'une analyse locale.

1.4.1. Courbes principales de points orientés principaux

La définition des courbes principales de points orientés, proposée par Delicado [64] et étendue dans Delicado et Huerta [65], généralise une propriété des composantes principales pour la loi normale multivariée, qui exprime que la variance

totale conditionnelle de $\mathbf{X}$, sachant que $\mathbf{X}$ appartient à un hyperplan, est minimale lorsque l'hyperplan est orthogonal à la première composante principale. Pour $\mathbf{x} \in \mathbb{R}^d$ et $\mathbf{y}$ un vecteur unitaire de $\mathbb{R}^d$, soit

$$H(\mathbf{x}, \mathbf{y}) = \{\mathbf{z} \in \mathbb{R}^d, {}^t(\mathbf{z} - \mathbf{x})\mathbf{y} = 0\},$$

où ${}^t(\cdot)$ désigne l'opérateur de transposition usuel. Ainsi défini, $H(\mathbf{x}, \mathbf{y})$ est l'hyperplan orthogonal à $\mathbf{y}$ passant par $\mathbf{x}$. Notons

$$m(\mathbf{x}, \mathbf{y}) = \mathbb{E}[\mathbf{X}|\mathbf{X} \in H(\mathbf{x}, \mathbf{y})]$$

et

$$\phi(\mathbf{x}, \mathbf{y}) = VT(\mathbf{X}|\mathbf{X} \in H(\mathbf{x}, \mathbf{y})),$$

où $VT(\mathbf{Y})$ désigne la variance totale d'un vecteur aléatoire $\mathbf{Y}$, c'est-à-dire la trace de sa matrice de covariance. Si $\mathbf{y}^\star(\mathbf{x})$ désigne l'ensemble des vecteurs unitaires qui minimisent $\phi(\mathbf{x}, \mathbf{y})$, notons $m^\star(\mathbf{x}) = m(\mathbf{x}, \mathbf{y}^\star(\mathbf{x}))$ l'ensemble des espérances conditionnelles associées. Lorsque $m^\star(\mathbf{x})$ est réduit à un singleton, son unique élément sera encore noté $m^\star(\mathbf{x})$.

Nous pouvons à présent énoncer plus précisément la propriété de la loi normale multivariée qui constitue le point de départ de la définition d'une courbe principale de points orientés.

Proposition 1.4.1. *Soit $\mathbf{X}$ un vecteur gaussien de $\mathbb{R}^d$, de moyenne m et de matrice de covariance Σ. On note $\mathbf{v}_1$ le vecteur propre unitaire associé à la plus grande valeur propre de Σ. Alors,*

- *Il existe un unique vecteur unitaire qui minimise $\phi(\mathbf{x}, \mathbf{y})$ pour tout $\mathbf{x} \in \mathbb{R}^d$, et ce vecteur est $\mathbf{v}_1$.*
- *Pour tout $\mathbf{x} \in \mathbb{R}^d$, $m^\star(\mathbf{x})$ appartient à la première composante principale $\mathbf{y}(t) = m + t\mathbf{v}_1$.*
- *Un élément $\mathbf{x}$ de $\mathbb{R}^d$ appartient à la première composante principale si, et seulement si, $\mathbf{x} = m^\star(\mathbf{x})$.*

Par analogie avec le cas gaussien, Delicado [64] introduit le concept de points orientés principaux.

Définition 1.4.1 (Points orientés principaux)**.** *L'ensemble $\Gamma(\mathbf{X})$ des points orientés principaux de $\mathbf{X}$ est l'ensemble des éléments $\mathbf{x} \in \mathbb{R}^d$ tels que $\mathbf{x} \in m^\star(\mathbf{x})$.*

Une courbe principale peut alors être définie à partir de la notion de points orientés principaux.

Définition 1.4.2. *La courbe paramétrée $\mathbf{f}$ est une courbe principale (de points orientés) pour $\mathbf{X}$ si son image est incluse dans $\Gamma(\mathbf{X})$.*

Sous certaines hypothèses, Delicado [64] démontre l'existence de points orientés principaux et des courbes principales associées. Remarquons que si $\mathbf{y}^\star(\mathbf{x})$ est réduit à un singleton, comme dans le cas de la loi normale, la définition des points orientés principaux s'écrit $m^\star(\mathbf{x}) = \mathbf{x}$ et rappelle ainsi l'auto-consistance. Nous pouvons cependant noter que l'analyse est ici locale, puisque la propriété considérée ne s'applique pas à une courbe, mais à des points de $\mathbb{R}^d$.

Dans le cas d'observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$, il faut faire appel au même type de stratégie que dans Hastie et Stuetzle [104]. En effet, le calcul des espérances et variances totales conditionnelles ne peut se faire directement, puisqu'un hyperplan contient très peu d'observations $\mathbf{X}_i$ (une seule voire souvent aucune). Pour contourner cette difficulté, Delicado [64] a recours à une projection des données accompagnée d'une pondération. Plus précisément, étant donné un hyperplan $H = H(\mathbf{x}, \mathbf{y})$, notons $\mathbf{X}_i^H$ le projeté orthogonal sur H de l'observation $\mathbf{X}_i$. On définit des poids

$$w_i = w(|{}^t(\mathbf{X}_i - \mathbf{x})\mathbf{y}|) = w(\|\mathbf{X}_i - \mathbf{X}_i^H\|)$$

où w est une fonction strictement positive et décroissante. L'espérance conditionnelle est remplacée par la moyenne des observations projetées $\mathbf{X}_i^H$ pondérée par les w_i. La variance totale conditionnelle est également définie de cette manière.

1.4.2. Composantes principales locales

Dans l'approche de Einbeck, Tutz et Evers [82], une courbe principale est calculée à partir d'un ensemble de premières composantes principales locales. Ce point de vue donne lieu à l'algorithme suivant. Soit K_h la fonction définie par

$$K_h(\mathbf{x}) = \frac{K(\mathbf{x}/h)}{h^d}, \quad \mathbf{x} \in \mathbb{R}^d,$$

où K est un noyau en dimension d et $h > 0$.

Partant d'un point $\mathbf{x} = \mathbf{x}^{(0)} \in \mathbb{R}^d$, qui est par exemple choisi au hasard parmi les observations, on calcule la moyenne empirique locale

$$\hat{m}^{\mathbf{x}} = \sum_{i=1}^{n} w_i \mathbf{X}_i,$$

où $w_i = K_h(\mathbf{X}_i - \mathbf{x}) / \sum_{\ell=1}^{n} K_h(\mathbf{X}_\ell - \mathbf{x})$. Une analyse en composantes principales est alors effectuée localement autour de $\mathbf{x}$. Soit $\hat{\Sigma}^{\mathbf{x}} = (\hat{\sigma}_{jk}^{\mathbf{x}})_{1 \le j,k \le d}$ la matrice de covariance empirique locale, définie par

$$\hat{\sigma}_{jk}^{\mathbf{x}} = \sum_{i=1}^{n} w_i (X_{ij} - \hat{m}_j^{\mathbf{x}})(X_{ik} - \hat{m}_k^{\mathbf{x}}).$$

Si $\hat{\gamma}^{\mathbf{x}}$ désigne le premier vecteur propre de $\hat{\Sigma}^{\mathbf{x}}$, la première composante principale locale autour de $\mathbf{x}$ est donnée par $\hat{\mathbf{v}}^{\mathbf{x}}(t) = \hat{m}^{\mathbf{x}} + t\hat{\gamma}^{\mathbf{x}}$. Ensuite, la valeur de $\mathbf{x}$ est actualisée en posant $\mathbf{x}^{(1)} = \hat{m}^{\mathbf{x}} + h\hat{\gamma}^{\mathbf{x}}$, et ces étapes sont itérées jusqu'à ce que $\hat{m}^{\mathbf{x}}$ reste approximativement constante, la limite des observations ayant été atteinte. Afin de visiter tout le nuage de données, la même procédure est appliquée dans la direction opposée $-\hat{\gamma}^{\mathbf{x}}$, excepté dans le cas particulier d'une courbe fermée.

Mentionnons que l'algorithme de courbes principales proposé par Verbeek, Vlassis et Kröse [185], consistant à assembler des segments pour obtenir une courbe principale polygonale, est également basé sur une analyse en composantes principales locale.

1.5. Estimation de filaments

Enfin, la problématique de l'estimation de filaments analysée par Genovese, Perone-Pacifico, Verdinelli et Wasserman [93] peut être rapprochée de celle des courbes principales. Dans ce contexte, le modèle s'écrit

$$\mathbf{X}_i = \mathbf{f}(U_i) + \varepsilon_i,\ i = 1, \ldots, n,$$

où $\mathbf{f} : [0,1] \to \mathbb{R}^2$ est une courbe régulière sans point double, $U_1, \ldots, U_n$ sont des observations indépendantes distribuées selon une loi H sur $[0,1]$, et les ε_i, modélisant le bruit, sont des vecteurs aléatoires centrés indépendants et de même loi F. Sous certaines hypothèses sur les lois F et H, l'objectif est d'estimer l'image de $\mathbf{f}$.

La stratégie de Genovese *et al.* [93] consiste à construire un ensemble $\hat{M}$ aussi petit que possible contenant la courbe $\mathbf{f}$, en utilisant la distance de Hausdorff comme fonction de perte entre $\hat{M}$ et l'image de $\mathbf{f}$. Il s'agit alors d'extraire $\mathbf{f}$ à partir de cet ensemble $\hat{M}$. L'approche repose sur deux concepts géométriques, l'axe médian et le rayon de courbure minimal global. L'axe médian d'un compact $S \in \mathbb{R}^2$ est l'ensemble des centres des boules dont l'intérieur est strictement inclus dans S, mais qui touchent S en au moins deux points. Le rayon de courbure minimal global d'une courbe est la quantité $\min_{\mathbf{x},\mathbf{y},\mathbf{z}} r(\mathbf{x},\mathbf{y},\mathbf{z})$, où $r(\mathbf{x},\mathbf{y},\mathbf{z})$ est le rayon du cercle passant par trois points distincts $\mathbf{x},\mathbf{y},\mathbf{z}$ de la courbe. Ici, le rayon de courbure minimal global de $\mathbf{f}$ est supposé strictement supérieur à R, ce qui garantit une intensité du bruit raisonnable.

Plusieurs procédés d'estimation de l'image de $\mathbf{f}$ sont envisagés. L'idée générale consiste à estimer le support de la distribution, puis la frontière du support, pour en déterminer ensuite l'axe médian. Après avoir établi une borne inférieure minimax (voir par exemple Lehmann et Casella [127, Chapitre 5]) pour ce problème,

Genovese *et al.* [93] montrent que l'une de leurs méthodes permet d'atteindre la vitesse correspondante.

1.6. Plusieurs courbes principales

Une courbe principale, dans les différentes définitions auxquelles nous nous sommes intéressés jusqu'ici, correspond dans le cas linéaire à la première composante principale. Dans l'analyse en composantes principales, on extrait les composantes principales successives, chacune expliquant une certaine proportion décroissante de la variance. De même, une distribution ou un ensemble d'observations peut donner lieu à plusieurs courbes principales. La Figure 1.12 constitue un exemple typique de configuration pour laquelle il semble opportun de rechercher plus d'une courbe principale.

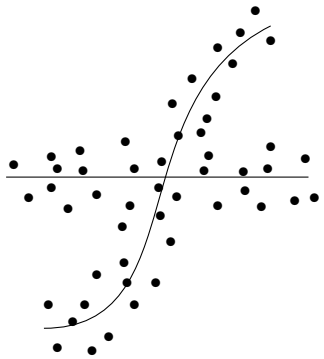


FIGURE 1.12.: Deux courbes principales pour un ensemble d'observations.

On peut trouver dans la littérature quelques mentions de cette notion de courbes principales successives.

Duchamp et Stuetzle [75], qui étudient les courbes principales (au sens de Hastie et Stuetzle [104]) dans le plan, constatent que les courbes principales multiples vérifient certaines propriétés analogues à l'orthogonalité des composantes principales.

Proposition 1.6.1. *Si* $\mathbf{f}_1$ *et* $\mathbf{f}_2$ *sont deux courbes principales pour* $\mathbf{X}$*, elles ne peuvent être séparées par un hyperplan.*

Les auteurs précisent ensuite ce résultat, en montrant que, sous certaines conditions de régularité des courbes et de convexité du support de la loi de $\mathbf{X}$, deux courbes principales s'intersectent toujours.

Dans Kégl et Krzyżak [116], pour traiter le cas des courbes principales multiples, l'algorithme de lignes polygonales de Kégl, Krzyżak, Linder et Zeger [117] est étendu à la notion de graphes principaux. Ce nouvel algorithme repose sur une nomenclature précise des types de sommets possibles. Comme le montre la Figure 1.13, on distingue, par exemple, les extrémités de la courbe, les nœuds en T ou encore les nœuds en étoile de degré 4. A chaque classe de sommets correspond une certaine pénalité, portant sur les angles ou la longueur des segments.

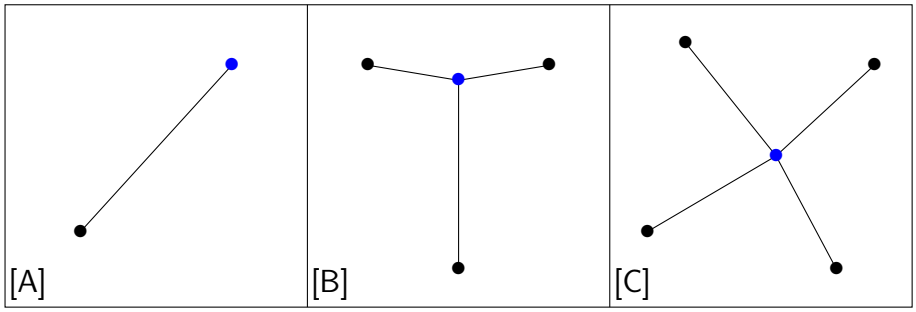


FIGURE 1.13.: Exemples de types de sommets. [A] Extrémité. [B] Nœud en T. [C] Etoile à 4 branches.

Remarquons que Delicado [64] propose une extension de sa définition à plusieurs courbes principales en utilisant une quantité généralisant la variance totale. Einbeck, Tutz et Evers indiquent dans [82] que leur algorithme de composantes principales locales permet de trouver plusieurs courbes principales par le biais de différents choix d'initialisation et suggèrent dans [81] de considérer des composantes principales locales correspondant au deuxième axe principal.

1.7. Quelques domaines d'application

Les applications de la notion de courbe principale sont nombreuses et variées. Nous proposons ici un survol qui sans se prétendre exhaustif cherche à donner un aperçu de cette grande diversité de champs d'application.

La première application, décrite dans l'article originel de Hastie et Stuetzle [104], a été mise en place au centre de l'accélérateur linéaire de Stanford, où un « collisionneur linéaire » composé de deux arcs formés d'aimants est chargé de faire entrer en collision deux faisceaux de particules finement focalisés. Par le biais d'une courbe principale, il est possible d'ajuster la position de ces aimants, ce qui est crucial pour obtenir une bonne focalisation (voir également Friedsam et Oren [87]).

Les courbes principales sont utilisées par Kégl et Krzyżak [116] en squelettisation, tâche importante qui constitue généralement une étape préalable à la reconnaissance optique de caractères. Dans Reinhard et Niranjan [160], elles interviennent dans une technique de reconnaissance vocale, tandis que Ozertem et Erdogmus [150] développent une procédure de débruitage basée sur les courbes principales. Couplée à un algorithme de déformation de temps, celle-ci peut être employée pour comparer des séries temporelles qui ne sont pas alignées sur le même axe de temps. En sonification, principe qui consiste à représenter des données sous forme de signaux acoustiques, l'axe de temps peut être défini à l'aide d'une courbe principale (Hermann, Meinicke et Ritter [107]).

D'autres applications ont trait à la géographie ou à la géologie. Ainsi, avec la multiplication des systèmes GPS (*Global Positioning System*), on dispose facilement d'observations donnant la trace d'individus qui se déplacent à pied ou dans un véhicule. Ces données peuvent être utilisées en cartographie, comme par exemple dans le projet *OpenStreetMap* (`http://www.openstreetmap.org`). Or, un même tronçon de route peut donner lieu à plusieurs traces différentes. L'idée mise en œuvre par Brunsdon [43] est alors de moyenner ces observations par le biais des courbes principales en vue d'améliorer la précision du tracé de route obtenu. D'autre part, les courbes principales sont utilisées par Stanford et Raftery [173] pour détecter des failles sismiques ou des champs de mines sur des images de reconnaissance aérienne, et par Banfield et Raftery [18], comme nous l'avons mentionné plus haut, pour identifier les contours de morceaux de banquise sur des images satellite. Einbeck, Tutz et Evers [82], quant à eux, analysent à l'aide des courbes principales une zone de plaines inondables, retrouvant les rivières et les vallées correspondantes. Dans [81], ces auteurs reconstruisent la côte européenne à partir des coordonnées d'hôtels du littoral européen.

Les sciences du vivant constituent une importante sphère d'application des courbes principales. Ces outils sont notamment utiles en écologie. Ainsi, Einbeck, Tutz et Evers [81] s'appuient sur les courbes principales pour observer la répartition sous-marine de colonies de coquilles Saint-Jacques près d'une côte. Une importante branche de l'écologie consiste à étudier la réponse écologique des espèces par rapport à des variables environnementales afin d'évaluer de manière quantitative la niche écologique d'une espèce. Il peut s'agir, par une méthode d'analyse de données multivariées, de démontrer le rôle d'une variable écologique particulière (analyse gradient directe) ou d'expliquer au mieux les gradients de biodiversité à partir des variables écologiques disponibles (analyse gradient indirecte). De'ath [62] observe que les courbes principales mènent à de meilleurs résultats que d'autres méthodes. Après lui, d'autres auteurs se sont appuyés sur les courbes principales pour analyser la relation existant entre deux jeux de données multivariées en biologie, comme

Corkeron, Anthony et Martin [55], qui étudient les dauphins et s'intéressent au lien entre profondeur maximale et temps maximal de plongée.

Les courbes principales peuvent être utilisées dans le domaine médical, en lien avec les différentes techniques d'imagerie. Par exemple, lors d'une angiographie, extraire la ligne médiane des vaisseaux sanguins permet de détecter et comprendre des dysfonctionnements du système cardio-vasculaire (Wong et Chung [190], Wong, So et Chung [191]). Dans la perspective d'étudier l'impact de produits microbicides sur la transmission du VIH, Caffo, Crainiceanu, Deng et Hendrix [45] développent une méthode non-invasive alternative à la sigmoïdoscopie en ajustant une courbe principale sur une image du côlon obtenue par tomographie d'émission monophotonique.

Les courbes principales se révèlent également efficaces en économie, dans l'industrie ou le commerce. Hastie et Stuetzle [104] présentent une méthode de comparaison d'estimations de titrage en or ou autres métaux de deux laboratoires différents. Par ailleurs, Zayed et Einbeck [194] proposent une méthode basée sur les courbes principales destinée à construire à partir de plusieurs sous-indices un nouvel indice qui en soit un résumé convenable. Dans le cadre du trafic autoroutier, les courbes principales peuvent aider à l'analyse de graphiques de la vitesse en fonction de la circulation (Einbeck et Dwyer [80]). Dans l'industrie, l'analyse en composantes principales utilisée en surveillance de procédés peut avantageusement être remplacée par les courbes principales lorsque les processus sont non-linéaires (voir par exemple Dong et McAvoy [72] et Wilson, Irwin et Lightbody [189]). Enfin, Zhang, Wu, Zhang, Huang et Tian [195] emploient des courbes principales pour reconstruire les canaux de navigation dans les eaux intérieures d'un Etat, avec pour objectif de détecter les bateaux circulant en dehors de ces canaux, sur lesquels pèse une suspicion d'activité de pêche non autorisée.

2. Choix d'une courbe principale

Sommaire

Dans le Chapitre 1, nous avons pu constater que le choix de certains paramètres ou des classes de courbes à considérer constitue un problème essentiel dans les différentes définitions de courbe principale et les algorithmes qui s'y rapportent. En effet, comme l'illustre la Figure 2.1, le but est d'obtenir une courbe résumant le mieux possible la forme des données, sans pour autant passer par tous les points.

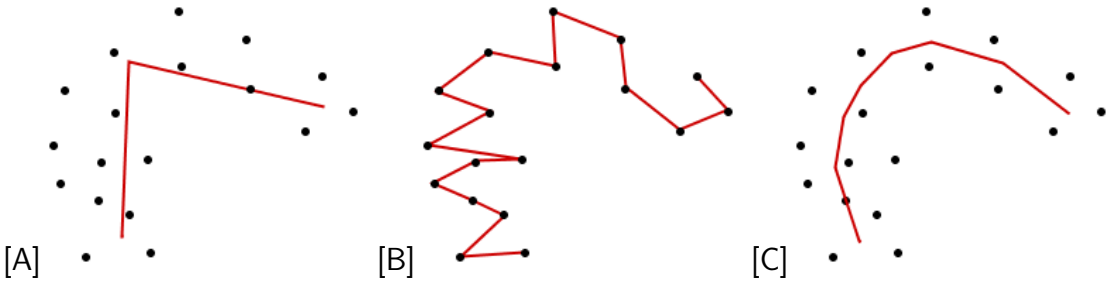


FIGURE 2.1.: Différents résultats de courbe principale selon les paramètres retenus. [A] Courbe trop sommaire. [B] Interpolation. [C] Courbe convenable.

Dans ce chapitre, nous adoptons une fois pour toutes le point de vue de Kégl *et al.* [117] et Sandilya et Kulkarni [166] qui définissent les courbes principales à partir de la minimisation du critère de moindres carrés

$$\Delta(\mathbf{f}) = \mathbb{E}\left[\inf_t \|\mathbf{X} - \mathbf{f}(t)\|^2\right] = \mathbb{E}\left[\|\mathbf{X} - \mathbf{f}(t_{\mathbf{f}}(\mathbf{X}))\|^2\right]$$

et de sa version empirique

$$\Delta_n(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n} \inf_t \|\mathbf{X}_i - \mathbf{f}(t)\|^2 = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - \mathbf{f}(t_{\mathbf{f}}(\mathbf{X}_i))\|^2 \tag{2.1}$$

(critères introduits dans la Section 1.3), et abordons la question du choix d'une bonne classe de courbes. Comme au Chapitre 3 de la première partie, notre approche repose sur la théorie de sélection de modèle par pénalisation de Birgé et Massart [36] et Barron, Birgé et Massart [21]. Le lecteur est invité à se reporter à l'Annexe B pour une présentation du contexte de la sélection de modèle et à consulter la monographie de Massart [141] pour davantage de détails.

Deux approches différentes sont présentées dans les Sections 2.1 et 2.2. Les démonstrations des résultats correspondants sont rassemblées dans les Sections 2.4 et 2.5 respectivement, tandis que la Section 2.3 est consacrée à quelques expériences sur données réelles et simulées relatives à la Section 2.2.

2.1. Un modèle gaussien

2.1.1. Choix de la longueur

Tout d'abord, nous étudions une méthode de sélection de modèle gaussienne pour choisir la longueur d'une courbe principale. Le contexte est similaire à celui de Caillerie et Michel [46]. Dans toute cette section, $\mathbb{R}^d$ est muni du produit scalaire défini par

$$\langle \mathbf{u}, \mathbf{v} \rangle = \frac{1}{d} \sum_{j=1}^{d} u_j v_j, \tag{2.2}$$

et $\| \cdot \|$ désigne la norme euclidienne associée.

Nous supposons que nous observons des vecteurs aléatoires $\mathbf{X}_1, \ldots, \mathbf{X}_n$ à valeurs dans $\mathbb{R}^d$ suivant un modèle gaussien

$$\mathbf{X}_i = \mathbf{x}_i^\star + \sigma \boldsymbol{\xi}_i, \quad i = 1, \ldots, n, \tag{2.3}$$

où les $\mathbf{x}_i^\star$ sont inconnus, les $\boldsymbol{\xi}_i$ sont des vecteurs gaussiens standards de $\mathbb{R}^d$ indépendants et $\sigma > 0$ désigne l'intensité du bruit, supposée connue. Notons $\overrightarrow{\mathbf{X}} = {}^t({}^t\mathbf{X}_1, \ldots, {}^t\mathbf{X}_n)$ le vecteur (colonne) constitué de l'ensemble des coordonnées des vecteurs aléatoires $\mathbf{X}_i$, $i = 1, \ldots, n$. En définissant de même $\overrightarrow{\mathbf{x}^\star}$ et $\overrightarrow{\boldsymbol{\xi}}$, le modèle (2.3) se récrit sous la forme

$$\overrightarrow{\mathbf{X}} = \overrightarrow{\mathbf{x}^\star} + \sigma \overrightarrow{\boldsymbol{\xi}}.$$

Soient F et G deux points fixés de $\mathbb{R}^d$ et $\mathcal{L}$ un sous-ensemble dénombrable de $]0, +\infty[$. Nous introduisons une collection dénombrable $\{\mathcal{F}_\ell\}_{\ell \in \mathcal{L}}$ où chaque ensemble $\mathcal{F}_\ell$ est une classe de courbes paramétrées $\mathbf{f} : I \to \mathbb{R}^d$ de longueur ℓ et d'extrémités F et G. Notre objectif est de sélectionner la longueur ℓ. Nous considérons pour ce faire le critère Δ_n' défini par

$$\begin{aligned} \Delta_n'(\mathbf{f}) &= \frac{1}{n} \sum_{i=1}^{n} \inf_{t \in I} \|\mathbf{X}_i - \mathbf{f}(t)\|^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \inf_{\mathbf{x}_i \in \mathcal{I}_\mathbf{f}} \|\mathbf{X}_i - \mathbf{x}_i\|^2, \end{aligned}$$

où $\mathcal{I}_\mathbf{f}$ désigne l'image de $\mathbf{f}$. Etant donné la définition de la norme $\| \cdot \|$ que nous avons choisie (2.2), il s'agit du critère $\Delta_n(\mathbf{f})$ (2.1) normalisé par la dimension d, dans le but de faire apparaître dans la suite la norme euclidienne normalisée de $\mathbb{R}^{nd}$. Supposons que pour tout $\ell \in \mathcal{L}$, $\overrightarrow{\mathbf{x}}_\ell^{(n)} = (\hat{\mathbf{x}}_{1\ell}, \ldots, \hat{\mathbf{x}}_{n\ell})$ minimise

$$\frac{1}{n} \sum_{i=1}^{n} \|\mathbf{X}_i - \mathbf{x}_i\|^2$$

en

$$\overrightarrow{\mathbf{x}} \in \mathcal{C}_\ell = \bigcup_{\mathbf{f}\in\mathcal{F}_\ell} (\mathcal{I}_\mathbf{f})^n.$$

Pour déterminer ℓ, nous cherchons alors à minimiser en ℓ un critère du type

$$\mathrm{crit}(\ell) = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - \hat{\mathbf{x}}_{i\ell}\|^2 + \mathrm{pen}(\ell),$$

où pen : $\mathcal{L} \to \mathbb{R}^+$ est une fonction de pénalité, destinée à empêcher le choix d'une longueur ℓ trop grande, c'est-à-dire de nature à permettre l'interpolation. Le but est de trouver une fonction de pénalité convenable.

Lorsque les modèles considérés sont linéaires, la pénalité peut être choisie proportionnelle à la dimension du modèle (Birgé et Massart [36]). Ici, les modèles $\mathcal{C}_\ell$ ne sont pas des sous-espaces vectoriels de $\mathbb{R}^{nd}$ et la dimension doit être remplacée par une autre quantité. Pour mesurer la complexité de ces modèles non linéaires, nous utiliserons l'entropie métrique.

Définition 2.1.1 (Nombre de recouvrement et entropie métrique)**.** *Le nombre de recouvrement $\mathcal{N}(S, \|\cdot\|, \varepsilon)$ d'un ensemble S est le nombre minimal de boules de rayon ε pour la norme $\|\cdot\|$ nécessaires pour recouvrir S. L'entropie métrique de S est donnée par*

$$\mathcal{H}(S, \|\cdot\|, \varepsilon) = \ln \mathcal{N}(S, \|\cdot\|, \varepsilon).$$

Notre approche est basée sur un théorème général de sélection de modèle pour des modèles gaussiens non linéaires (Massart [141]) rappelé dans l'Annexe B.2. Notons $\|\cdot\|_{nd}$ la norme de $\mathbb{R}^{nd}$ normalisée, définie par $\langle \overrightarrow{\mathbf{u}}, \overrightarrow{\mathbf{v}} \rangle = \frac{1}{nd}\sum_{i=1}^{nd} u_i v_i$.

Pour tout $\ell \in \mathcal{L}$, soit φ_ℓ une fonction telle que $\varphi_\ell \geq \phi_\ell$, où ϕ_ℓ est donnée par

$$\phi_\ell(u) = \kappa \int_0^u \sqrt{\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)}\mathrm{d}\varepsilon, \tag{2.4}$$

avec κ une constante absolue. On définit d_ℓ par l'équation

$$\varphi_\ell\left(2\sigma\frac{\sqrt{d_\ell}}{\sqrt{nd}}\right) = \frac{\sigma d_\ell}{\sqrt{nd}}.$$

Supposons qu'il existe une famille de poids $\{w_\ell\}_{\ell\in\mathcal{L}}$ vérifiant

$$\sum_{\ell\in\mathcal{L}} e^{-w_\ell} = \Sigma < \infty.$$

Sous ces hypothèses et avec ces notations, le Théorème 4.18 de Massart [141] s'écrit de la façon suivante.

Théorème 2.1.1. *Soient $\eta > 1$ et*

$$\mathrm{pen}(\ell) \geq \eta \frac{\sigma^2}{nd}(\sqrt{d_\ell} + \sqrt{2w_\ell})^2.$$

Alors, presque sûrement, il existe un minimiseur $\hat{\ell}$ du critère pénalisé

$$\mathrm{crit}(\ell) = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - \hat{\mathbf{x}}_{i\ell}\|^2 + \mathrm{pen}(\ell).$$

Si l'on note $\tilde{\mathbf{x}}_i = \hat{\mathbf{x}}_{i\hat{\ell}}$ pour tout i=1,...,n, on a

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\|\tilde{\mathbf{x}}_i - \mathbf{x}_i^\star\|^2 \leq c(\eta)\Big[\inf_{\ell\in\mathcal{L}}(d^2(\overrightarrow{\mathbf{x}^\star}, \mathcal{C}_\ell) + \mathrm{pen}(\ell)) + \frac{\sigma^2}{nd}(\Sigma + 1)\Big],$$

où $d^2(\overrightarrow{\mathbf{x}^\star}, \mathcal{C}_\ell) = \inf_{\overrightarrow{\mathbf{y}}\in\mathcal{C}_\ell} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{x}_i^\star\|^2$.

Ce résultat établit, pour une pénalité $\mathrm{pen}(\ell)$ assez grande, une inégalité de type oracle en espérance pour les $\tilde{\mathbf{x}}_i$, $i = 1, \ldots, n$. Si nous parvenons à évaluer l'intégrale de Dudley [77] (2.4), nous pourrons employer ce théorème dans notre contexte afin de sélectionner la longueur ℓ. Nous aurons besoin pour cela de quelques lemmes intermédiaires, dont les preuves sont reportées à la Section 2.4 pour la clarté de l'exposition.

La première étape consiste à contrôler l'entropie métrique des classes $\mathcal{C}_\ell$, $\ell \in \mathcal{L}$. Pour ce faire, remarquons que pour tout $\ell \in \mathcal{L}$, $\bigcup_{\mathbf{f}\in\mathcal{F}_\ell} \mathcal{I}_\mathbf{f}$ correspond à un ellipsoïde de $\mathbb{R}^d$, noté dans la suite $\mathcal{E}_\ell$.

Lemme 2.1.1. *Toute courbe paramétrée de $\mathbb{R}^d$ d'extrémités F et G, de longueur ℓ ($\ell > FG$), est incluse dans un ellipsoïde $\mathcal{E}_\ell$ de premier axe principal de longueur $a = \ell$, les autres axes étant de longueur $b = \sqrt{\ell^2 - FG^2}$.*

En particulier, dans $\mathbb{R}^2$, $\mathcal{E}_\ell$ est une ellipse ayant F et G pour foyers, et dans $\mathbb{R}^3$, un ellipsoïde de révolution autour de l'axe correspondant à ces deux points.

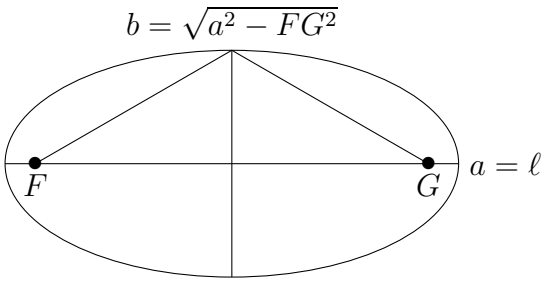


FIGURE 2.2.: Dans le plan $\mathbb{R}^2$, ellipse $\mathcal{E}_\ell$ de foyers F et G et d'axes a et b.

Nous obtenons alors la majoration suivante pour $\mathcal{N}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)$, $\ell \in \mathcal{L}$.

Lemme 2.1.2. *Supposons $a \geq b \geq \varepsilon$. Le nombre de recouvrement de $\mathcal{C}_\ell$ pour la norme $\|\cdot\|_{nd}$ normalisée de $\mathbb{R}^{nd}$ vérifie*

$$\mathcal{N}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^{nd} (ab^{d-1})^n.$$

En majorant, pour tout $\ell \in \mathcal{L}$, l'intégrale

$$\phi_\ell(u) = \kappa \int_0^u \sqrt{\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)} \mathrm{d}\varepsilon,$$

nous pouvons alors définir une fonction φ_ℓ adéquate.

Lemme 2.1.3. *La fonction φ_ℓ donnée par*

$$\varphi_\ell(r) = \begin{cases} \kappa r \sqrt{nd} \left(\sqrt{\ln\left(\frac{2a^{1/d}b^{1-1/d}}{r}\right)} + \sqrt{\pi} \right) & si \quad r \leq b \\ \varphi_\ell(b) + (r-b)\varphi'_\ell(b) & si \quad r \geq b \end{cases}$$

vérifie, pour tout r,

$$\varphi_\ell(r) \geq \phi_\ell(r).$$

Enfin, pour appliquer le Théorème 2.1.1, il nous faut évaluer d_ℓ, défini par l'équation

$$\varphi_\ell\left(\frac{2\sigma\sqrt{d_\ell}}{\sqrt{nd}}\right) = \frac{\sigma d_\ell}{\sqrt{nd}},$$

ce qui fait l'objet du lemme suivant.

Lemme 2.1.4. *Soit φ_ℓ donnée par le Lemme 2.1.3. Supposons que*

$$\sigma \leq \frac{b}{4\kappa} \left[\sqrt{\ln 2 + \frac{1}{d} \ln\left(\frac{a}{b}\right)} + \sqrt{\pi} \right]^{-1}.$$

Alors, l'équation

$$\varphi_\ell\left(\frac{2\sigma\sqrt{d_\ell}}{\sqrt{nd}}\right) = \frac{\sigma d_\ell}{\sqrt{nd}}$$

admet une solution d_ℓ vérifiant

$$d_\ell \leq 8\kappa^2 nd \left(\ln\left(\frac{a^{1/d}b^{1-1/d}}{2\sigma\kappa\sqrt{\pi}}\right) + \pi \right).$$

Nous sommes maintenant en mesure d'énoncer le résultat principal de cette section.

Théorème 2.1.2. *Supposons qu'il existe des poids* $\{w_\ell\}_{\ell\in\mathcal{L}}$ *tels que*

$$\sum_{\ell\in\mathcal{L}} e^{-w_\ell} = \Sigma < \infty,$$

et que, pour tout $\ell \in \mathcal{L}$*,*

$$\sigma \leq \frac{b}{4\kappa}\left[\sqrt{\ln 2 + \frac{1}{d}\ln\left(\frac{a}{b}\right)} + \sqrt{\pi}\right]^{-1}. \tag{2.5}$$

Alors, il existe des constantes c_1*,* c_2 *telles que, pour tout* $\eta > 1$*, si*

$$\mathrm{pen}(\ell) \geq \eta\sigma^2\left[c_1\left(\ln\left(\frac{a^{1/d}b^{1-1/d}}{\sigma}\right) + c_2\right) + \frac{4w_\ell}{nd}\right], \tag{2.6}$$

alors, presque sûrement, il existe un minimiseur $\hat{\ell}$ *du critère pénalisé*

$$\mathrm{crit}(\ell) = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{X}_i - \hat{\mathbf{x}}_{i\ell}\|^2 + \mathrm{pen}(\ell).$$

En outre, si l'on note $\tilde{\mathbf{x}}_i = \hat{\mathbf{x}}_{i\hat{\ell}}$ *pour tout i=1,...,n, on a*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\tilde{\mathbf{x}}_i - \mathbf{x}_i^\star\|^2 \leq c(\eta)\Big[\inf_{\ell\in\mathcal{L}}\{d^2(\overrightarrow{\mathbf{x}^\star}, \mathcal{C}_\ell) + \mathrm{pen}(\ell)\} + \frac{\sigma^2}{nd}(\Sigma+1)\Big],$$

où $d^2(\overrightarrow{\mathbf{x}^\star}, \mathcal{C}_\ell) = \inf_{\overrightarrow{\mathbf{y}}\in\mathcal{C}_\ell}\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{y}_i - \mathbf{x}_i^\star\|^2$.

Cet énoncé appelle quelques commentaires.

La première remarque concerne le fait que le Théorème 2.1.2 fait apparaître des constantes inconnues. Si le Lemme 2.1.4 montre que nous pouvons choisir $c_1 \leq 16\kappa^2$ et $c_2 \leq \pi - \ln(2\kappa\sqrt{\pi})$, ces valeurs n'apportent pas de véritable information, car il s'agit de majorations, qui sont probablement trop larges. En outre, nous avons supposé la variance du bruit σ connue, et constatons que σ intervient dans la pénalité. Or, le niveau de bruit n'est généralement pas connu en pratique. En fait, l'expression (2.6) ne fournit pas directement une fonction de pénalité, mais son intérêt est d'en donner la forme. Il est possible d'estimer σ séparément et de procéder ensuite par *plug-in*. Cependant, une autre solution pour trouver c_1, c_2 et σ consiste à s'appuyer sur la méthode de la pente, déjà utilisée dans les simulations du Chapitre 3 et présentée plus en détail dans l'Annexe B.3.

D'après les formules reliant ℓ à a et b, la quantité $\ln(a^{1/d}b^{1-1/d})$ dans la pénalité caractérise chaque modèle de courbes de longueur ℓ. Les autres éléments variant sur

la collection de modèles sont les poids $\{w_\ell\}_{\ell\in\mathcal{L}}$. Rappelons que dans le cas linéaire, où chaque modèle S_ℓ est de dimension D_ℓ, un choix possible pour w_ℓ est $w_\ell = w(D_\ell)$ où $w(D) = cD + \ln|\{\ell \in \mathcal{L}, D_\ell = D\}|$ et $c > 0$ (voir Massart [141], Section 4.2.1). S'il n'y a pas de redondance dans la dimension des modèles, cette stratégie revient à choisir w_ℓ proportionnel à D_ℓ. Par analogie, w_ℓ peut ici être pris proportionnel à $\ln(a^{1/d}b^{1-1/d})$. Plus formellement, nous pouvons poser $w_\ell = c\ln a^{1/d}b^{1-1/d}$, où la constante $c > 0$ est telle que $\sum_{\ell\in\mathcal{L}} \frac{1}{a^{c/d}b^{c(1-1/d)}} = \Sigma < +\infty$. Dans ces conditions, la pénalité est finalement proportionnelle à $\ln(a^{1/d}b^{1-1/d})$ et pourrait donc en pratique être calibrée en utilisant l'heuristique de pente.

Observons par ailleurs que la condition (2.5) exprime que le niveau de bruit σ ne doit pas être trop grand par rapport à b. Autrement dit, si $b = \sqrt{\ell^2 - FG^2}$ est de l'ordre de σ, il n'est pas possible d'obtenir une courbe principale convenable de longueur ℓ.

Enfin, nous pouvons noter qu'en raison de l'exposant n dans le nombre de recouvrement dans le Lemme 2.1.2 — un commentaire à ce sujet est donné à la suite de la preuve du lemme dans la Remarque 2.4.1 —, la forme de la pénalité obtenue ne tend pas vers 0 lorsque n tend vers l'infini. Ce point est intrinsèquement lié à la géométrie du problème. En effet, sa résolution n'est pas facilitée par l'augmentation de la taille de l'échantillon dans la mesure où nous n'avons rien spécifié sur la répartition des $\mathbf{x}_i^\star$. Une piste de recherche future consisterait à supposer que ces derniers sont distribués sur la courbe selon une loi uniforme et à regarder le problème dans le cadre de la sélection de modèle en estimation de densité. Ce point de vue met en jeu des calculs d'entropie à crochets de classes de densités de mélanges gaussiens continus en dimension d.

2.1.2. Arbre couvrant de poids minimal

Rappelons que nous avons supposé fixées les extrémités F et G de la courbe principale recherchée. D'un point de vue pratique, il nous faut déterminer F et G à partir des observations. Une solution possible consiste à choisir les points les plus éloignés dans l'arbre couvrant de poids minimal de l'ensemble (ou d'un sous-ensemble) des données. Quelques définitions préalables sont nécessaires.

Définition 2.1.2 (Arbre)**.** *Un arbre est un graphe connexe non orienté, qui ne contient aucun cycle.*

La définition d'un arbre est illustrée dans la Figure 2.3. Un arbre couvrant de poids minimal est un type d'arbre particulier.

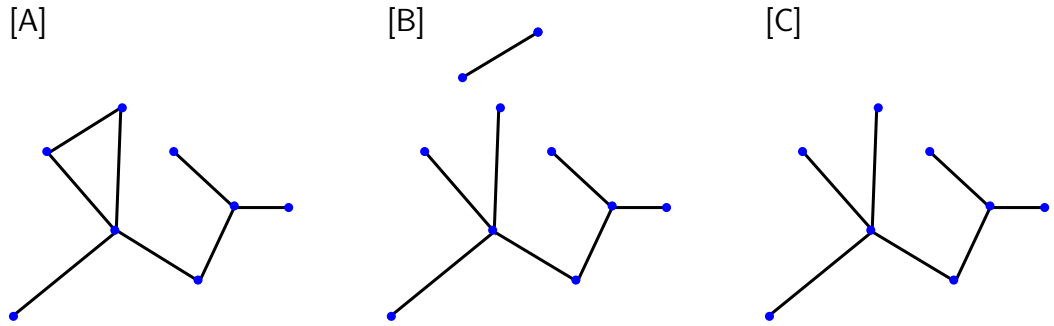


FIGURE 2.3.: [A] Ce graphe contient un cycle. [B] Non connexe. [C] Exemple d'arbre.

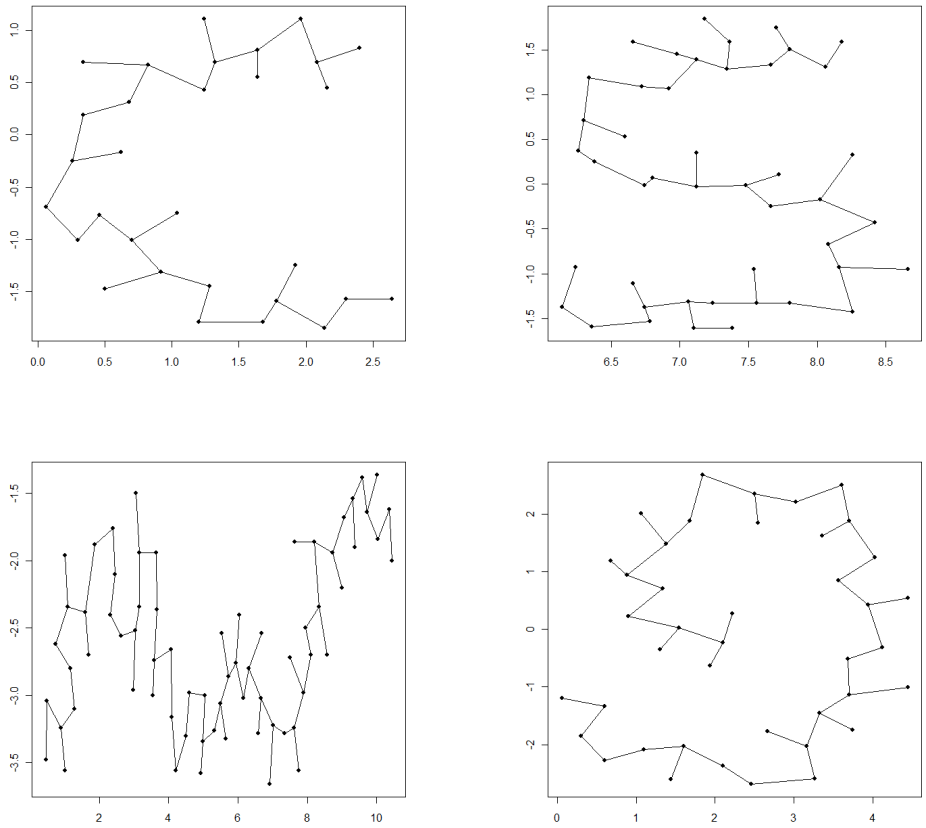


FIGURE 2.4.: Quelques exemples d'arbres couvrants de poids minimal.

Définition 2.1.3 (Arbre couvrant de poids minimal)**.** *Un arbre couvrant associé à un ensemble de sommets $\mathcal{V}$ est un arbre connectant tous ces sommets. Dans le cas où les arêtes sont valuées, un arbre couvrant dont la somme des poids des arêtes est minimale est appelé arbre couvrant de poids minimal.*

Dans notre contexte, chaque arête potentielle reliant deux observations se voit attribuer un poids égal à sa longueur. Après avoir construit un arbre couvrant de poids minimal, qui se trouve être ici un arbre couvrant de longueur minimale, nous pouvons définir F et G comme les deux points les plus éloignés dans cet arbre.

Les deux principaux algorithmes permettant de construire un arbre couvrant de poids minimal sont l'algorithme de Kruskal [123] et celui de Prim [158]. Les arbres couvrants de longueur minimale représentés dans la Figure 2.4 ont été obtenus via l'algorithme de Prim. Un sommet est choisi au hasard et relié au sommet le plus proche. Puis, à chaque étape, l'arbre croît d'une arête, qui doit être la plus courte possible. Les étapes de l'algorithme de Prim sont illustrées dans la Figure 2.5. Dans la méthode de Kruskal, les arêtes, préalablement rangées par ordre croissant, sont ajoutées une à une jusqu'à obtenir un arbre couvrant. Pour ce faire, seuls sont autorisés les ajouts ne formant pas de cycle.

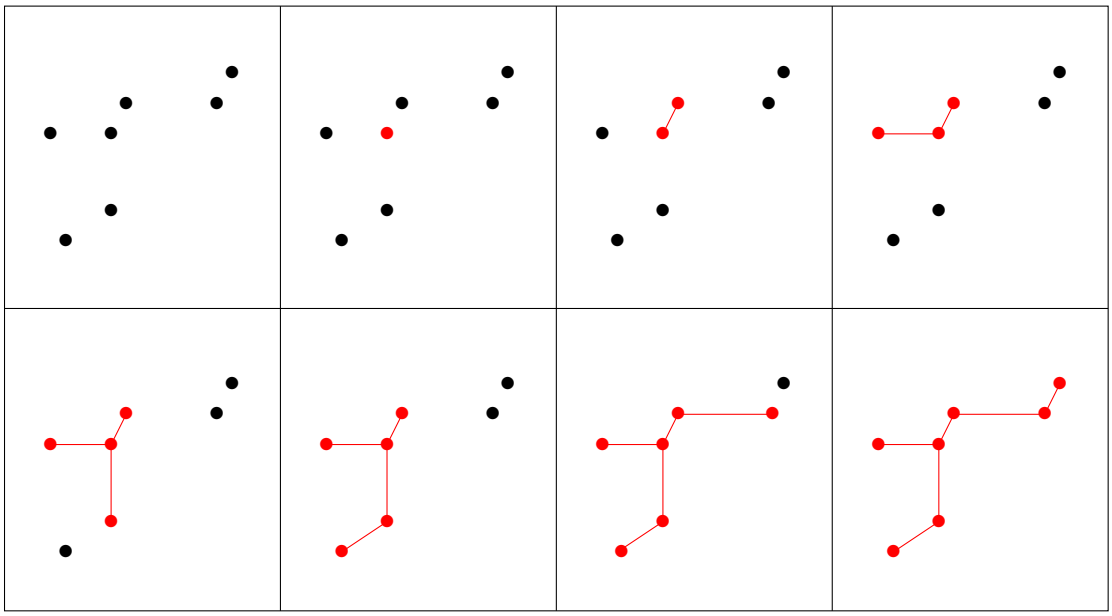


FIGURE 2.5.: Algorithme de Prim. Etapes de la construction d'un arbre couvrant de longueur minimale.

2.2. Modèles bornés*

Dans la section précédente, nous avons inscrit notre problème dans un cadre gaussien. En intégrant dans la forme des modèles la contrainte introduite par la borne inférieure sur t, le critère à minimiser a été écrit de manière à faire apparaître la norme euclidienne normalisée de $\mathbb{R}^{nd}$, ce qui nous a permis d'utiliser des résultats de sélection de modèle pour des modèles gaussiens non linéaires. Or, ainsi défini, ce critère porte sur les points échantillonnés sur la courbe, et donc cette approche conduit à une inégalité de type oracle pour les estimateurs $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n$ de $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Pour établir un résultat portant sur un estimateur de la courbe principale elle-même, il nous faut à présent mettre en œuvre une autre stratégie, reposant sur le critère $\Delta_n(\mathbf{f})$ à minimiser en $\mathbf{f} : I \to \mathbb{R}^d$. Pour ce faire, nous allons nous appuyer, comme dans le Chapitre 3 de la première partie, sur le résultat général de sélection de modèle de Massart [141] rappelé dans l'Annexe B.2 (Théorème B.2.2). Dans cette optique, nous supposons dans toute la section que

$$\mathbb{P}\{\mathbf{X} \in \mathcal{C}\} = 1,$$

où $\mathcal{C}$ est un convexe compact de $\mathbb{R}^d$, de diamètre δ. Notons qu'en posant une telle hypothèse, nous sortons ipso facto du cadre gaussien. Dans la suite, $\|\cdot\|$ désigne la norme euclidienne standard de $\mathbb{R}^d$ et $\mathbf{X}_1, \ldots, \mathbf{X}_n$ un échantillon d'un vecteur aléatoire générique $\mathbf{X}$ de $\mathbb{R}^d$ tel que $\mathbb{E}\|\mathbf{X}\|^2 < +\infty$. Les démonstrations des résultats de cette section sont reportés à la Section 2.5.

2.2.1. Courbes principales de longueur bornée

Dans un premier temps, nous nous plaçons dans le cadre de Kégl *et al.* [117] et considérons des courbes de longueur bornée. Par le Lemme 1 dans Kégl [115], l'hypothèse $\mathbb{P}\{\mathbf{X} \in \mathcal{C}\} = 1$ implique que, pour tout $L > 0$, il existe une courbe principale pour $\mathbf{X}$ de longueur au plus L dans $\mathcal{C}$, c'est-à-dire une courbe paramétrée $\mathbf{f}^\star : I \to \mathbb{R}^d$ (non nécessairement unique) de longueur inférieure ou égale à L et d'image dans $\mathcal{C}$ réalisant le minimum de $\mathbb{E}\left[\inf_{t\in I} \|\mathbf{X} - \mathbf{f}(t)\|^2\right]$. En conséquence, nous ne nous intéresserons dans ce qui suit qu'aux courbes principales dont l'image est incluse dans $\mathcal{C}$. Notons $\mathcal{F}$ l'ensemble de toutes les courbes $\mathbf{f} = (f_1, \ldots, f_d)$ paramétrées sur I d'image dans $\mathcal{C}$.

Nous considérons le contraste

$$\Delta(\mathbf{f}, \mathbf{x}) = \inf_{t\in I} \|\mathbf{x} - \mathbf{f}(t)\|^2, \quad \mathbf{f} \in \mathcal{F}, \mathbf{x} \in \mathbb{R}^d,$$

*Cette section a donné lieu à un article repris dans l'Annexe F, écrit en collaboration avec Gérard Biau.

et le risque empirique associé

$$\Delta_n(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n} \inf_{t\in I} \|\mathbf{X}_i - \mathbf{f}(t)\|^2.$$

Pour une longueur $L > 0$, nous posons

$$\mathbf{f}^\star \in \underset{\mathbf{f}\in\mathcal{F},\mathscr{L}(\mathbf{f})\leq L}{\arg\min}\ \mathbb{E}[\Delta(\mathbf{f},\mathbf{X})].$$

Soit $\mathcal{L}$ un sous-ensemble dénombrable de $]0, L]$ et $\mathcal{Q}$ une grille assez fine sur $\mathcal{C}$. Pour tous $k \geq 1$ et $\ell \in \mathcal{L}$, le modèle $\mathcal{F}_{k,\ell}$ est défini comme la collection de toutes les lignes polygonales à k segments et de longueur au plus ℓ, dont les sommets appartiennent à $\mathcal{Q}$. Remarquons que la collection $\{\mathcal{F}_{k,\ell}\}_{k\geq 1,\ell\in\mathcal{L}}$ est dénombrable et qu'il en va de même pour chaque modèle $\mathcal{F}_{k,\ell}$. Pour tous $k \geq 1$ et $\ell \in \mathcal{L}$, soit

$$\hat{\mathbf{f}}_{k,\ell} \in \underset{\mathbf{f}\in\mathcal{F}_{k,\ell}}{\arg\min}\, \Delta_n(\mathbf{f})$$

une ligne polygonale minimisant le critère empirique $\Delta_n(\mathbf{f})$ sur tous les éléments de la classe $\mathcal{F}_{k,\ell}$.

A ce stade, nous disposons d'une famille d'estimateurs $\{\hat{\mathbf{f}}_{k,\ell}\}_{k\geq 1,\ell\in\mathcal{L}}$ et cherchons à sélectionner la meilleure courbe principale parmi tous les éléments de cette collection. Pour cela, notre approche repose à nouveau sur la théorie de sélection de modèle par pénalisation. Il s'agit de construire une fonction de pénalité convenable $\mathrm{pen} : \mathbb{N}^* \times \mathcal{L} \to \mathbb{R}^+$ et de retenir ensuite la courbe correspondant au couple $(\hat{k}, \hat{\ell})$ qui minimise le critère pénalisé

$$\mathrm{crit}(k,\ell) = \Delta_n(\hat{\mathbf{f}}_{k,\ell}) + \mathrm{pen}(k,\ell).$$

La performance de cet estimateur $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k},\hat{\ell}}$ est alors évaluée au moyen de la perte

$$\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) = \mathbb{E}[\Delta(\tilde{\mathbf{f}}, \mathbf{X}) - \Delta(\mathbf{f}^\star, \mathbf{X})].$$

Dans ce contexte, nous pouvons énoncer le résultat suivant, qui est une adaptation du Théorème 8.1 de Massart [141] (voir Annexe B.2).

Théorème 2.2.1. *Considérons une famille* $\{x_{k,\ell}\}_{k\geq 1,\ell\in\mathcal{L}}$ *de poids positifs tels que*

$$\sum_{k\geq 1,\ell\in\mathcal{L}} e^{-x_{k,\ell}} = \Sigma < +\infty,$$

et une fonction de pénalité $\mathrm{pen} : \mathbb{N}^* \times \mathcal{L} \to \mathbb{R}^+$. *Soit* $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k},\hat{\ell}}$. *Si pour tout* $(k,\ell) \in \mathbb{N}^* \times \mathcal{L}$,

$$\mathrm{pen}(k,\ell) \geq \mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}} (\mathbb{E}[\Delta(\mathbf{f},\mathbf{X})] - \Delta_n(\mathbf{f}))\right] + \delta^2\sqrt{\frac{x_{k,\ell}}{2n}},$$

alors

$$\mathbb{E}[\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}})] \leq \inf_{k\geq 1, \ell\in\mathcal{L}} \Big[\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\ell}) + \mathrm{pen}(k,\ell)\Big] + \frac{\delta^2\Sigma}{2^{3/2}}\sqrt{\frac{\pi}{n}},$$

où $\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\ell}) = \inf_{\mathbf{f}\in\mathcal{F}_{k,\ell}} \mathcal{D}(\mathbf{f}^\star, \mathbf{f})$.

Le Théorème 2.2.1 propose une forme de pénalité en fonction de la quantité $\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}} (\mathbb{E}[\Delta(\mathbf{f},\mathbf{X})] - \Delta_n(\mathbf{f}))\right]$ et établit une inégalité de type oracle en espérance pour le minimiseur du contraste empirique pénalisé correspondant, pour la perte $\mathcal{D}(\mathbf{f}^\star, \mathbf{f}) = \mathbb{E}[\Delta(\mathbf{f},\mathbf{X}) - \Delta(\mathbf{f}^\star,\mathbf{X})]$.

Dans le but d'appliquer le Théorème 2.2.1 à notre problème, nous cherchons donc à obtenir une borne supérieure pour

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}} (\mathbb{E}[\Delta(\mathbf{f},\mathbf{X})] - \Delta_n(\mathbf{f}))\right].$$

Pour ce faire, remarquons tout d'abord que, par symétrisation, il suffit de majorer la moyenne de Rademacher

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \Delta(\mathbf{f},\mathbf{X}_i)\right]. \tag{2.7}$$

Lemme 2.2.1 (Symétrisation)**.** *Soient* $\varepsilon_1,\dots,\varepsilon_n$ *des variables aléatoires de Rademacher indépendantes, indépendantes de* $\mathbf{X}_1,\dots,\mathbf{X}_n$*. On a*

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}} (\mathbb{E}[\Delta(\mathbf{f},\mathbf{X})] - \Delta_n(\mathbf{f}))\right] \leq 2\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \Delta(\mathbf{f},\mathbf{X}_i)\right].$$

La moyenne de Rademacher (2.7) peut ensuite être majorée par une intégrale de Dudley [78]. Plus précisément, en notant

$$S_{k,\ell} = \{\Delta(\mathbf{f},\cdot), \mathbf{f}\in\mathcal{F}_{k,\ell}\},$$

il existe une constante $c>0$ telle que, pour tous $\mathbf{X}_1,\dots,\mathbf{X}_n$,

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \Delta(\mathbf{f},\mathbf{X}_i)\right] \leq \frac{c}{\sqrt{n}}\int_0^{+\infty} \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon.$$

Pour évaluer le nombre de recouvrement de $S_{k,\ell}$, nous utilisons alors la borne suivante, établie par Kégl *et al.* [117] :

$$\mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon) \leq 2^{\ell\delta/\varepsilon + 3k+1} V_d^{k+1} \left(\frac{\delta^2\sqrt{d}}{\varepsilon} + \sqrt{d}\right)^d \left(\frac{\ell\delta\sqrt{d}}{k\varepsilon} + 3\sqrt{d}\right)^{kd}, \tag{2.8}$$

où V_d désigne le volume de la boule unité en dimension d. Finalement, nous obtenons la proposition suivante.

Proposition 2.2.1. *Il existe des constantes positives $a_0, \dots, a_2$, ne dépendant que de L, d et δ, telles que*

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}} \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta(\mathbf{f},\mathbf{X}_i)\right] \leq \frac{1}{\sqrt{n}}\left[a_1\sqrt{k} + a_2\frac{\ell}{\sqrt{k}} + a_0\right].$$

En combinant le Lemme 2.2.1 et la Proposition 2.2.1, le Théorème 2.2.1 se récrit de la manière suivante.

Théorème 2.2.2. *Considérons une famille de poids positifs $\{x_{k,\ell}\}_{k\geq 1,\ell\in\mathcal{L}}$ tels que*

$$\sum_{k\geq 1,\ell\in\mathcal{L}} e^{-x_{k,\ell}} = \Sigma < +\infty,$$

et une fonction de pénalité $\mathrm{pen} : \mathbb{N}^*\times\mathcal{L} \to \mathbb{R}^+$. *Soit* $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k},\hat{\ell}}$. *Il existe des constantes positives $c_0, \dots, c_2$, ne dépendant que de la longueur maximale L, de la dimension d et du diamètre δ du convexe $\mathcal{C}$, telles que, si pour tout $(k,\ell) \in \mathbb{N}^*\times\mathcal{L}$,*

$$\mathrm{pen}(k,\ell) \geq \frac{1}{\sqrt{n}}\left[c_1\sqrt{k} + c_2\frac{\ell}{\sqrt{k}} + c_0\right] + \delta^2\sqrt{\frac{x_{k,\ell}}{2n}},$$

alors

$$\mathbb{E}[\mathcal{D}(\mathbf{f}^\star,\tilde{\mathbf{f}})] \leq \inf_{k\geq 1,\ell\in\mathcal{L}}\left[\mathcal{D}(\mathbf{f}^\star,\mathcal{F}_{k,\ell}) + \mathrm{pen}(k,\ell)\right] + \frac{\delta^2\Sigma}{2^{3/2}}\sqrt{\frac{\pi}{n}},$$

où $\mathcal{D}(\mathbf{f}^\star,\mathcal{F}_{k,\ell}) = \inf_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\mathcal{D}(\mathbf{f}^\star,\mathbf{f})$.

Dans la pénalité obtenue interviennent un terme proportionnel à $\sqrt{k/n}$ et un terme proportionnel à $\ell/\sqrt{kn}$. Cette forme de pénalité paraît tout à fait pertinente, dans la mesure où le nombre k de segments et la longueur ℓ des courbes mesurent effectivement la complexité des modèles. De plus, nous pouvions nous attendre à ce que k et ℓ soient liés, comme le suggère le terme $\ell/\sqrt{kn}$. Si la longueur ℓ augmente, il faut autoriser davantage de segments afin de conserver une certaine régularité. Notons que la pénalité, proportionnelle à $1/\sqrt{n}$, tend vers 0 avec n, contrairement à celle obtenue dans la Section 2.1.

Remarquons que la démonstration de la Proposition 2.2.1 fournit des valeurs possibles pour les constantes $c_0, \dots, c_2$. Cependant, comme dans le contexte gaussien, il s'agit de majorations. Notons simplement que $c_1 = c_1'\delta^2$, $c_2 = c_2'\delta$ et $c_0 = c_0'\delta^2$, où c_0', c_1' et c_2' sont des constantes sans dimension, de sorte que la forme de la pénalité est homogène au carré d'une longueur, comme l'est le critère $\Delta_n(\mathbf{f})$.

Concernant le choix des poids $\{x_{k,\ell}\}_{k\geq 1,\ell\in\mathcal{L}}$, si le cardinal de la collection de modèles ne dépasse pas n^2 (ce qui sera le cas dans les exemples pratiques de la Section 2.3), nous pouvons poser $x_{k,\ell} = 2\ln n$ pour tout (k,ℓ). Ce choix ne change pas la forme de la pénalité (bien que la vitesse soit modifiée) et entraîne $\Sigma = 1$.

Remarque 2.2.1. Si la longueur ℓ des lignes polygonales est fixée, l'objectif étant de sélectionner le nombre k de segments, c'est le terme en $\sqrt{k/n}$ qui traduit la complexité des modèles dans la pénalité, et puisqu'il n'y a pas de redondance dans la dimension des modèles, les poids peuvent également être choisis de cet ordre, comme indiqué dans la Section 2.1. Dans ce cas, nous obtenons ainsi une pénalité en $\sqrt{k/n}$.

2.2.2. Courbes principales de courbure intégrale bornée

Nous nous sommes concentrés dans ce qui précède sur le cadre des courbes principales de longueur bornée de Kégl *et al.* [117]. Comme nous l'avons vu dans le Chapitre 1, Sandilya et Kulkarni [166] ont proposé une approche alternative, basée sur le contrôle de la courbure intégrale. Nous considérons à présent la question du choix d'une bonne courbe principale dans le contexte des courbes principales de courbure intégrale bornée. Grâce à un résultat exprimant qu'une courbe de courbure intégrale bornée a également une longueur bornée, nous pourrons établir un résultat de sélection de modèle analogue à celui de la section précédente.

L'hypothèse $\mathbb{P}\{\mathbf{X} \in \mathcal{C}\} = 1$ assure qu'il existe une courbe de courbure intégrale bornée minimisant le critère $\mathbb{E}[\Delta(\mathbf{f}, X)]$ (voir le Chapitre 1). Pour $K \geq 0$, supposons

$$\mathbf{f}^\star \in \underset{\mathbf{f}\in\mathcal{F}, \mathscr{K}(\mathbf{f})\leq K}{\arg\min}\ \mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})],$$

où $\mathscr{K}(\mathbf{f})$ désigne la courbure intégrale de la courbe $\mathbf{f}$. Pour un sous-ensemble dénombrable $\mathcal{K}$ de $[0, K]$, nous définissons une collection dénombrable de modèles $\{\mathcal{F}_{k,\kappa}\}_{k\geq 1, \kappa\in\mathcal{K}}$, où chaque $\mathcal{F}_{k,\kappa}$ est formé de lignes polygonales à k segments, de courbure intégrale au plus κ, ayant ses sommets sur une grille $\mathcal{Q}$ de $\mathcal{C}$. Pour tous $k \geq 1$ et $\kappa \in \mathcal{K}$, soit

$$\hat{\mathbf{f}}_{k,\kappa} \in \underset{\mathbf{f}\in\mathcal{F}_{k,\kappa}}{\arg\min}\, \Delta_n(\mathbf{f})$$

une ligne polygonale minimisant le critère empirique $\Delta_n(\mathbf{f})$ sur tous les éléments de la classe $\mathcal{F}_{k,\kappa}$. Afin de choisir la meilleure courbe principale dans la famille $\{\hat{\mathbf{f}}_{k,\kappa}\}_{k\geq 1,\kappa\in\mathcal{K}}$, notre objectif est de trouver une fonction de pénalité appropriée $\mathrm{pen} : \mathbb{N}^* \times \mathcal{K} \to \mathbb{R}^+$. Plus formellement, la minimisation du critère

$$\mathrm{crit}(k, \kappa) = \Delta_n(\hat{\mathbf{f}}_{k,\kappa}) + \mathrm{pen}(k, \kappa)$$

doit fournir une courbe principale convenable, dont la qualité sera mesurée par la perte

$$\mathcal{D}(\mathbf{f}^\star, \mathbf{f}) = \mathbb{E}[\Delta(\mathbf{f}, \mathbf{X}) - \Delta(\mathbf{f}^\star, \mathbf{X})].$$

Pour une fonction pen(k, κ), notons $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k},\hat{\kappa}}$, où $(\hat{k}, \hat{\kappa})$ minimise le critère pénalisé crit(k, κ).

Pour obtenir un résultat de la forme du Théorème 2.2.2, nous savons qu'il suffit de trouver une borne supérieure pour l'intégrale

$$\int_0^{+\infty} \sqrt{\ln \mathcal{N}(S_{k,\kappa}, \|\cdot\|_\infty, \varepsilon)} \mathrm{d}\varepsilon, \tag{2.9}$$

où

$$S_{k,\kappa} = \{\Delta(\mathbf{f}, \cdot), \mathbf{f} \in \mathcal{F}_{k,\kappa}\}.$$

A cette fin, nous utilisons la majoration suivante, due à Sandilya et Kulkarni [166] :

$$\mathcal{N}(S_{k,\kappa}, \|\cdot\|_\infty, \varepsilon) \leq 2^{\delta^2\zeta(\kappa)/\varepsilon + 2k+1} V_d^{k+1} ((\delta^2/\varepsilon + 1)\sqrt{d})^d ((\delta^2\zeta(\kappa)/k\varepsilon + 3)\sqrt{d})^{kd},$$

où V_d est le volume de la boule unité en dimension d. Cette adaptation du résultat de Kégl *et al.* [117] (2.8) au contexte des courbes principales de courbure intégrale bornée repose sur le lemme suivant (voir par exemple le livre d'Alexandrov et Reshetnyak [7, Chapitre 2]), qui établit un lien intéressant entre la longueur d'une courbe et sa courbure intégrale,

Lemme 2.2.2. *Soient* $\mathbf{f}$ *une courbe de courbure intégrale* κ *et* δ *le diamètre de* $\mathcal{C}$. *Alors,* $\mathscr{L}(\mathbf{f}) \leq \delta\zeta(\kappa)$, *où la fonction* ζ *est définie par*

$$\zeta(x) = \begin{cases} \dfrac{1}{\cos(x/2)} & \text{si} \quad 0 \leq x \leq \dfrac{\pi}{2} \\ 2\sin(x/2) & \text{si} \quad \dfrac{\pi}{2} \leq x \leq \dfrac{2\pi}{3} \\ \dfrac{x}{2} - \dfrac{\pi}{3} + \sqrt{3} & \text{si} \quad x \geq \dfrac{2\pi}{3}. \end{cases}$$

Le graphe de la fonction ζ est représenté dans la Figure 2.6.

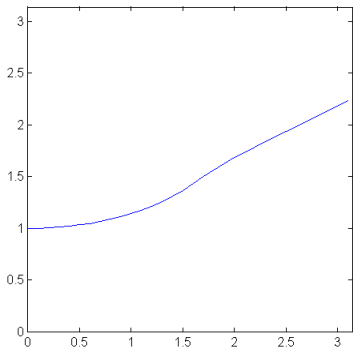


FIGURE 2.6.: Représentation graphique de la fonction ζ.

Remarque 2.2.2. L'inégalité donnée par le Lemme 2.2.2 est exacte dans le sens suivant : pour tout $\kappa \geq 0$ et tout $\delta > 0$, il existe une courbe de courbure intégrale κ, dont l'image est incluse dans une boule de diamètre δ et dont la longueur est exactement égale à $\delta\zeta(\kappa)$.

L'approche développée dans la Section 2.2.1 s'adapte ainsi au cadre où la longueur est remplacée par la courbure intégrale. La Proposition 2.2.2 constitue le pendant de la Proposition 2.2.1 de la section précédente.

Proposition 2.2.2. *Il existe des constantes positives $a_0, \ldots, a_4$, ne dépendant que de d, telles que*

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\kappa}} \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta(\mathbf{f},\mathbf{X}_i)\right] \leq \delta^2\left(a_1\sqrt{k} + a_2\sqrt{\zeta(\kappa)} + a_3\frac{\zeta(\kappa)}{\sqrt{k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}<1\}} + a_4\sqrt{k\ln\frac{\zeta(\kappa)}{k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}\geq 1\}} + a_0\right).$$

Finalement, nous sommes en mesure d'énoncer le théorème suivant.

Théorème 2.2.3. *Considérons une famille de poids positifs $\{x_{k,\kappa}\}_{k\geq 1,\kappa\in\mathcal{K}}$ tels que*

$$\sum_{k\geq 1,\kappa\in\mathcal{K}} e^{-x_{k,\kappa}} = \Sigma < +\infty,$$

et une fonction de pénalité $\mathrm{pen} : \mathbb{N}^* \times \mathcal{K} \to \mathbb{R}^+$. *Soit* $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k},\hat{\kappa}}$. *Il existe des constantes positives $c_0, \ldots, c_2$, ne dépendant que de la dimension d, telles que, si pour tout $(k,\kappa) \in \mathbb{N}^* \times \mathcal{K}$,*

$$\mathrm{pen}(k,\kappa) \geq \frac{\delta^2}{\sqrt{n}}\left[c_1\sqrt{k} + c_2\max\left(\frac{\zeta(\kappa)}{\sqrt{k}}, \sqrt{k\ln\frac{\zeta(\kappa)}{k}}\right) + c_0 + \sqrt{\frac{x_{k,\kappa}}{2}}\right],$$

alors

$$\mathbb{E}[\mathcal{D}(\mathbf{f}^\star,\tilde{\mathbf{f}})] \leq \inf_{k\geq 1,\kappa\in\mathcal{K}}\left[\mathcal{D}(\mathbf{f}^\star,\mathcal{F}_{k,\kappa}) + \mathrm{pen}(k,\kappa)\right] + \frac{\delta^2\Sigma}{2^{3/2}}\sqrt{\frac{\pi}{n}},$$

où $\mathcal{D}(\mathbf{f}^\star,\mathcal{F}_{k,\kappa}) = \inf_{\mathbf{f}\in\mathcal{F}_{k,\kappa}}\mathcal{D}(\mathbf{f}^\star,\mathbf{f})$.

Dans l'expression de la pénalité intervient un terme en $\sqrt{k/n}$, comme dans le cas des courbes de longueur bornée, tandis que la longueur ℓ est remplacée par $\zeta(\kappa)$, fonction croissante de la courbure intégrale κ. La forme obtenue est cohérente, puisque le nombre de segments k et la courbure intégrale κ reflètent la complexité des modèles. En outre, comme précédemment, un terme lie ces deux quantités, ce qui traduit le fait que, pour obtenir une courbe principale correcte, k et κ doivent être bien choisis l'un relativement à l'autre. Grosso modo, une courbure plus importante implique davantage de segments.

2.3. Résultats expérimentaux

Cette section présente quelques simulations et expériences sur données réelles, effectuées avec le logiciel MATLAB, ayant pour objectif d'illustrer la procédure de sélection de modèle suggérée par le Théorème 2.2.2 et le Théorème 2.2.3. Les pénalités proposées par ces théorèmes font intervenir des constantes, qui doivent être déterminées en pratique. Pour ce faire, une direction possible consiste à utiliser l'heuristique de pente comme dans le Chapitre 3 de la première partie. Plus précisément, nous avons choisi de procéder par estimation directe de la pente. Rappelons que l'heuristique de pente, introduite par Birgé et Massart [37] et développée entre autres par Arlot et Massart [10]), est une méthode de calibration permettant d'ajuster une pénalité connue à constante multiplicative près. L'hypothèse sous-jacente pour appliquer cette technique, à savoir que le contraste empirique diminue lorsque la complexité des modèles augmente, est effectivement vérifiée dans notre contexte de courbes principales.

D'un point de vue algorithmique, deux stratégies différentes ont été implémentées, désignées dans la suite par **MS1** et **MS2** (l'acronyme « MS » signifie « sélection de modèle »).

- La méthode **MS1**, qui correspond de plus près à la théorie développée dans la Section 2.2, est basée sur le choix simultané du nombre k de segments et de la longueur ℓ de la courbe. Pour chaque nombre de segments $k = 1, \ldots, 80$ et pour une gamme de valeurs de la longueur ℓ (la longueur maximale L et le pas dépendent de l'échelle du jeu de données), nous avons calculé le critère

$$\Delta_n(\hat{\mathbf{f}}_{k,\ell}) = \frac{1}{n}\sum_{i=1}^{n} \Delta(\hat{\mathbf{f}}_{k,\ell}, \mathbf{X}_i).$$

  Alors, considérant pour la commodité des calculs une pénalité simplifiée de la forme $c_1\sqrt{k} + c_2\ell$, nous avons sélectionné les constantes c_1 et c_2 en implémentant une version bivariée de l'heuristique de pente. Plus précisément, nous supposons que pour les grandes valeurs de k et ℓ, le critère $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ se comporte comme $c_1\sqrt{k} + c_2\ell$, et les constantes c_1 et c_2 sont choisies grâce à une étape de régression.

- Le second algorithme, **MS2**, est une adaptation du *Polygonal Line Algorithm* de Kégl, Krzyżak, Linder et Zeger [117]. Dans cette procédure, les sommets de la courbe principale sont optimisés l'un après l'autre de manière cyclique, comme nous l'avons vu dans le Chapitre 1, et la courbure est contrôlée à chaque étape au moyen d'une pénalité locale sur les angles. Nous n'avons pas cherché ici à optimiser cette dernière et l'avons fixée comme préconisé dans

[117]. Donc, dans cette seconde approche, la forme de notre pénalité se réduit finalement à $c\sqrt{k/n}$. Pour calibrer la constante c, nous avons utilisé l'interface CAPUSHE de Baudry, Maugis et Michel [25] comme dans le Chapitre 3 de la première partie.

2.3.1. Données simulées

Dans cette première série d'expériences, nous considérons des données planaires distribuées autour d'une courbe de référence avec un certain bruit. Plus formellement, les observations ont été générées à partir du modèle

$$\mathbf{X} = \mathbf{Y} + \boldsymbol{\varepsilon}, \tag{2.10}$$

où $\mathbf{Y}$ est distribué uniformément sur une courbe planaire $\mathbf{f}$ et $\boldsymbol{\varepsilon}$ est un bruit gaussien bivarié, indépendant de $\mathbf{Y}$. Même si, comme expliqué dans le Chapitre 1, la courbe générative $\mathbf{f}$ n'est pas une courbe principale en général (en raison du biais de modèle décrit dans la Section 1.1.2), ce modèle gaussien est considéré comme une référence pour les simulations dans la littérature sur les courbes principales.

Dans un premier exemple, soit $\mathbf{f}$ un demi-cercle de rayon 1. La variance du bruit est égale à 0.004 et le nombre n d'observations est fixé à 100 (voir Figure 2.7).

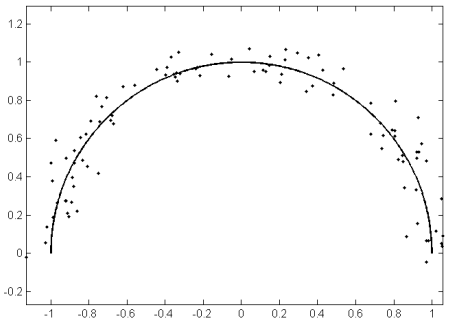
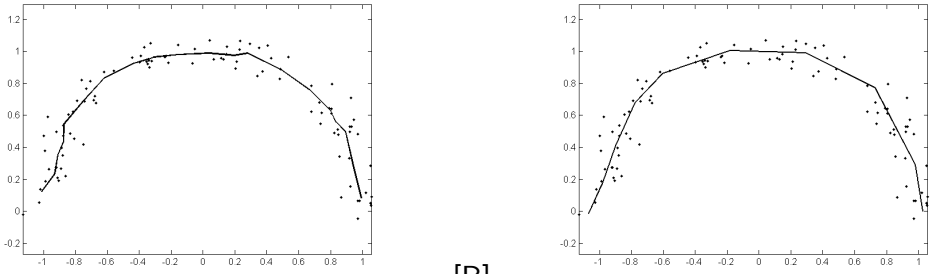


FIGURE 2.7.: 100 observations distribuées autour d'un demi-cercle de rayon 1.

Rappelons que l'algorithme **MS1** calcule le critère $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ pour une table de valeurs de k et ℓ et sélectionne les constantes intervenant dans la pénalité d'après une méthode de la pente bivariée. La Figure 2.8 montre la surface $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ en fonction de $\sqrt{k}$ et ℓ.

Les deux algorithmes ont été appliqués au jeu de données et les courbes principales résultantes sont visibles dans la Figure 2.9. A des fins de comparaison, les

Figures 2.10 et 2.11 montrent quelques courbes obtenues en spécifiant d'autres valeurs pour k et ℓ.



[A] [B]

FIGURE 2.9.: Courbes principales sélectionnées pour les données autour du demi-cercle (n=100). [A] Méthode **MS1** : $\hat{k} = 20$ et $\hat{\ell} = 3$. [B] Méthode **MS2** : $\hat{k} = 9$.

Nous notons que les sorties des deux algorithmes ont approximativement la même qualité. En effet, la courbe principale de **MS1** présente quelques irrégularités non visibles sur le résultat de **MS2**, qui semble cependant plus sommaire, du fait d'une relativement petite valeur de $\hat{k}$.

Les méthodes **MS1** et **MS2** ont aussi été testées sur un ensemble de données plus grand, représenté dans la Figure 2.12. Les résultats pour le jeu de données du demi-cercle avec $n = 250$ sont donnés dans la Figure 2.13. Nous observons que les deux courbes principales obtenues dans cet exemple sont tout à fait correctes.

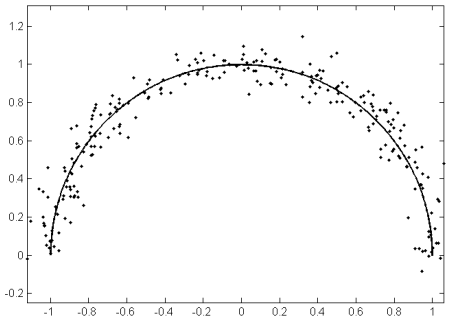


FIGURE 2.12.: 250 observations distribuées autour d'un demi-cercle de rayon 1.

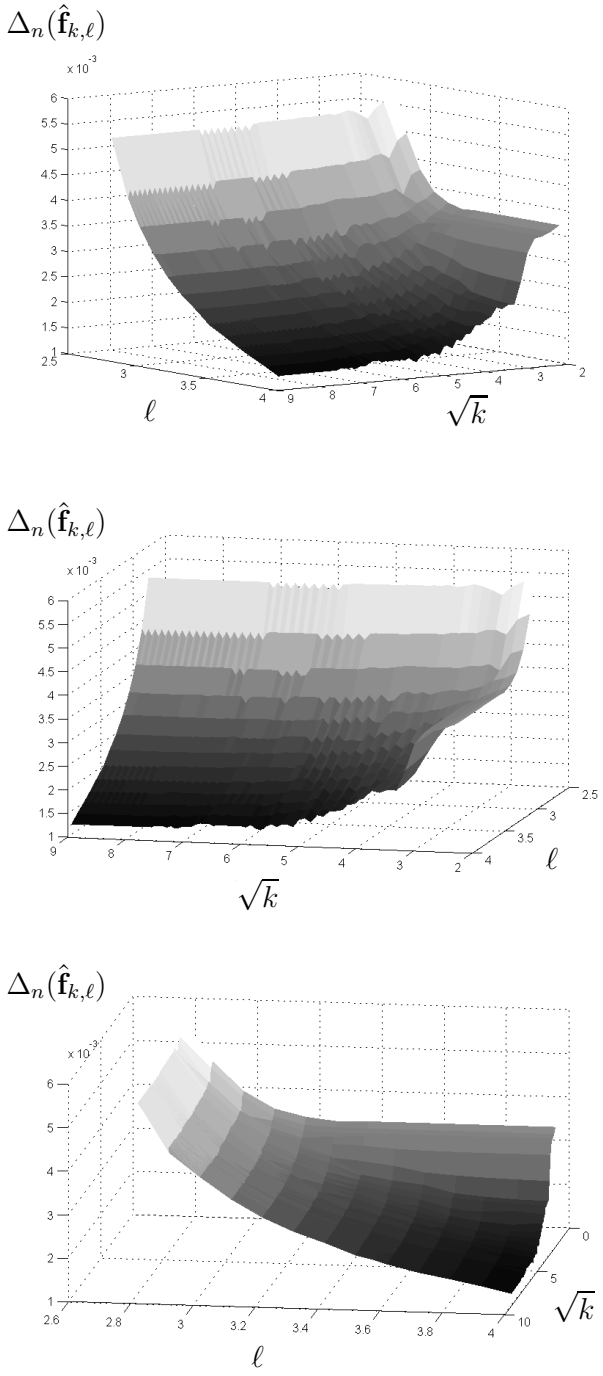


FIGURE 2.8.: Critère $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ en fonction de $\sqrt{k}$ et ℓ pour les données autour du demi-cercle.

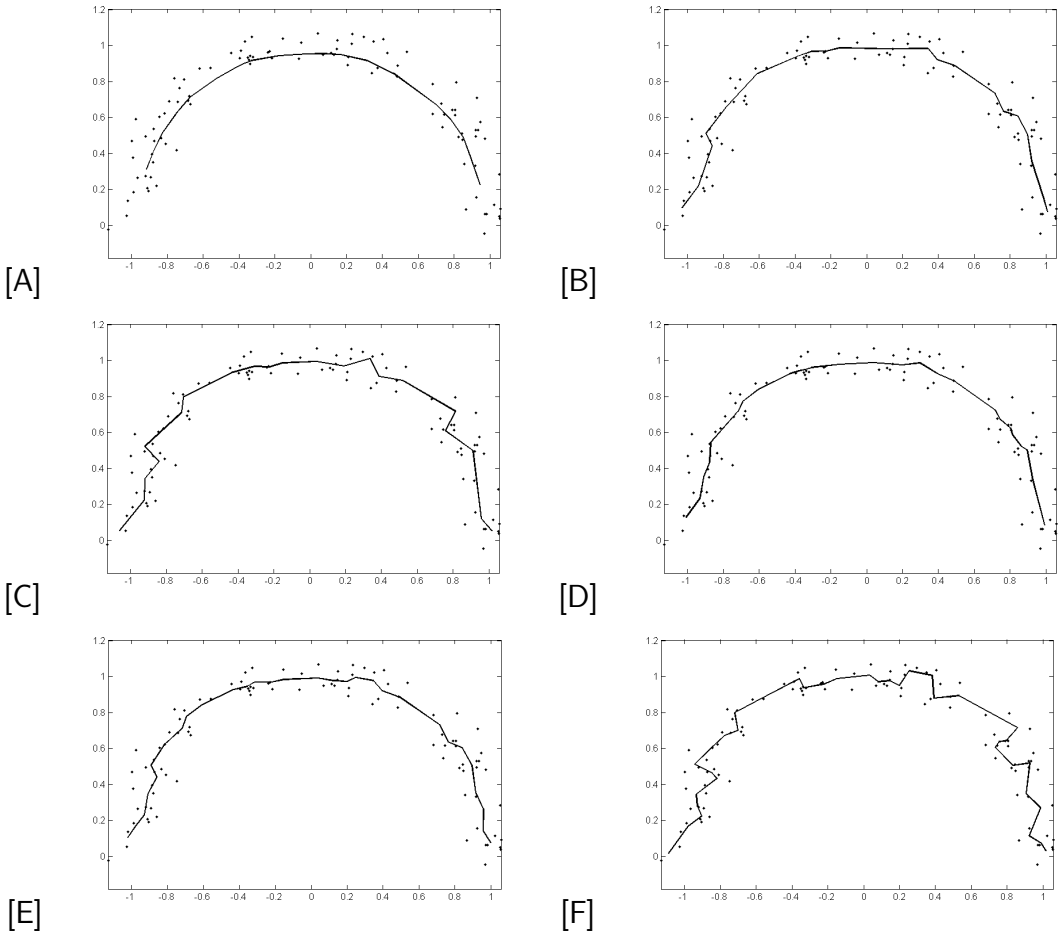


FIGURE 2.10.: Méthode **MS1** : exemples de courbes principales pour quelques valeurs de k et ℓ (n=100). [A] $k = 20$, $\ell = 2.5$. [B] $k = 20$, $\ell = 3.1$. [C] $k = 20$, $\ell = 3.4$. [D] $k = 25$, $\ell = 3$. [E] $k = 30$, $\ell = 3.1$. [F] $k = 35$, $\ell = 4$.

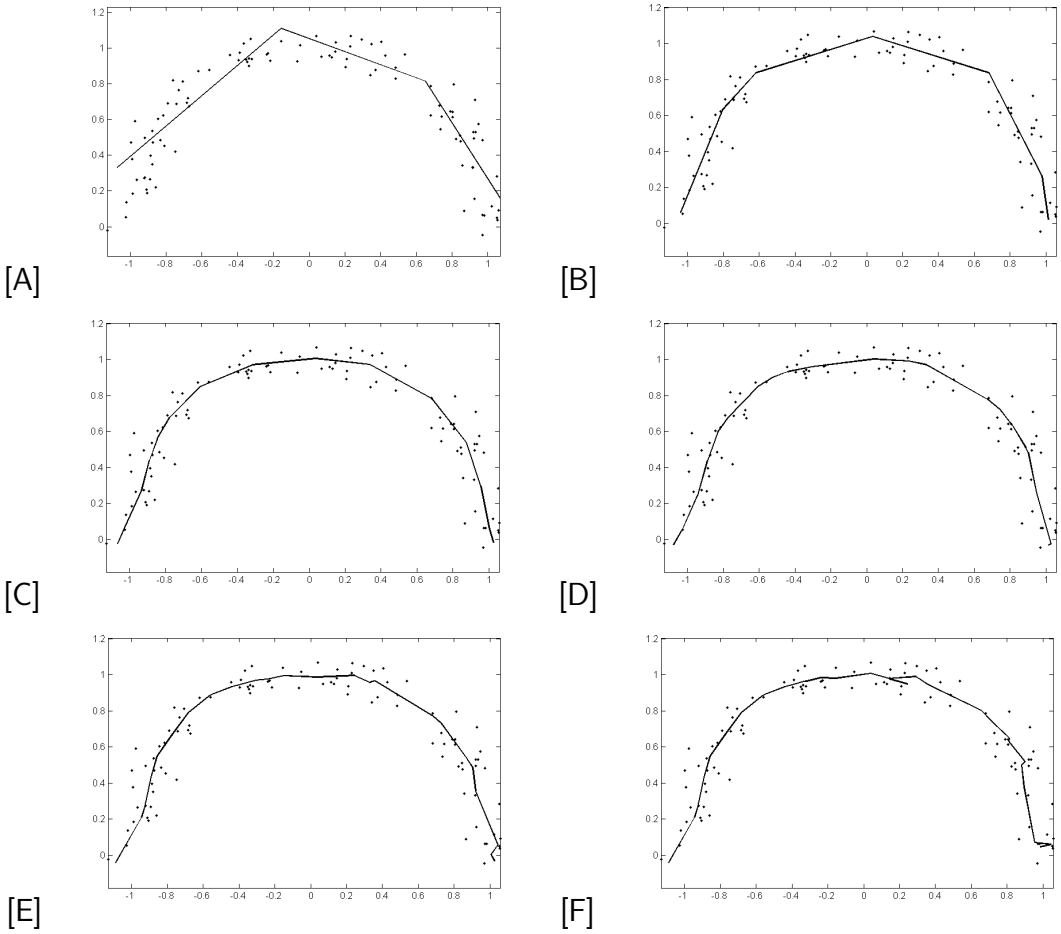
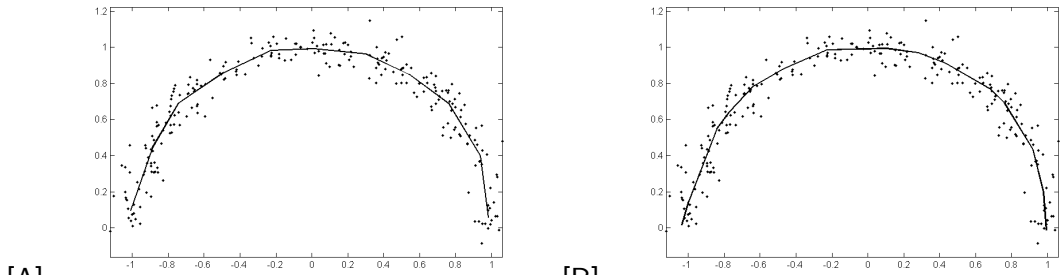


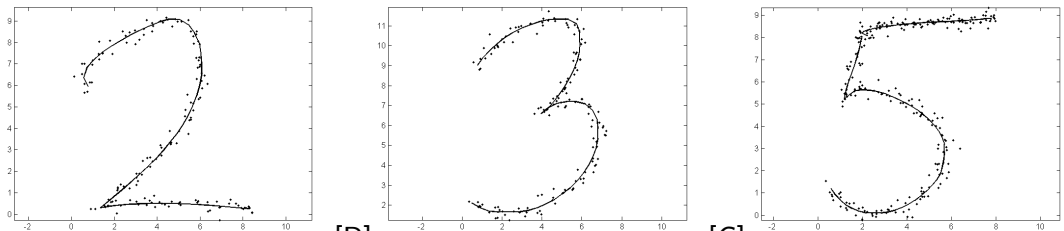
FIGURE 2.11.: Méthode **MS2** : exemples de courbes principales pour quelques valeurs de k (n=100). [A] $k = 3$. [B] $k = 6$. [C] $k = 14$. [D] $k = 20$. [E] $k = 26$. [F] $k = 35$.



[A] [B]

FIGURE 2.13.: Courbes principales sélectionnées pour les données autour du demi-cercle (n=250). [A] Méthode **MS1** : $\hat{k} = 12$, $\hat{\ell} = 3$. [B] Méthode **MS2** : $\hat{k} = 14$.

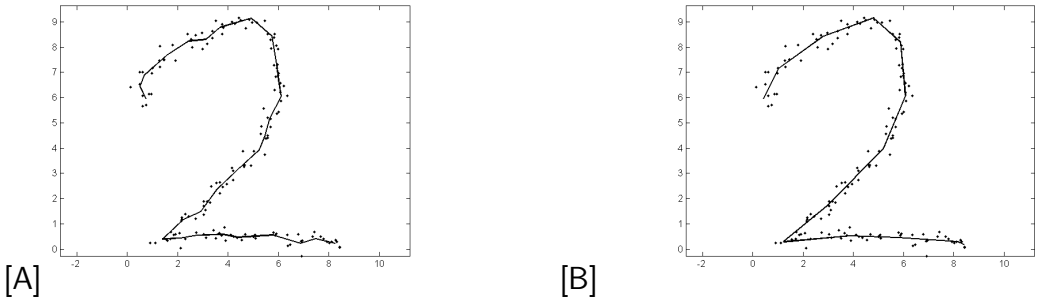
Dans un second ensemble d'exemples, nous avons choisi pour courbes génératives des chiffres, avec un bruit de variance 0.04. Comme le montre la Figure 2.14, 150 observations ont été échantillonnées autour du chiffre 2 et du chiffre 3 et 250 observations autour du chiffre 5.



[A] [B] [C]

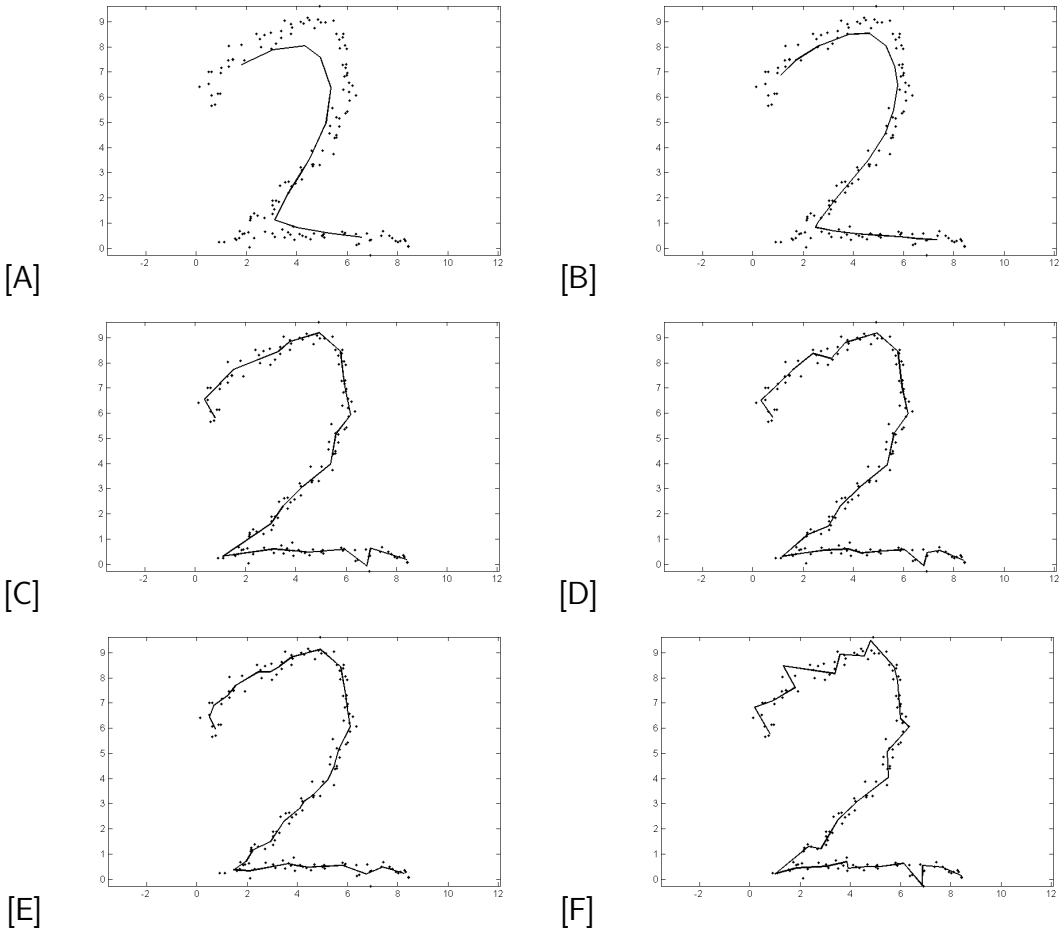
FIGURE 2.14.: [A] 150 observations autour du chiffre 2. [B] 150 observations autour du chiffre 3. [C] 250 observations autour du chiffre 5.

La Figure 2.15 présente le résultat obtenu pour le chiffre 2 avec les deux algorithmes **MS1** et **MS2**, alors que les Figures 2.16 et 2.17 montrent les courbes correspondant à d'autres choix des paramètres.



[A] [B]

FIGURE 2.15.: Courbes principales sélectionnées pour le chiffre 2 (n=150). [A] Méthode **MS1** : $\hat{k} = 27$, $\hat{\ell} = 24$. [B] Méthode **MS2** : $\hat{k} = 12$.



[A] [B] [C] [D] [E] [F]

FIGURE 2.16.: Méthode **MS1** : exemples de courbes principales pour quelques valeurs de k et ℓ (n=150). [A] $k = 12$, $\ell = 14$. [B] $k = 20$, $\ell = 18$. [C] $k = 20$, $\ell = 26$. [D] $k = 27$, $\ell = 26$. [E] $k = 35$, $\ell = 24$. [F] $k = 30$, $\ell = 30$.

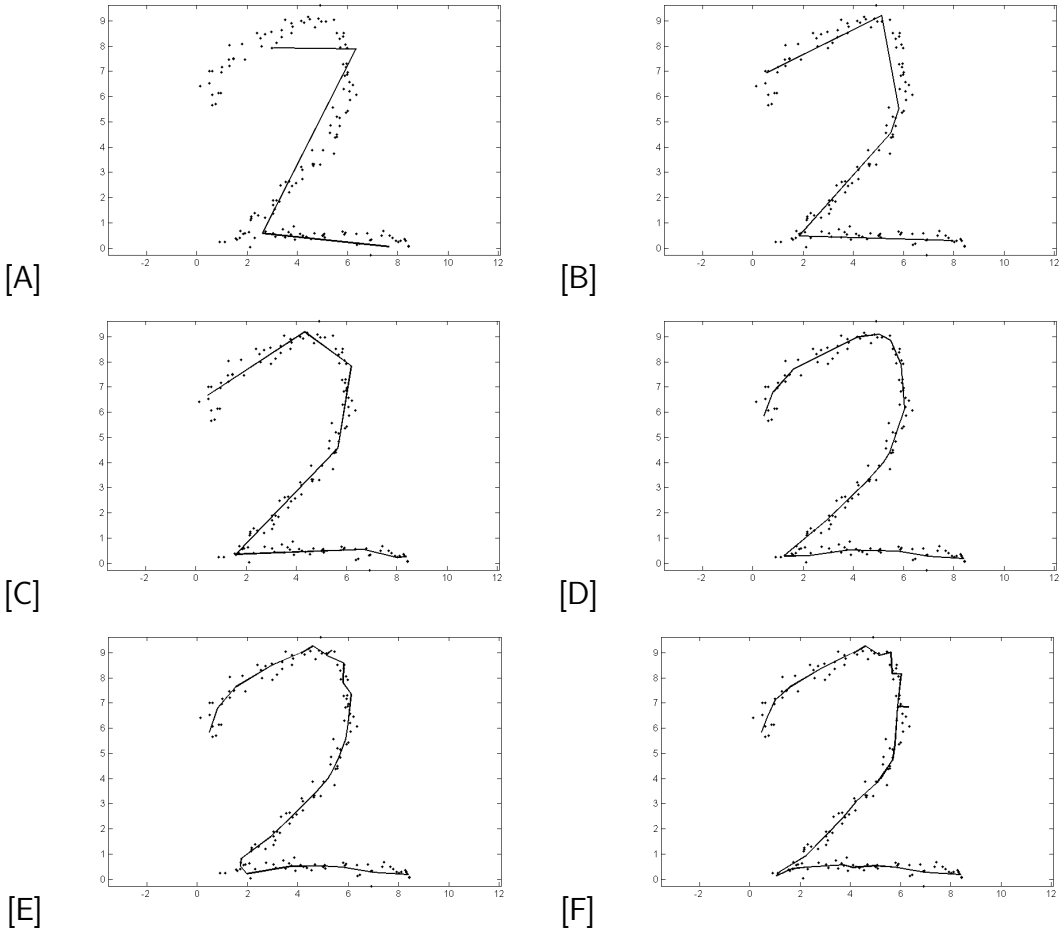
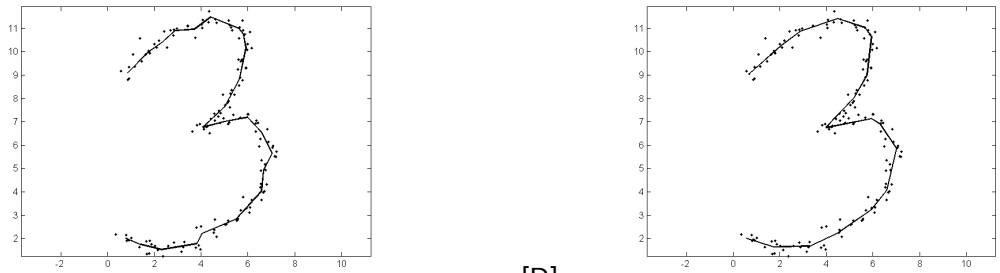


FIGURE 2.17.: Méthode **MS2** : exemples de courbes principales pour quelques valeurs de k (n=150). [A] $k = 3$. [B] $k = 5$. [C] $k = 7$. [D] $k = 20$. [E] $k = 30$. [F] $k = 40$.

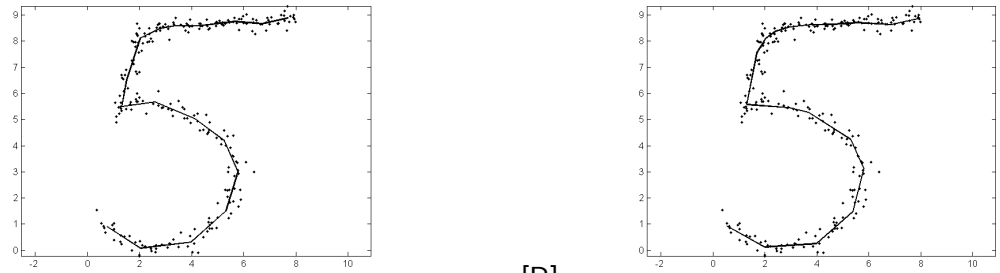
Concernant les données autour du chiffre 2, la courbe principale de **MS1** suit les observations de plus près que ce à quoi l'on s'attendrait. En revanche, la sortie de **MS2** paraît meilleure, bien qu'un $\hat{k}$ légèrement supérieur rendrait sans doute la courbe principale un peu plus régulière.

Les courbes principales ajustées pour les chiffres 3 et 5 sont visibles dans la Figure 2.18 et la Figure 2.19. Pour le chiffre 3, nous observons à nouveau que l'algorithme **MS1** présente une tendance au surapprentissage, tandis que la sortie de **MS2** paraît un peu sommaire. Cependant, les courbes principales obtenues pour le chiffre 5 sont visuellement très satisfaisantes. Sur ce dernier exemple, les deux algorithmes ont abouti à un résultat similaire et il est intéressant de noter qu'ils ont presque sélectionné la même valeur de $\hat{k}$.



[A] [B]

FIGURE 2.18.: Courbes principales sélectionnées pour le chiffre 3 (n=150). [A] Méthode **MS1** : $\hat{k} = 23$, $\hat{\ell} = 23$. [B] Méthode **MS2** : $\hat{k} = 18$.



[A] [B]

FIGURE 2.19.: Courbes principales sélectionnées pour le chiffre 5 (n=250). [A] Méthode **MS1** : $\hat{k} = 17$, $\hat{\ell} = 21$. [B] Méthode **MS2** : $\hat{k} = 18$.

Cette petite étude sur simulations confirme l'importance d'un bon choix des paramètres $\hat{k}$ et $\hat{\ell}$ pour obtenir une courbe principale convenable. Dans l'ensemble, la qualité visuelle des sorties des deux algorithmes est satisfaisante, même si ceux-ci mènent parfois à des résultats quelque peu différents. En fait, les courbes principales ajustées par **MS1** suivent souvent les données d'assez près, en particulier lorsque la taille de l'échantillon n'est pas très grande, alors que les sorties de **MS2** ont tendance à être un peu anguleuses en raison de la sélection d'un $\hat{k}$ plutôt petit. Par ailleurs, notons que du point de vue de la complexité algorithmique, l'exécution de **MS1** est, par construction, moins rapide que celle de **MS2**, puisque le premier des deux algorithmes comprend le calcul du critère $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ pour une gamme de valeurs de la longueur $\ell \leq L$.

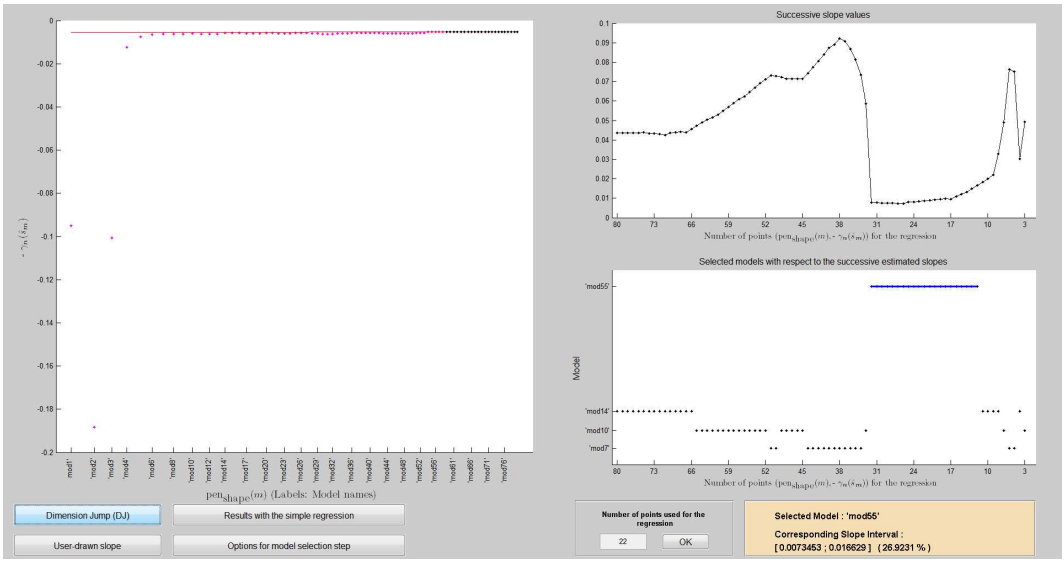
Remarque 2.3.1. Concernant l'heuristique de pente dans la méthode **MS2**, nous avons observé sur ces simulations que le tracé du critère Δ_n en fonction de k n'est pas toujours suffisamment lisse pour obtenir des résultats pertinents avec l'interface CAPUSHE. En effet, par construction du *Polygonal Line Algorithm* de Kégl *et al.* [117], Δ_n ne décroît pas nécessairement d'une valeur de k à la suivante, bien que globalement décroissant. Un certain nombre de points aberrants peuvent ainsi se trouver dans la zone des grandes valeurs de k et perturber l'estimation de la pente. Pour tenter de pallier ce problème, deux stratégies ont été envisagées.

La première idée réside dans le fait de supprimer tous les points du graphe de $-\Delta_n$ correspondant aux valeurs de k telles que $-\Delta_n(\mathbf{f}_k) < -\Delta_n(\mathbf{f}_{k-1})$. Le principal inconvénient de cette approche est qu'elle peut conduire à retirer un nombre de points conséquent. En particulier, toute une zone de valeurs de k autour du bon nombre de segments peut être supprimée, de sorte qu'il devient impossible en pratique de sélectionner un $\hat{k}$ adéquat.

Une autre tentative consiste à effectuer plusieurs fois l'étape d'optimisation des sommets, en choisissant aléatoirement l'ordre dans lequel ils sont successivement actualisés, et à conserver le résultat correspondant au critère le plus faible. Cette méthode permet effectivement de lisser le critère Δ_n, mais induit une variabilité problématique. Nous avons en effet constaté que des courbes principales différentes pouvaient être obtenues avec le même nombre de segments, donnant lieu à des résultats plus ou moins satisfaisants.

Aucune de ces deux stratégies ne résout entièrement le problème. Néanmoins, un tracé de la pente « à l'oeil nu » apporte la plupart du temps un résultat convenable, comme l'illustre la Figure 2.20, qui présente un exemple de sorties de CAPUSHE pour la méthode d'estimation directe de la pente et la méthode du saut de dimension, et la Figure 2.21, correspondant au tracé « visuel » de la pente. Sur cet exemple, cette dernière méthode donne le même résultat que le saut de dimension, résultat différent de celui obtenu par l'estimation automatique de la pente.

[A]



[B]

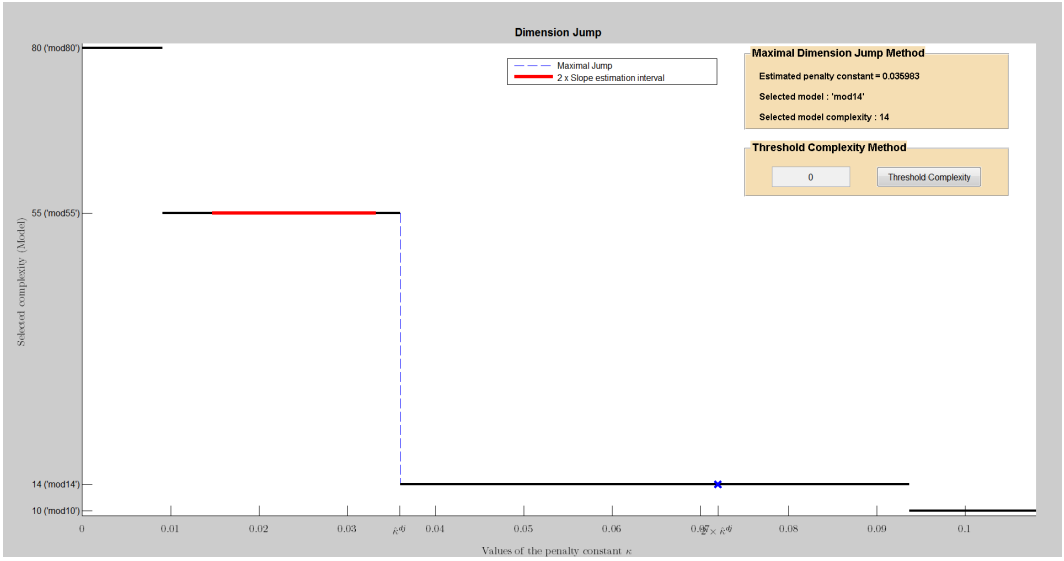


FIGURE 2.20.: Exemple de sorties de CAPUSHE pour les données autour du demi-cercle (n=250). [A] Estimation de la pente : $\hat{k} = 55$. [B] Saut de dimension : $\hat{k} = 14$.

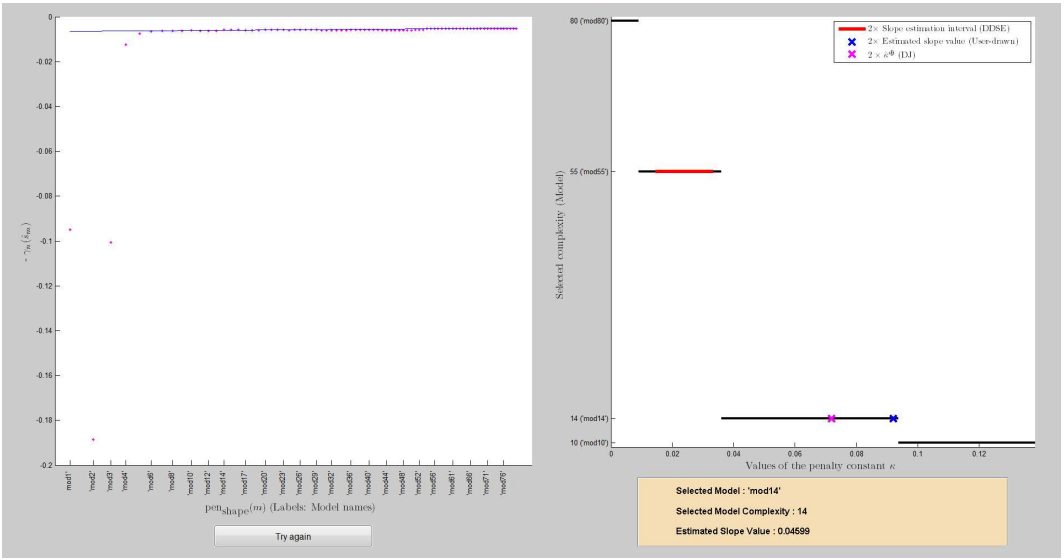


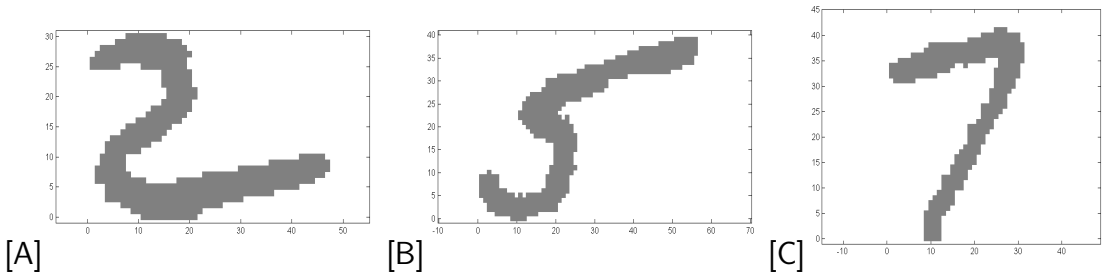
FIGURE 2.21.: Exemple de sortie de CAPUSHE pour les données autour du demi-cercle (n=250). Tracé visuel de la pente : $\hat{k} = 14$.

2.3.2. Données réelles

Chiffres de la base de données NIST

Le premier jeu de données réel utilisé dans cette seconde série d'expériences a pour origine la base de données NIST Special Database 19 (`http://www.nist.gov/srd/nistsd19.cfm`), contenant des caractères manuscrits provenant de 3600 personnes. Les données consistent en des images binaires scannées à 11.8 points par millimètre (300 dpi). Chacune de ces images peut être vue comme un ensemble d'observations dans $\mathbb{R}^2$ réparties uniformément dans une zone correspondant à l'épaisseur du trait de stylo. Comme nous l'avons mentionné dans le Chapitre 1, déterminer l'axe médian de tels caractères manuscrits constitue souvent une étape préliminaire dans le domaine de la reconnaissance de caractères (voir aussi Deutsch [69] et Alcorn et Hoggar [6]).

Les algorithmes **MS1** et **MS2** ont été appliqués aux trois chiffres de la base de données NIST visibles dans la Figure 2.22. La Figure 2.23 montre la surface correspondant au tracé du critère $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ en fonction de $\sqrt{k}$ et ℓ. Les courbes principales obtenues pour ce chiffre avec les deux méthodes sont représentées dans la Figure 2.24, tandis que les Figures 2.25 et 2.26 montrent quelques résultats pour d'autres valeurs des paramètres.



[A] [B] [C]

FIGURE 2.22.: Trois chiffres manuscrits de la base de données NIST.

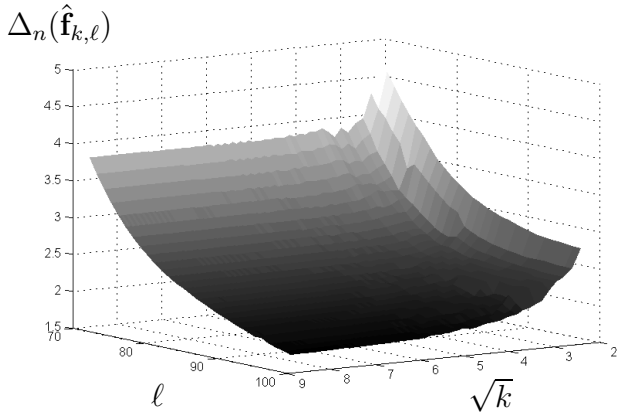
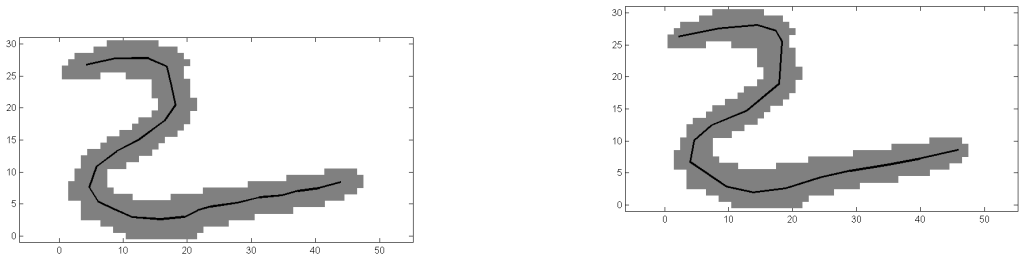
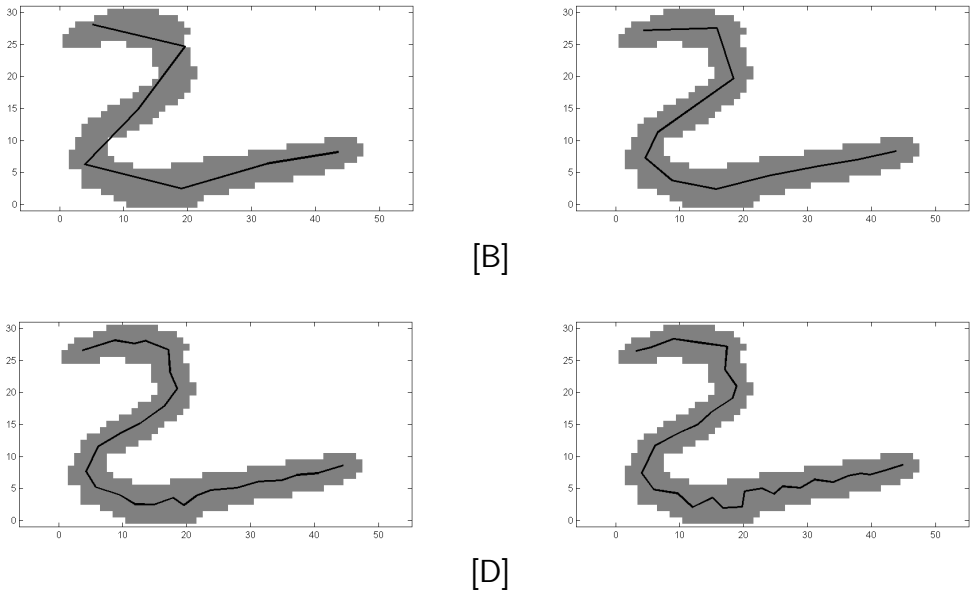


FIGURE 2.23.: Critère $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ en fonction de $\sqrt{k}$ et ℓ pour le chiffre 2 de la base de données NIST (n=458).



[A] [B]

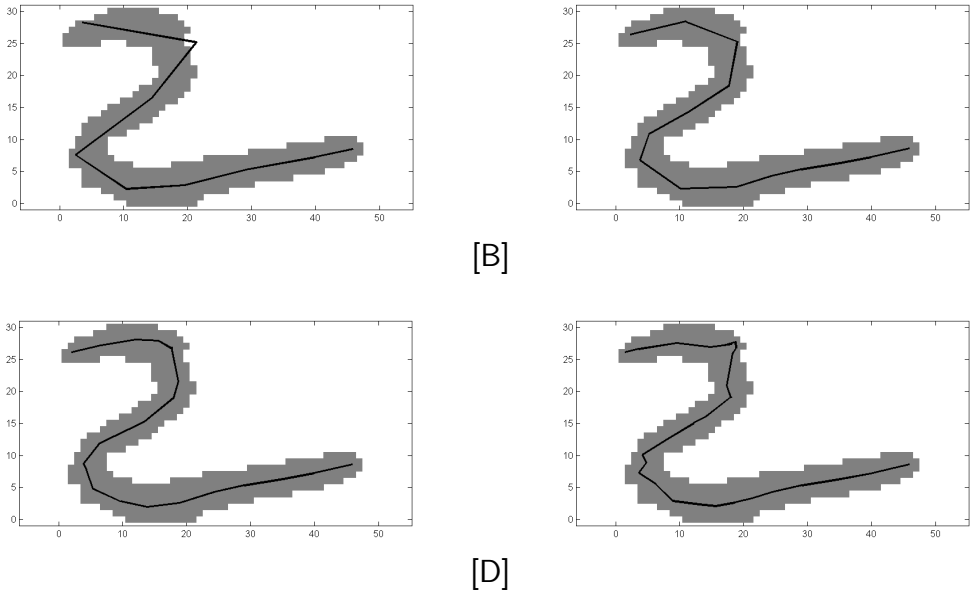
FIGURE 2.24.: Courbes principales sélectionnées pour le chiffre 2 (n=458). [A] Méthode **MS1** : $\hat{k} = 23$, $\hat{\ell} = 80$. [B] Méthode **MS2** : $\hat{k} = 17$.



[A] [B]

[C] [D]

FIGURE 2.25.: Méthode **MS1** : exemples de courbes principales pour quelques valeurs de k et ℓ. [A] $k = 6$, $\ell = 80$. [B] $k = 10$, $\ell = 80$. [C] $k = 25$, $\ell = 84$. [D] $k = 30$, $\ell = 90$.



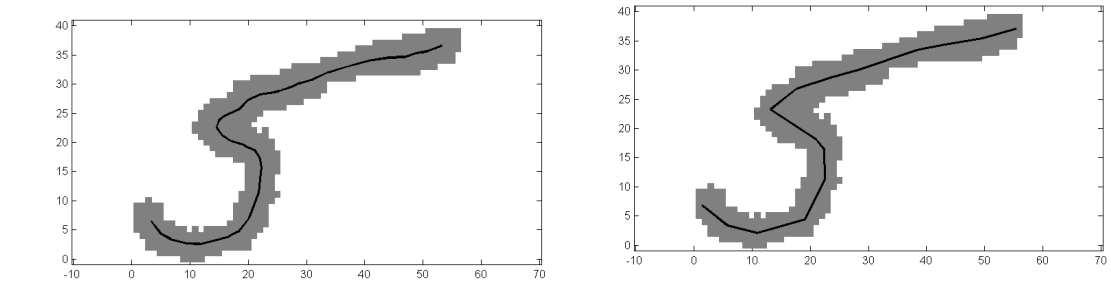
[A] [B]

[C] [D]

FIGURE 2.26.: Méthode **MS2** : exemples de courbes principales pour quelques valeurs de k. [A] $k = 8$. [B] $k = 13$. [C] $k = 20$. [D] $k = 30$.

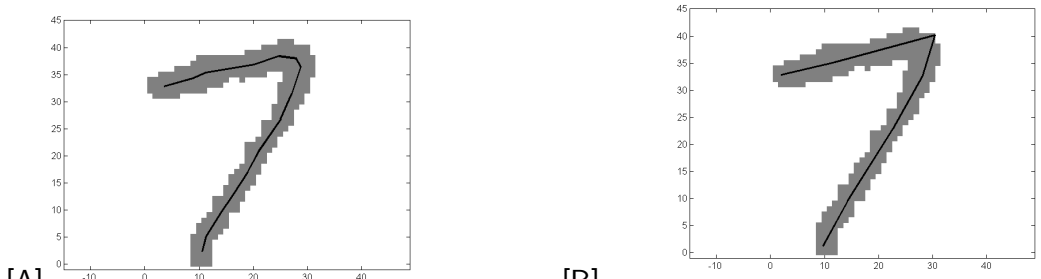
Nous observons que les deux résultats pour le chiffre 2 sont assez similaires, avec cependant un léger avantage de **MS1**. En effet, cet algorithme donne lieu à une courbe plus régulière qui retrouve plus précisément la boucle du chiffre 2.

Les Figures 2.27 et 2.28 montrent les sorties des algorithmes **MS1** et **MS2** pour les chiffres NIST 5 et 7.



[A] [B]

FIGURE 2.27.: Courbes principales sélectionnées pour le chiffre 5 (n=513). [A] Méthode **MS1** : $\hat{k} = 38$, $\hat{\ell} = 82$. [B] Méthode **MS2** : $\hat{k} = 14$.



[A] [B]

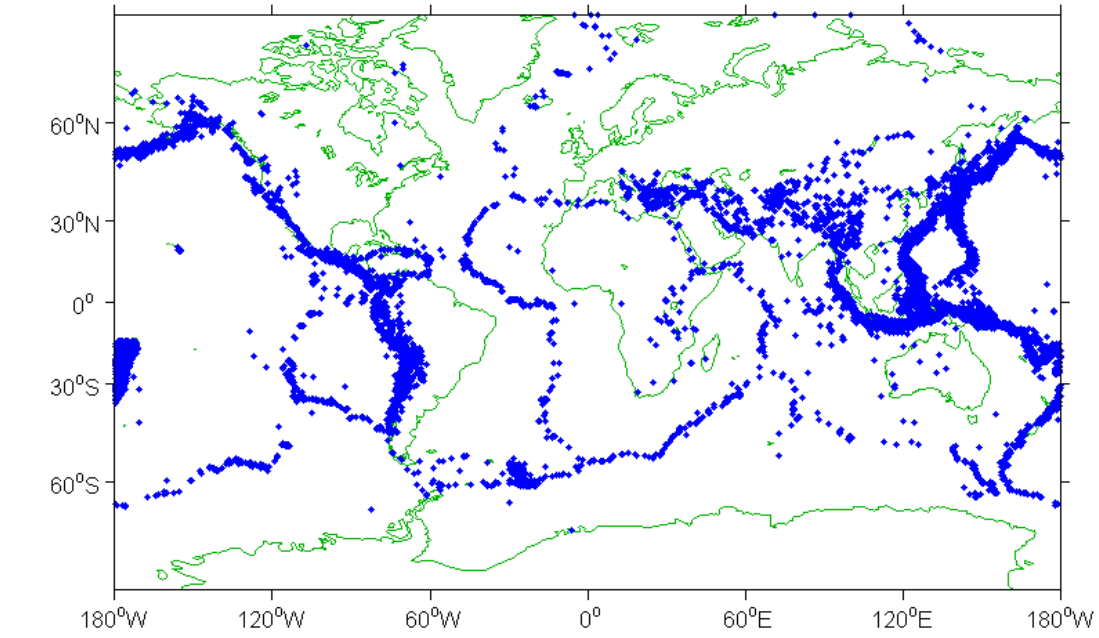
FIGURE 2.28.: Courbes principales sélectionnées pour le chiffre 7 (n=334). [A] Méthode **MS1** : $\hat{k} = 15$, $\hat{\ell} = 66$. [B] Méthode **MS2** : $\hat{k} = 6$.

Finalement, sur ces jeux de données de chiffres NIST, nous avons constaté que les deux méthodes fonctionnent plutôt bien. Ici, **MS1** ne semble pas provoquer de surapprentissage, probablement parce que la taille de l'échantillon est suffisamment grande. Comme dans les exemples avec données simulées, il apparaît néanmoins que les courbes estimées par l'algorithme **MS2** pourraient être plus régulières avec un nombre $\hat{k}$ de segments un peu plus grand.

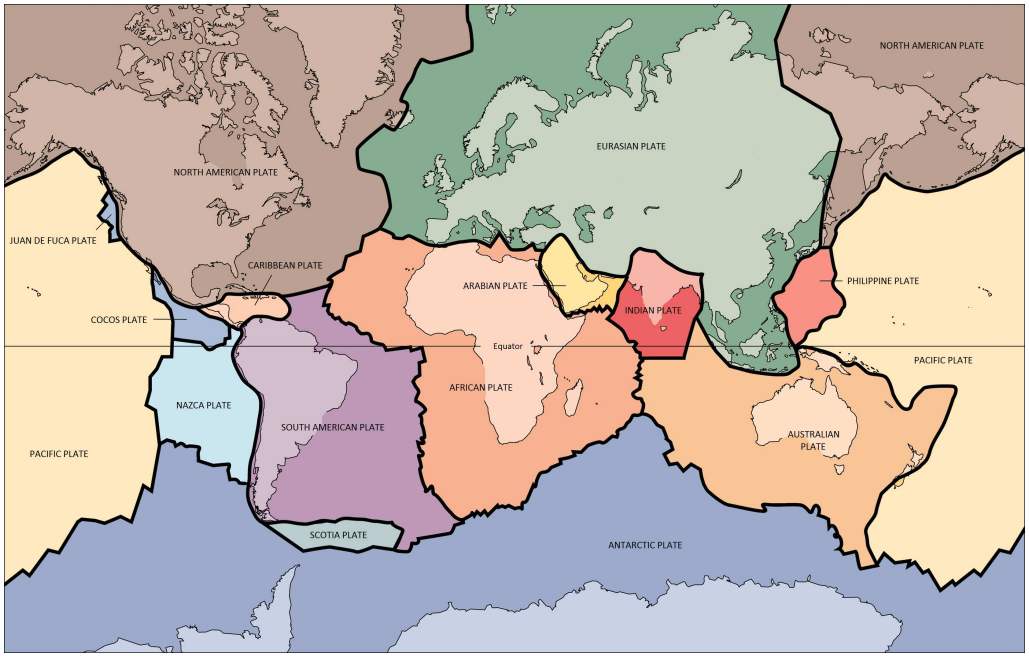
Données sismiques

Avec les images satellite, la localisation des séismes enregistrés à la surface du globe constitue en géologie la seconde source d'information essentielle dans la connaissance des failles sismiques, qu'elles soient liées à un mécanisme de subduction ou d'accrétion. Un aperçu du lien entre zones sismiques et limites de plaques est donné par la Figure 2.29, qui représente des impacts sismiques dans le monde — la carte est dessinée en utilisant la projection de Miller —, ainsi qu'une carte du monde de l'UGS (*United States Geological Survey*) montrant les différentes plaques lithosphériques. Le jeu de données, qui peut être téléchargé sur la page de l'USGS (`http://earthquake.usgs.gov/research/data/centennial.php`), provient du *Centennial Catalog*, qui liste de grands tremblements de terre enregistrés depuis 1900 (Engdahl et Villaseñor [83]) Nous utilisons ici les algorithmes **MS1** et **MS2** pour retrouver le tracé de limites de plaques lithosphériques à partir des données de localisation de séismes de la Figure 2.29.

Nous considérons deux zones d'activité sismique. La première (notée dans la suite **Z1**) est située dans l'océan Atlantique, à l'ouest du continent africain (de 60°S 50°W à 40°N 0° environ), et la seconde (**Z2**) va du sud de l'Afrique jusqu'au sud de l'Australie (de 65°S 0° à 25°S 160°E environ). Ces deux zones sont localisées sur la carte du monde dans la Figure 2.30. La Figure 2.31 montre les résultats de courbes principales pour **Z1** et la Figure 2.32 ceux pour **Z2**.



[A]



[B]

FIGURE 2.29.: [A] Impacts sismiques. [B] Plaques lithosphériques.

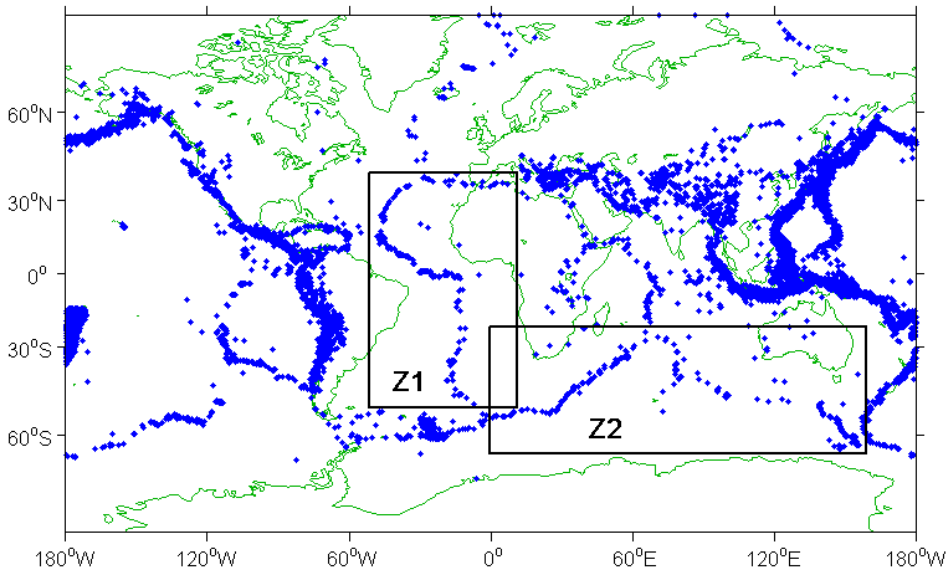


FIGURE 2.30.: Localisation des deux zones sismiques considérées **Z1** (de 60°S 50°W à 40°N 0°) et **Z2** (de 65°S 0° à 25°S 160°E).

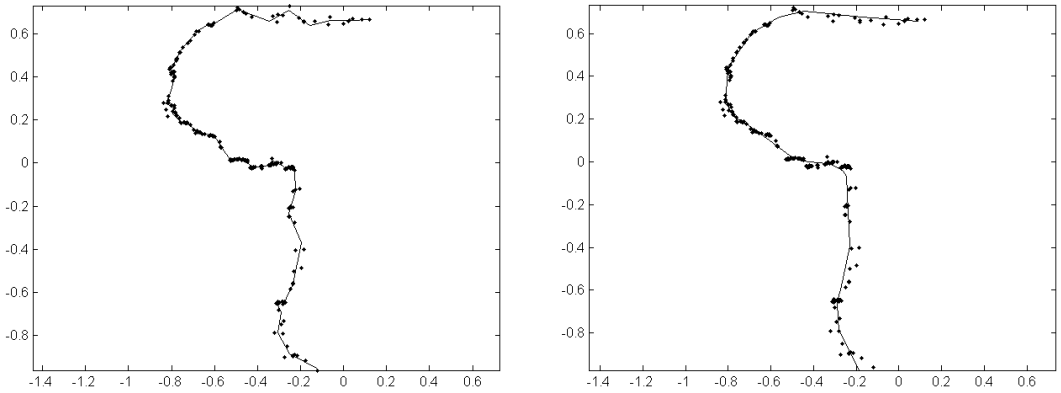


FIGURE 2.31.: Courbes principales sélectionnées pour la zone sismique **Z1** (n=252). [A] Méthode **MS1** : $\hat{k} = 55$, $\hat{\ell} = 31$. [B] Méthode **MS2** : $\hat{k} = 30$.

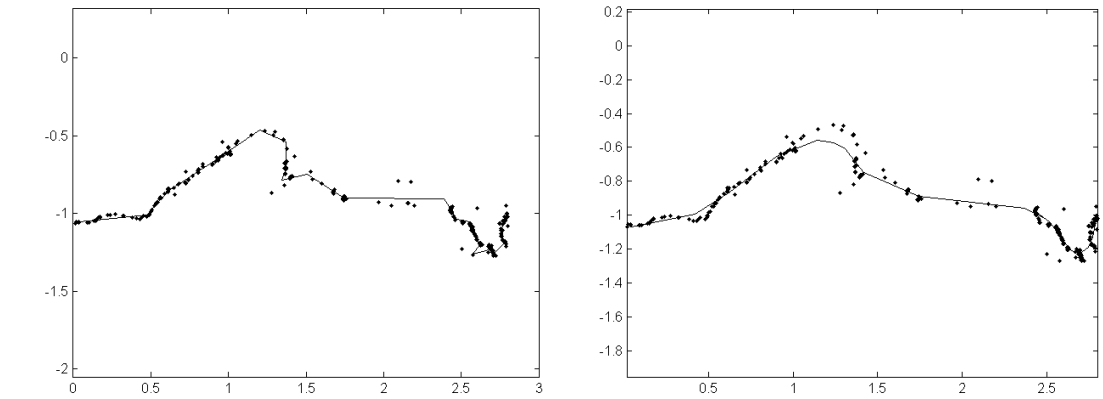


FIGURE 2.32.: Courbes principales sélectionnées pour la zone sismique **Z2** (n=322). [A] Méthode **MS1** : $\hat{k} = 22$, $\hat{\ell} = 38$. [B] Méthode **MS2** : $\hat{k} = 20$.

Dans la Figure 2.31, nous constatons, pour la zone sismique **Z1**, que la méthode **MS1** produit une nouvelle fois une courbe principale suivant les données d'assez près. L'algorithme **MS2** donne au contraire une courbe plus régulière, qui paraît plus satisfaisante à première vue. Cependant, la limite de la plaque lithosphérique a plus de chances de ressembler à la courbe principale plus irrégulière obtenue avec **MS1**, comme le suggère la Figure 2.29 [B]. La même observation vaut pour **Z2** (Figure 2.32). En outre, dans ce cas, la sortie de **MS2** ne retrouve pas la forme de la limite de la plaque, qui passe certainement le long des quelques observations situées les plus au nord et non plusieurs degrés au sud. La pénalité locale sur les angles est apparemment trop forte dans ce contexte. Par conséquent, sur ces données sismiques, les résultats de **MS1** semble plus pertinents.

Il est intéressant de noter que l'utilisation de ce type de données de tremblement de terre pour le tracé de failles pourrait en particulier s'avérer très utile dans le but de localiser certaines failles peu repérables, mais qu'il convient de surveiller de près dans le cadre de la prévention des risques sismiques. Harding et Berghoff [100], qui emploient une méthode basée sur la cartographie par laser aéroporté étudient par exemple les risques sismiques dans une zone densément recouverte de végétation, située dans le *Puget Lowland* (Etats-Unis, Washington).

2.4. Preuves des résultats de la Section 2.1

2.4.1. Preuve du Lemme 2.1.1

Posons $c = FG/2$. Notons que $\ell > 2c$. Dans un repère orthonormal de $\mathbb{R}^d$ bien choisi, F a pour coordonnées $(-c, 0, \ldots, 0)$ et G $(c, 0, \ldots, 0)$. Une courbe de longueur ℓ, d'extrémités F et G, est incluse dans l'ensemble délimité par les points $M(x_1, \ldots, x_d)$ tels que

$$MF + MG = \ell.$$

Montrons qu'il s'agit d'un ellipsoïde de premier axe principal de longueur ℓ, les autres axes étant de longueur $\sqrt{\ell^2 - FG^2}$. Soit donc $M(x_1, \ldots, x_d)$ tel que $MF + MG = \ell$. Alors,

$$MF^2 = (x_1 + c)^2 + \sum_{j=2}^{d} x_j^2$$

et

$$MG^2 = (x_1 - c)^2 + \sum_{j=2}^{d} x_j^2.$$

Ainsi

$$MF - MG = \frac{MF^2 - MG^2}{MF + MG} = \frac{(x_1 + c)^2 - (x_1 - c)^2}{\ell} = \frac{4x_1 c}{\ell}.$$

Il en résulte d'une part l'égalité

$$(MF + MG)^2 + (MF - MG)^2 = \ell^2 + \frac{16 x_1^2 c^2}{\ell^2},$$

et d'autre part

$$(MF + MG)^2 + (MF - MG)^2 = 2(MF^2 + MG^2) = 4\sum_{j=1}^{d} x_j^2 + 4c^2.$$

D'où

$$\ell^2 + \frac{16 x_1^2 c^2}{\ell^2} = 4\sum_{j=1}^{d} x_j^2 + 4c^2,$$

ce qui se récrit

$$x_1^2 \left(1 - \frac{4c^2}{\ell^2}\right) + \sum_{j=2}^{d} x_j^2 = \frac{\ell^2}{4} - c^2, \tag{2.11}$$

ou encore

$$\frac{x_1^2}{\ell^2/4} + \sum_{j=2}^{d} \frac{x_j^2}{\ell^2/4 - c^2} = 1, \tag{2.12}$$

où $\ell^2/4 - c^2 > 0$ puisque $\ell > 2c$. En d'autres termes, le point M appartient à un ellipsoïde ayant un axe de longueur ℓ et $d-1$ axes de longueur $2\sqrt{\ell^2/4 - c^2} = \sqrt{\ell^2 - FG^2}$.

Réciproquement, si $M(x_1, \dots, x_d)$ vérifie l'équation (2.11), avec $\dfrac{\ell^2}{4} - c^2 > 0$, alors

$$\begin{aligned} MF^2 &= (x_1 + c)^2 + \sum_{j=2}^{d} x_j^2 \\ &= (x_1 + c)^2 + \frac{\ell^2}{4} - c^2 - x_1^2 + \frac{4x_1^2 c^2}{\ell^2} \\ &= 2x_1 c + \frac{\ell^2}{4} + \frac{4x_1^2 c^2}{\ell^2} \\ &= \left(\frac{\ell}{2} + \frac{2x_1 c}{\ell} \right)^2, \end{aligned}$$

d'où $MF = \left| \dfrac{\ell}{2} + \dfrac{2x_1 c}{\ell} \right|$. De même, $MG = \left| \dfrac{\ell}{2} - \dfrac{2x_1 c}{\ell} \right|$. Or, $|x_1| \leq \dfrac{\ell^2}{4c}$, car sinon $\dfrac{x_1^2}{\ell^2/4} > \dfrac{\ell^2}{4c^2} > 1$, ce qui contredit l'équation (2.12). Finalement, $MF + MG = \dfrac{\ell}{2} + \dfrac{2x_1 c}{\ell} + \dfrac{\ell}{2} - \dfrac{2x_1 c}{\ell} = \ell$.

2.4.2. Preuve du Lemme 2.1.2

Nous cherchons tout d'abord à calculer le nombre de recouvrement d'un ellipsoïde $\mathcal{E}_\ell$ de $\mathbb{R}^d$ de foyers F et G. Le lemme suivant est un cas particulier de la Proposition 5 de von Luxburg, Bousquet et Schölkopf [186].

Lemme 2.4.1. *Supposons que $a \geq b \geq \varepsilon$. Le nombre de boules de rayon ε nécessaires pour recouvrir $\mathcal{E}_\ell$, ellipsoïde de dimension d d'axes principaux $a, b, \dots, b$, vérifie*

$$\mathcal{N}(\mathcal{E}_\ell, \|\cdot\|, \varepsilon) \leq \left(\frac{2}{\varepsilon} \right)^d a b^{d-1}.$$

Preuve du Lemme 2.4.1. Le nombre de boules de rayon ε nécessaires pour recouvrir $\mathcal{E}_\ell$ vérifie

$$\mathcal{N}(\mathcal{E}_\ell, \|\cdot\|, \varepsilon) \leq \left(\left\lfloor \frac{a}{\varepsilon} \right\rfloor + 1 \right) \left(\left\lfloor \frac{b}{\varepsilon} \right\rfloor + 1 \right)^{d-1},$$

où $\lfloor y \rfloor$ désigne la partie entière de y. En effet, l'ellipsoïde $\mathcal{E}_\ell$ est inscrit dans un parallélépipède de côtés de longueur $a, b, \ldots, b$. Or, le nombre de boules de rayon ε nécessaires pour recouvrir un parallélépipède de côtés de longueur $c_1, \ldots, c_d$ est

$$\prod_{j=1}^{d} \left(\left\lfloor \frac{c_j}{\varepsilon} \right\rfloor + 1 \right).$$

Comme par hypothèse $a \geq b \geq \varepsilon$, on a $\left\lfloor \frac{a}{\varepsilon} \right\rfloor + 1 \leq \frac{2a}{\varepsilon}$ et $\left\lfloor \frac{b}{\varepsilon} \right\rfloor + 1 \leq \frac{2b}{\varepsilon}$, d'où

$$\mathcal{N}(\mathcal{E}_\ell, \|\cdot\|, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^d ab^{d-1}.$$

□

D'après le lemme 2.4.1, nous savons que

$$\mathcal{N}(\mathcal{E}_\ell, \|\cdot\|, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^d ab^{d-1}.$$

Soit $\mathcal{U}$ une collection d'au plus $\left(\frac{2}{\varepsilon}\right)^d ab^{d-1}$ centres de boules de $(\mathbb{R}^d, \|\cdot\|)$ correspondant à un ε-recouvrement de $\mathcal{E}_\ell$. Pour chaque vecteur $\overrightarrow{\mathbf{u}} = {}^t({}^t\mathbf{u}_1, \ldots, {}^t\mathbf{u}_n) \in \mathbb{R}^{nd}$, où les $\mathbf{u}_i$ sont des éléments de $\mathcal{U}$, on a $\prod_{i=1}^{n} B(\mathbf{u}_i, \varepsilon) \subset B(\overrightarrow{\mathbf{u}}, \varepsilon)$ (boules pour la norme normalisée de $\mathbb{R}^d$ et de $\mathbb{R}^{nd}$ respectivement). Par conséquent,

$$\mathcal{N}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^{nd} (ab^{d-1})^n.$$

Remarque 2.4.1. Il n'est pas possible de s'affranchir de l'exposant n dans ce calcul de nombre de recouvrement. En effet, si l'on considère une seule courbe $\mathbf{f}$ de longueur ℓ, recouverte par N boules de $\mathbb{R}^d$, il faut N^n boules de $\mathbb{R}^{nd}$ pour recouvrir $(\mathcal{I}_\mathbf{f})^n$, puisque chacun des n points considérés sur la courbe peut se trouver dans n'importe laquelle des N boules. Cependant, notre majoration du nombre de boules de $\mathbb{R}^{nd}$ nécessaires pour recouvrir $\mathcal{C}_\ell$ est probablement trop large. En effet, dans la mesure où les n points sont contraints à être situés sur la même courbe de longueur ℓ, nous n'aurions pas besoin d'utiliser toutes les boules de $\mathbb{R}^{nd}$ obtenues en combinant les centres des boules de $\mathbb{R}^d$ recouvrant $\mathcal{E}_\ell$. On pourrait sans aucun doute obtenir une meilleure majoration en résolvant le problème combinatoire consistant à dénombrer toutes les combinaisons admissibles de boules de $\mathbb{R}^d$ pour une classe $\mathcal{C}_\ell$.

2.4.3. Preuve du Lemme 2.1.3

Commençons par un lemme technique qui nous sera utile dans les majorations d'intégrale.

Lemme 2.4.2. *Pour $x \in]0,1]$,*

$$\int_0^x \sqrt{\ln \frac{1}{t}}\mathrm{d}t \leq x\left(\sqrt{\ln \frac{1}{x}} + \sqrt{\pi}\right).$$

Preuve du Lemme 2.4.2. On a

$$\begin{aligned}\int_0^x \sqrt{\ln \frac{1}{t}}\mathrm{d}t &= \left[t\sqrt{\ln 1/t}\right]_0^x + \int_0^x \frac{1}{2\sqrt{\ln 1/t}}\mathrm{d}t \\ &= x\sqrt{\ln 1/x} + \frac{1}{\sqrt{2}}\int_{\sqrt{2\ln 1/x}}^{+\infty} e^{-u^2/2}\mathrm{d}u \\ &\leq x(\sqrt{\ln 1/x} + \sqrt{\pi}).\end{aligned}$$

Cette inégalité provient du fait que, pour $a \geq 0$,

$$\frac{1}{\sqrt{2\pi}}\int_a^{+\infty} e^{-u^2/2}\mathrm{d}u \leq e^{-a^2/2}. \tag{2.13}$$

En effet, si g désigne la fonction définie par $g(a) = e^{-a^2/2} - \frac{1}{\sqrt{2\pi}}\int_a^{+\infty} e^{-u^2/2}\mathrm{d}u$, alors $g'(a) = e^{-a^2/2}(\frac{1}{\sqrt{2\pi}} - a)$. Ainsi, g est strictement croissante sur $[0, 1/\sqrt{2\pi}]$ et décroissante sur $[1/\sqrt{2\pi}, +\infty]$. Comme $g(0) = 1/2$ et $\lim_{+\infty} g = 0$, on obtient $g(a) \geq 0$ pour tout $a \geq 0$, d'où l'inégalité (2.13). □

Revenons à la preuve du Lemme 2.1.3. D'après le Lemme 2.1.2, l'entropie métrique $\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)$ vérifie

$$\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon) \leq nd \ln\left(\frac{2a^{1/d}b^{1-1/d}}{\varepsilon}\right).$$

Si $r \leq b$,

$$\begin{aligned}\phi_\ell(r) &= \kappa\int_0^r \sqrt{\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)}\mathrm{d}\varepsilon \\ &\leq \kappa\sqrt{nd}\int_0^r \sqrt{\ln\left(\frac{2a^{1/d}b^{1-1/d}}{\varepsilon}\right)}\mathrm{d}\varepsilon.\end{aligned}$$

Le changement de variables $t = \frac{\varepsilon}{2a^{1/d}b^{1-1/d}}$ donne

$$\phi_\ell(r) \leq 2\kappa\sqrt{nd}a^{1/d}b^{1-1/d}\int_0^{\frac{r}{2a^{1/d}b^{1-1/d}}}\sqrt{\ln\frac{1}{t}}\mathrm{d}t.$$

Or, le Lemme 2.4.2 indique que pour $x \in]0,1]$,

$$\int_0^x \sqrt{\ln\frac{1}{t}}\mathrm{d}t \leq x\left(\sqrt{\ln\frac{1}{x}} + \sqrt{\pi}\right).$$

Donc

$$\phi_\ell(r) \leq \kappa r\sqrt{nd}\left(\sqrt{\ln\left(\frac{2a^{1/d}b^{1-1/d}}{r}\right)} + \sqrt{\pi}\right).$$

Pour $r \leq b$, soit

$$\varphi_\ell(r) = \kappa r\sqrt{nd}\left(\sqrt{\ln\left(\frac{2a^{1/d}b^{1-1/d}}{r}\right)} + \sqrt{\pi}\right).$$

Si $r \geq b$,

$$\begin{aligned}
\phi_\ell(r) &= \kappa\int_0^r \sqrt{\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)}\mathrm{d}\varepsilon \\
&= \kappa\int_0^b \sqrt{\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)}\mathrm{d}\varepsilon + \kappa\int_b^r \sqrt{\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)}\mathrm{d}\varepsilon \\
&\leq \phi_\ell(b) + (r-b)H(b) \\
&\leq \varphi_\ell(b) + (r-b)\varphi'_\ell(b),
\end{aligned}$$

où $H(b) = \kappa\sqrt{\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, b)}$. En effet, $\varphi_\ell(b) \geq \phi_\ell(b)$ par définition de φ_ℓ, et d'autre part

$$\begin{aligned}
\frac{\varphi'_\ell(b) - H(b)}{\kappa\sqrt{nd}} &\geq \sqrt{\ln\left[2\left(\frac{a}{b}\right)^{1/d}\right]} + \sqrt{\pi} - \frac{1}{2}\left(\ln\left[2\left(\frac{a}{b}\right)^{1/d}\right]\right)^{-1/2} - \sqrt{\ln\left[2\left(\frac{a}{b}\right)^{1/d}\right]} \\
&\geq \sqrt{\pi} - \frac{1}{2}\left(\ln\left(2\left(\frac{a}{b}\right)^{1/d}\right)\right)^{-1/2} \\
&\geq \sqrt{\pi} - \frac{1}{2\sqrt{\ln 2}} \\
&\geq 0,
\end{aligned}$$

ce qui montre que $\varphi'_\ell(b) \geq H(b)$.

Posons alors $\varphi_\ell(r) = \varphi_\ell(b) + (r-b)\varphi'_\ell(b)$ pour $r \geq b$, de sorte que, finalement, $\phi_\ell(r) \leq \varphi_\ell(r)$ pour tout r.

2.4.4. Preuve du lemme 2.1.4

Remarquons pour commencer que φ_ℓ est concave. En effet, la dérivée seconde de la restriction $\varphi_\ell|_{]0,b]}$ de φ_ℓ à $]0,b]$ est égale à

$$-\frac{\kappa\sqrt{nd}}{2r}\left[\frac{1}{2}\left(\ln\left(\frac{2a^{1/d}b^{1-1/d}}{r}\right)\right)^{-3/2}+\ln\left(\frac{2a^{1/d}b^{1-1/d}}{r}\right)^{-1/2}\right]\leq 0,$$

ce qui implique que $\varphi_\ell|_{]0,b]}$ est concave. Comme φ_ℓ est obtenue en prolongeant $\varphi_\ell|_{]0,b]}$ au moyen de la tangente à cette fonction en b, φ_ℓ est encore concave.

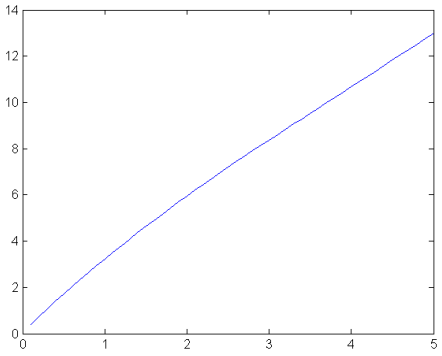


FIGURE 2.33.: Fonction $\dfrac{\varphi_\ell}{\kappa\sqrt{nd}}$ pour $d=2$, $a=6$ et $b=3$.

Revenons à l'équation

$$\varphi_\ell\left(\frac{2\sigma\sqrt{d_\ell}}{\sqrt{nd}}\right)=\frac{\sigma d_\ell}{\sqrt{nd}}.$$

Chercher une solution d_ℓ revient à résoudre, pour $r>0$,

$$\varphi_\ell(r)=\frac{\sqrt{nd}}{4\sigma}r^2,$$

qui admet une unique solution $r_\ell=2\sigma\sqrt{\frac{d_\ell}{nd}}$, puisque φ_ℓ est concave et $r\mapsto r^2$ convexe. En outre, la solution r_ℓ vérifie $r_\ell\leq b$ si, et seulement si,

$$\varphi_\ell(b)\leq\frac{\sqrt{nd}}{4\sigma}b^2,$$

c'est-à-dire

$$\kappa b\sqrt{nd}\left(\sqrt{\ln 2+\frac{1}{d}\ln\left(\frac{a}{b}\right)}+\sqrt{\pi}\right)\leq\frac{\sqrt{nd}}{4\sigma}b^2,$$

ce qui signifie

$$\sigma \leq \frac{b}{4\kappa}\left[\sqrt{\ln 2 + \frac{1}{d}\ln\left(\frac{a}{b}\right)} + \sqrt{\pi}\right]^{-1}.$$

Si cette condition est satisfaite, l'équation devient

$$\kappa r_\ell \sqrt{nd}\left(\sqrt{\ln\left(\frac{2a^{1/d}b^{1-1/d}}{r_\ell}\right)} + \sqrt{\pi}\right) = \frac{\sqrt{nd}}{4\sigma} r_\ell^2,$$

ce qui est équivalent à

$$4\sigma\kappa\left(\sqrt{\ln\left(\frac{2a^{1/d}b^{1-1/d}}{r_\ell}\right)} + \sqrt{\pi}\right) = r_\ell.$$

On constate que $4\sigma\kappa\sqrt{\pi} \leq r_\ell$. Ainsi,

$$r_\ell \leq 4\sigma\kappa\left(\sqrt{\ln\left(\frac{a^{1/d}b^{1-1/d}}{2\sigma\kappa\sqrt{\pi}}\right)} + \sqrt{\pi}\right).$$

Comme $r_\ell = 2\sigma\sqrt{\frac{d_\ell}{nd}}$, on obtient

$$d_\ell \leq 8\kappa^2 nd\left(\ln\left(\frac{a^{1/d}b^{1-1/d}}{2\sigma\kappa\sqrt{\pi}}\right) + \pi\right).$$

2.5. Preuves des résultats de la Section 2.2

2.5.1. Démonstration du Théorème 2.2.1

Ce théorème, adapté de Massart [141, Théorème 8.1] (voir Annexe B.2), se démontre de la même manière que le Théorème 3.2.1 dans le Chapitre 3 de la première partie.

Notons $\bar{\Delta}_n(\mathbf{f}) = \Delta_n(\mathbf{f}) - \mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})]$ le processus empirique centré. Pour tous $k \geq 1$, $\ell \in \mathcal{L}$, pour toute $\mathbf{f}_{k,\ell} \in \mathcal{F}_{k,\ell}$, on a, par définition de $\tilde{\mathbf{f}}$,

$$\Delta_n(\tilde{\mathbf{f}}) + \mathrm{pen}(\hat{k}, \hat{\ell}) \leq \Delta_n(\mathbf{f}_{k,\ell}) + \mathrm{pen}(k, \ell).$$

De manière équivalente,

$$\Delta_n(\tilde{\mathbf{f}}) - \Delta_n(\mathbf{f}_{k,\ell}) \leq \mathrm{pen}(k, \ell) - \mathrm{pen}(\hat{k}, \hat{\ell}).$$

Comme $\Delta_n(\tilde{\mathbf{f}}) = \mathbb{E}[\Delta(\tilde{\mathbf{f}}, \mathbf{X})] + \bar{\Delta}_n(\tilde{\mathbf{f}})$ et $\Delta_n(\mathbf{f}_{k,\ell}) = \mathbb{E}[\Delta(\mathbf{f}_{k,\ell}, \mathbf{X})] + \bar{\Delta}_n(\mathbf{f}_{k,\ell})$, cette inégalité devient

$$\mathbb{E}[\Delta(\tilde{\mathbf{f}}, \mathbf{X})] - \mathbb{E}[\Delta(\mathbf{f}_{k,\ell}, \mathbf{X})] \leq \bar{\Delta}_n(\mathbf{f}_{k,\ell}) - \bar{\Delta}_n(\tilde{\mathbf{f}}) + \mathrm{pen}(k,\ell) - \mathrm{pen}(\hat{k},\hat{\ell}). \tag{2.14}$$

Puisque, pour toute courbe $\mathbf{f}$,

$$\mathcal{D}(\mathbf{f}^\star, \mathbf{f}) = \mathbb{E}[\Delta(\mathbf{f}, \mathbf{X}) - \Delta(\mathbf{f}^\star, \mathbf{X})],$$

on a

$$\mathbb{E}[\Delta(\tilde{\mathbf{f}}, \mathbf{X})] - \mathbb{E}[\Delta(\mathbf{f}_{k,\ell}, \mathbf{X})] = \mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) - \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}). \tag{2.15}$$

En combinant (2.14) et (2.15),

$$\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) \leq \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) - \bar{\Delta}_n(\tilde{\mathbf{f}}) + \mathrm{pen}(k,\ell) - \mathrm{pen}(\hat{k},\hat{\ell}). \tag{2.16}$$

Considérons à présent des poids positifs $\{x_{k,\ell}\}_{k\geq 1, \ell\in\mathcal{L}}$ tels que

$$\sum_{k\geq 1,\ell\in\mathcal{L}} e^{-x_{k,\ell}} = \Sigma < +\infty$$

et fixons $z > 0$. Par le Lemme 3.4.1 (voir la démonstration du Théorème 3.2.1 dans le Chapitre 3), nous obtenons, pour tous $k' \geq 1, \ell' \in \mathcal{L}$ et $\varepsilon \geq 0$,

$$\mathbb{P}\left\{ \sup_{\mathbf{f}\in\mathcal{F}_{k',\ell'}} (-\bar{\Delta}_n(\mathbf{f})) \geq \mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k',\ell'}} (-\bar{\Delta}_n(\mathbf{f}))\right] + \varepsilon \right\} \leq \exp\left(-\frac{2n\varepsilon^2}{\delta^4}\right),$$

ce qui se récrit, pour $\varepsilon = \delta^2\sqrt{\frac{x_{k',\ell'}+z}{2n}}$,

$$\mathbb{P}\left\{ \sup_{\mathbf{f}\in\mathcal{F}_{k',\ell'}} (-\bar{\Delta}_n(\mathbf{f})) \geq \mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k',\ell'}} (-\bar{\Delta}_n(\mathbf{f}))\right] + \delta^2\sqrt{\frac{x_{k',\ell'}+z}{2n}} \right\} \leq e^{-x_{k',\ell'}-z}.$$

Posant $E_{k',\ell'} = \mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k',\ell'}}(-\bar{\Delta}_n(\mathbf{f}))\right]$, on a donc, pour tous $k' \geq 1$, $\ell' \in \mathcal{L}$,

$$\sup_{\mathbf{f}\in\mathcal{F}_{k',\ell'}} (-\bar{\Delta}_n(\mathbf{f})) \leq E_{k',\ell'} + \delta^2\sqrt{\frac{x_{k',\ell'}+z}{2n}},$$

sauf sur un ensemble de probabilité au plus Σe^{-z}. Puis, d'après l'inégalité (2.16), il vient

$$\begin{aligned}
\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) &\leq \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + E_{\hat{k},\hat{\ell}} + \delta^2\sqrt{\frac{x_{\hat{k},\hat{\ell}}+z}{2n}} - \mathrm{pen}(\hat{k},\hat{\ell}) + \mathrm{pen}(k,\ell) \\
&\leq \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + E_{\hat{k},\hat{\ell}} + \delta^2\sqrt{\frac{x_{\hat{k},\hat{\ell}}}{2n}} - \mathrm{pen}(\hat{k},\hat{\ell}) + \mathrm{pen}(k,\ell) + \delta^2\sqrt{\frac{z}{2n}},
\end{aligned}$$

sauf sur un ensemble de probabilité au plus Σe^{-z}. Donc, si pour tous $k' \geq 1$, $\ell' \in \mathcal{L}$,

$$\mathrm{pen}(k', \ell') \geq E_{k',\ell'} + \delta^2 \sqrt{\frac{x_{k',\ell'}}{2n}},$$

alors

$$\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) \leq \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + \mathrm{pen}(k, \ell) + \delta^2 \sqrt{\frac{z}{2n}},$$

sauf sur un ensemble de probabilité au plus Σe^{-z}. Ceci se récrit

$$\mathbb{P}\left\{[\delta^{-2}\sqrt{2n}[\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) - \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + \mathrm{pen}(k, \ell)] \geq \sqrt{z}\right\} \leq \Sigma e^{-z},$$

ou encore, en posant $z = u^2$,

$$\mathbb{P}\left\{[\delta^{-2}\sqrt{2n}[\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) - \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + \mathrm{pen}(k, \ell)] \geq u\right\} \leq \Sigma e^{-u^2}.$$

Comme $\int_0^{+\infty} e^{-u^2} \mathrm{d}u = \frac{\sqrt{\pi}}{2}$, on a

$$\mathbb{E}\Big[(\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) - \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + \mathrm{pen}(k, \ell))_+\Big] \leq \frac{\delta^2 \Sigma}{2^{3/2}} \sqrt{\frac{\pi}{n}},$$

où $g_+ = \max(g, 0)$. Puisque $\mathbb{E}[\bar{\Delta}_n(\mathbf{f}_{k,\ell})] = 0$, il en résulte

$$\mathbb{E}[\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}})] \leq \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \mathrm{pen}(k, \ell) + \frac{\delta^2 \Sigma}{2^{3/2}} \sqrt{\frac{\pi}{n}}.$$

Finalement, comme cette inégalité est vraie pour tous k et ℓ,

$$\mathbb{E}[\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}})] \leq \inf_{k \geq 1, \ell \in \mathcal{L}} \Big[\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\ell}) + \mathrm{pen}(k, \ell)\Big] + \frac{\delta^2 \Sigma}{2^{3/2}} \sqrt{\frac{\pi}{n}},$$

où $\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\ell}) = \inf_{\mathbf{f} \in \mathcal{F}_{k,\ell}} \mathcal{D}(\mathbf{f}^\star, \mathbf{f})$.

2.5.2. Preuve du Lemme 2.2.1

Soient $\varepsilon_1, \ldots, \varepsilon_n$ des variables aléatoires de Rademacher indépendantes, et indépendantes de $\mathbf{X}_1, \ldots, \mathbf{X}_n$, et soient $\mathbf{X}'_1, \ldots, \mathbf{X}'_n$ des copies indépendantes de $\mathbf{X}_1, \ldots, \mathbf{X}_n$, qui sont aussi indépendantes de $\varepsilon_1, \ldots \varepsilon_n$. Un argument de symétrisa-

tion nous donne

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\big(\mathbb{E}[\Delta(\mathbf{f},\mathbf{X})]-\Delta_n(\mathbf{f})\big)\Big]\\
&\quad=\mathbb{E}\Big[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\Big(\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n\Delta(\mathbf{f},\mathbf{X}_i')\Big|\mathbf{X}_1,\dots,\mathbf{X}_n\Big]-\frac{1}{n}\sum_{i=1}^n\Delta(\mathbf{f},\mathbf{X}_i)\Big)\Big]\\
&\quad\le\mathbb{E}\Big[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\frac{1}{n}\sum_{i=1}^n\Big(\Delta(\mathbf{f},\mathbf{X}_i')-\Delta(\mathbf{f},\mathbf{X}_i)\Big)\Big]\\
&\quad=\mathbb{E}\Big[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\Big(\Delta(\mathbf{f},\mathbf{X}_i')-\Delta(\mathbf{f},\mathbf{X}_i)\Big)\Big]\\
&\quad\le\mathbb{E}\Big[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\Delta(\mathbf{f},\mathbf{X}_i')\Big]+\mathbb{E}\Big[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\frac{1}{n}\sum_{i=1}^n(-\varepsilon_i)\Delta(\mathbf{f},\mathbf{X}_i)\Big]\\
&\quad=2\mathbb{E}\Big[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\Delta(\mathbf{f},\mathbf{X}_i)\Big].
\end{aligned}$$

2.5.3. Démonstration de la Proposition 2.2.1

Nous savons qu'il suffit pour démontrer la proposition de majorer l'intégrale de Dudley

$$\int_0^{+\infty}\sqrt{\ln\mathcal{N}(S_{k,\ell},\|\cdot\|_\infty,\varepsilon)}\mathrm{d}\varepsilon.$$

Or, pour tout $\varepsilon>0$,

$$\begin{aligned}
&\ln\mathcal{N}(S_{k,\ell},\|\cdot\|_\infty,\varepsilon)\\
&\quad\le\Big[\frac{\ell\delta}{\varepsilon}+3k+1\Big]\ln 2+(k+1)\ln V_d+d\ln\Big[\frac{\delta^2\sqrt{d}}{\varepsilon}+\sqrt{d}\Big]+kd\ln\Big[\frac{\ell\delta\sqrt{d}}{k\varepsilon}+3\sqrt{d}\Big]\\
&\quad=\frac{\ell\delta}{\varepsilon}\ln 2+(3k+1)\ln 2+(k+1)\ln V_d+d(k+1)\ln\sqrt{d}+d\ln\Big(\frac{\delta^2}{\varepsilon}+1\Big)\\
&\qquad+kd\ln 3+kd\ln\Big(\frac{\ell\delta}{3k\varepsilon}+1\Big)\\
&\quad=\frac{\ell\delta}{\varepsilon}\ln 2+d\ln\Big(\frac{\delta^2}{\varepsilon}+1\Big)+kd\ln\Big(\frac{\ell\delta}{3k\varepsilon}+1\Big)+kd\ln 3+(3k+1)\ln 2\\
&\qquad+(k+1)(\ln V_d+\frac{d}{2}\ln d).
\end{aligned}$$

Donc, comme l'image de $\mathbf{f}$ est incluse dans un convexe $\mathcal{C}$ de diamètre δ, nous obtenons

$$\begin{aligned}
\int_0^{+\infty}\sqrt{\ln\mathcal{N}(S_{k,\ell},\|\cdot\|_\infty,\varepsilon)}\mathrm{d}\varepsilon&=\int_0^{\delta^2}\sqrt{\ln\mathcal{N}(S_{k,\ell},\|\cdot\|_\infty,\varepsilon)}\mathrm{d}\varepsilon\\
&\le\sqrt{\ell\delta\ln 2}I_1+\sqrt{d}I_2+\sqrt{kd}I_3+\delta^2A(k,d),
\end{aligned}$$

où $I_1 = \int_0^{\delta^2} \frac{1}{\sqrt{\varepsilon}} \mathrm{d}\varepsilon$, $I_2 = \int_0^{\delta^2} \sqrt{\ln\Big(\frac{\delta^2}{\varepsilon} + 1\Big)} \mathrm{d}\varepsilon$, $I_3 = \int_0^{\delta^2} \sqrt{\ln\Big(\frac{\ell\delta}{3k\varepsilon} + 1\Big)} \mathrm{d}\varepsilon$, et

$$A(k,d) = \Big[kd \ln 3 + (3k+1)\ln 2 + (k+1)(\ln V_d + \frac{d}{2}\ln d)\Big]^{1/2}.$$

Intégrale $\mathbf{I_1}$. On a

$$I_1 = \int_0^{\delta^2} \frac{1}{\sqrt{\varepsilon}} \mathrm{d}\varepsilon = 2\delta.$$

Intégrale $\mathbf{I_2}$. Pour $\varepsilon \in]0, \delta^2]$, $\frac{\delta^2}{\varepsilon} \geq 1$, donc

$$\begin{aligned} I_2 &\leq \int_0^{\delta^2} \sqrt{\ln\Big(\frac{2\delta^2}{\varepsilon}\Big)} \mathrm{d}\varepsilon \\ &= 2\delta^2 \int_0^{1/2} \sqrt{\ln \frac{1}{u}} \mathrm{d}u \\ &\leq \delta^2(\sqrt{\ln 2} + \sqrt{\pi}). \end{aligned}$$

Intégrale $\mathbf{I_3}$. Soit $M = \max(3k, L/\delta)$. Pour tout $\ell \in L$, $\delta \geq \frac{\ell}{M}$, et donc $\delta^2 \geq \frac{\ell\delta}{M}$. On découpe l'intégrale I_3 en écrivant

$$\begin{aligned} I_3 &= \int_0^{\delta^2} \sqrt{\ln\Big(\frac{\ell\delta}{3k\varepsilon} + 1\Big)} \mathrm{d}\varepsilon \\ &= \int_0^{\ell\delta/M} \sqrt{\ln\Big(\frac{\ell\delta}{3k\varepsilon} + 1\Big)} \mathrm{d}\varepsilon + \int_{\ell\delta/M}^{\delta^2} \sqrt{\ln\Big(\frac{\ell\delta}{3k\varepsilon} + 1\Big)} \mathrm{d}\varepsilon. \end{aligned} \tag{2.17}$$

Observons que, comme $\varepsilon \leq \frac{\ell\delta}{M}$, on a $\frac{\ell\delta}{3k\varepsilon} \geq 1$. On obtient alors

$$\begin{aligned} \int_0^{\ell\delta/M} \sqrt{\ln\Big(\frac{\ell\delta}{3k\varepsilon} + 1\Big)} \mathrm{d}\varepsilon &\leq \int_0^{\ell\delta/M} \sqrt{\ln\Big(\frac{2\ell\delta}{3k\varepsilon}\Big)} \mathrm{d}\varepsilon \\ &= \frac{2\ell\delta}{3k} \int_0^{3k/2M} \sqrt{\ln \frac{1}{u}} \mathrm{d}u \\ &\leq \frac{\ell\delta}{M}\Big(\sqrt{\ln\Big(\frac{2M}{3k}\Big)} + \sqrt{\pi}\Big). \end{aligned}$$

Pour la seconde intégrale de l'égalité (2.17), comme la fonction à intégrer est décroissante en ε, on a

$$\begin{aligned} \int_{\ell\delta/M}^{\delta^2} \sqrt{\ln\Big(\frac{\ell\delta}{3k\varepsilon} + 1\Big)} \mathrm{d}\varepsilon &\leq \Big(\delta^2 - \frac{\ell\delta}{M}\Big)\sqrt{\ln\Big(\frac{M}{3k} + 1\Big)} \\ &\leq \Big(\delta^2 - \frac{\ell\delta}{M}\Big)\sqrt{\ln\Big(\frac{2M}{3k}\Big)}. \end{aligned}$$

Il en résulte que

$$I_3 \leq \delta^2 \sqrt{\ln\left(\frac{2M}{3k}\right)} + \frac{\ell\delta}{M}\sqrt{\pi}.$$

Donc,

$$\begin{aligned}
&\int_0^{+\infty} \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon \\
&\quad \leq 2\delta\sqrt{\delta\ell \ln 2} + \sqrt{d}\delta^2(\sqrt{\ln 2} + \sqrt{\pi}) + \frac{\ell\delta}{M}\sqrt{kd\pi} + \delta^2\sqrt{kd\ln\left(\frac{2M}{3k}\right)} + \delta^2 A(k,d) \\
&\quad = 2\delta\sqrt{\delta\ell \ln 2} + \frac{\ell\delta}{M}\sqrt{kd\pi} + a_0 \\
&\qquad + \sqrt{k}\delta^2\left[d\ln\left(\frac{2M}{3k}\right) + d\ln 3 + \frac{d}{2}\ln d + \ln V_d + 3\ln 2\right]^{1/2},
\end{aligned}$$

où a_0 est une constante positive. Finalement,

$$\int_0^{+\infty} \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon \leq a_1\sqrt{k} + a_2\frac{\ell}{\sqrt{k}} + a_0,$$

où les constantes positives $a_0, \ldots, a_2$ ne dépendent que de la longueur maximale L, de la dimension d et du diamètre δ du convexe $\mathcal{C}$.

2.5.4. Démonstration de la proposition 2.2.2

Procédant comme dans la démonstration de la Proposition 2.2.1, nous cherchons à majorer l'intégrale de Dudley

$$\int_0^{+\infty} \sqrt{\ln \mathcal{N}(S_{k,\kappa}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon.$$

Pour tout $\varepsilon > 0$, on a

$$\begin{aligned}
&\ln \mathcal{N}(S_{k,\kappa}, \|\cdot\|_\infty, \varepsilon) \\
&\quad \le \Big(\frac{\zeta(\kappa)\delta^2}{\varepsilon} + 2k + 1\Big)\ln 2 + (k+1)\ln V_d + d\ln\Big(\frac{\delta^2\sqrt{d}}{\varepsilon} + \sqrt{d}\Big) \\
&\qquad + kd\ln\Big(\frac{\zeta(\kappa)\delta^2\sqrt{d}}{k\varepsilon} + 3\sqrt{d}\Big) \\
&\quad = \frac{\zeta(k)\delta^2}{\varepsilon}\ln 2 + (2k+1)\ln 2 + (k+1)\ln V_d + d(k+1)\ln\sqrt{d} + d\ln\Big(\frac{\delta^2}{\varepsilon} + 1\Big) \\
&\qquad + kd\ln 3 + kd\ln\Big(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\Big) \\
&\quad = \frac{\zeta(k)\delta^2}{\varepsilon}\ln 2 + d\ln\Big(\frac{\delta^2}{\varepsilon} + 1\Big) + kd\ln\Big(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\Big) + kd\ln 3 + (2k+1)\ln 2 \\
&\qquad + (k+1)(\ln V_d + \frac{d}{2}\ln d).
\end{aligned}$$

Par conséquent,

$$\begin{aligned}
\int_0^{+\infty} \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon &= \int_0^{\delta^2} \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon \\
&\le \delta\sqrt{\zeta(\kappa)\ln 2}I_1 + \sqrt{d}I_2 + \sqrt{kd}I_3 + \delta^2 A(k,d),
\end{aligned}$$

où $I_1 = \displaystyle\int_0^{\delta^2} \frac{1}{\sqrt{\varepsilon}}\mathrm{d}\varepsilon$, $I_2 = \displaystyle\int_0^{\delta^2} \sqrt{\ln\Big(\frac{\delta^2}{\varepsilon} + 1\Big)}\mathrm{d}\varepsilon$, $I_3 = \displaystyle\int_0^{\delta^2} \sqrt{\ln\Big(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\Big)}\mathrm{d}\varepsilon$, et

$$A(k,d) = \delta^2\Big[kd\ln 3 + (2k+1)\ln 2 + (k+1)(\ln V_d + \frac{d}{2}\ln d)\Big]^{1/2}.$$

Intégrale $\mathbf{I_1}$. On a $I_1 = \displaystyle\int_0^{\delta^2} \frac{1}{\sqrt{\varepsilon}}\mathrm{d}\varepsilon = 2\delta$.

Intégrale $\mathbf{I_2}$. Pour $\varepsilon \in]0, \delta^2]$, $\dfrac{\delta^2}{\varepsilon} \ge 1$, donc

$$\begin{aligned}
I_2 &\le \int_0^{\delta^2} \sqrt{\ln\Big(\frac{2\delta^2}{\varepsilon}\Big)}\mathrm{d}\varepsilon \\
&= 2\delta^2 \int_0^{1/2} \sqrt{\ln\frac{1}{u}}\mathrm{d}u \\
&\le \delta^2(\sqrt{\ln 2} + \sqrt{\pi}).
\end{aligned}$$

Intégrale $\mathbf{I_3}$. Si $\dfrac{\zeta(\kappa)}{3k} \geq 1$, comme $\dfrac{\delta^2}{\varepsilon} \geq 1$ pour $\varepsilon \in]0, \delta^2]$, on a

$$\begin{aligned} I_3 &= \int_0^{\delta^2} \sqrt{\ln\Big(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\Big)}\mathrm{d}\varepsilon \\ &\leq \int_0^{\delta^2} \sqrt{\ln\Big(\frac{2\zeta(\kappa)\delta^2}{3k\varepsilon}\Big)}\mathrm{d}\varepsilon \\ &= \frac{2\zeta(\kappa)\delta^2}{3k}\int_0^{3k/2\zeta(\kappa)} \sqrt{\ln\frac{1}{u}}\mathrm{d}u \\ &\leq \delta^2\bigg(\sqrt{\ln\frac{2\zeta(\kappa)}{3k}} + \sqrt{\pi}\bigg) \end{aligned}$$

Si $\dfrac{\zeta(\kappa)}{3k} < 1$, on découpe I_3 en écrivant

$$\begin{aligned} I_3 &= \int_0^{\delta^2} \sqrt{\ln\Big(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\Big)}\mathrm{d}\varepsilon \\ &= \int_0^{\zeta(\kappa)\delta^2/3k} \sqrt{\ln\Big(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\Big)}\mathrm{d}\varepsilon + \int_{\zeta(\kappa)\delta^2/3k}^{\delta^2} \sqrt{\ln\Big(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\Big)}\mathrm{d}\varepsilon. \end{aligned} \tag{2.18}$$

La première intégrale est majorée en utilisant l'inégalité $\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} \geq 1$ pour $\varepsilon \in]0, \frac{\zeta(\kappa)\delta^2}{3k}]$. On obtient

$$\begin{aligned} \int_0^{\zeta(\kappa)\delta^2/3k} \sqrt{\ln\Big(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\Big)}\mathrm{d}\varepsilon &\leq \int_0^{\zeta(\kappa)\delta^2/3k} \sqrt{\ln\Big(\frac{2\zeta(\kappa)\delta^2}{3k\varepsilon}\Big)}\mathrm{d}\varepsilon \\ &= \frac{2\zeta(\kappa)\delta^2}{3k}\int_0^{1/2} \sqrt{\ln\frac{1}{u}}\mathrm{d}u \\ &\leq \frac{\zeta(\kappa)\delta^2}{3k}(\sqrt{\ln 2} + \sqrt{\pi}). \end{aligned}$$

Pour la seconde intégrale dans 2.18, comme la fonction considérée est décroissante en ε,

$$\int_{\zeta(\kappa)\delta^2/3k}^{\delta^2} \sqrt{\ln\Big(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\Big)}\mathrm{d}\varepsilon \leq \Big(\delta^2 - \frac{\zeta(\kappa)\delta^2}{3k}\Big)\sqrt{\ln 2}.$$

Finalement, on a

$$I_3 \leq \begin{cases} \delta^2\bigg(\sqrt{\ln\dfrac{\zeta(\kappa)}{3k}} + \sqrt{\pi} + \sqrt{\ln 2}\bigg) & \text{si} \quad \dfrac{\zeta(\kappa)}{3k} \geq 1 \\ \delta^2\bigg(\dfrac{\zeta(\kappa)}{3k}\sqrt{\pi} + \sqrt{\ln 2}\bigg) & \text{si} \quad \dfrac{\zeta(\kappa)}{3k} < 1. \end{cases}$$

En collectant les différents résultats, il vient

$$
\begin{aligned}
&\int_0^{+\infty} \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon \\
&\leq 2\delta^2\sqrt{\zeta(\kappa)\ln 2} + \sqrt{d}\delta^2(\sqrt{\ln 2} + \sqrt{\pi}) + \delta^2\sqrt{kd}\left(\sqrt{\ln\frac{\zeta(\kappa)}{3k}} + \sqrt{\pi} + \sqrt{\ln 2}\right)\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}\geq 1\}} \\
&\quad + \delta^2\sqrt{kd}\left(\frac{\zeta(\kappa)}{3k}\sqrt{\pi} + \sqrt{\ln 2}\right)\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}<1\}} + \delta^2 A(k,d) \\
&\leq \delta^2\Big(2\sqrt{\zeta(\kappa)\ln 2} + \frac{\zeta(\kappa)}{3\sqrt{k}}\sqrt{\pi d}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}<1\}} + \sqrt{kd\ln\frac{\zeta(\kappa)}{3k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}\geq 1\}} \\
&\quad + \sqrt{k}\Big[\sqrt{d}(\sqrt{\pi} + \sqrt{\ln 2}) + \Big(d\ln 3 + \frac{d}{2}\ln d + \ln V_d + 2\ln 2\Big)^{1/2} + a_0\Big]\Big),
\end{aligned}
$$

où a_0 est une constante positive. Finalement,

$$
\begin{aligned}
&\int_0^{+\infty} \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon \\
&\qquad \leq \delta^2\Big(a_1\sqrt{k} + a_2\sqrt{\zeta(\kappa)} + a_3\frac{\zeta(\kappa)}{\sqrt{k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}<1\}} + a_4\sqrt{k\ln\frac{\zeta(\kappa)}{k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}\geq 1\}} + a_0\Big),
\end{aligned}
$$

où les constantes $a_0, \ldots, a_4$ ne dépendent que de la dimension d.

Conclusion et perspectives

De par la nature même de cette technique, apprécier la pertinence d'un résultat d'apprentissage non supervisé est relativement difficile. En effet, il s'agit d'une méthode exploratoire qui vise à découvrir une structure intrinsèque cachée dans les données et à représenter celles-ci de manière à la fois précise et compacte, sans qu'il existe, de prime abord, de critère objectif indiquant comment effectuer le compromis entre ces deux requêtes. Pourtant, même lorsqu'elle semble apporter une information sommaire, cette démarche peut se révéler extrêmement utile. En effet, parfois suivie d'une procédure d'apprentissage supervisé, elle peut constituer un préalable indispensable à une analyse plus fine.

Le clustering, par exemple, qui a fait l'objet de la première partie de la thèse, permet de mettre en lumière des groupes qui pourront être utilisés en classification. Dans le Chapitre 1 de la première partie, à la suite de Banerjee, Merugu, Dhillon et Ghosh [17], dont l'approche est plus orientée informatique, nous nous sommes intéressés d'un point de vue théorique à la quantification et au clustering étendus à la classe des divergences de Bregman (Bregman [40]). Ce travail s'inscrit dans un contexte où les techniques d'apprentissage non supervisé remarquables de simplicité que sont la quantification et le clustering trouvent des applications dans des domaines toujours plus variés, pour des données de complexité croissante. L'objectif du Chapitre 1 était donc d'étendre à différentes divergences de Bregman des résultats théoriques de quantification valables dans un espace de Hilbert muni de sa norme hilbertienne. En particulier, nos résultats justifient l'utilisation en clustering de certaines mesures de distorsion qui font en fait partie de la famille des divergences de Bregman. Cependant, cette classe permet de fabriquer des mesures de distorsion bien plus nombreuses et diversifiées que celles que nous avons citées dans nos exemples et la question de savoir quelle divergence est adaptée à quel type de données est très intéressante, surtout dans la perspective des applications. Ce point pourrait être approfondi. En effet, la relation entre familles exponentielles et divergences de Bregman apporte un élément de réponse, mais elle est difficilement exploitable en pratique puisque la loi sous-jacente des observations est inconnue. Par ailleurs, si les conditions données dans le Chapitre 1 assurent l'existence d'un quantificateur optimal et la convergence de la distorsion à la vitesse $1/\sqrt{n}$, une étape ultérieure consisterait à déterminer s'il est possible d'affaiblir ces conditions. Il pourrait également s'avérer enrichissant d'adopter un point de vue différent et

d'explorer les problématiques de la quantification et du clustering dans l'esprit de Pollard [156], à la lumière de résultats sur les processus empiriques et les espacements (Shorack et Wellner [170], del Barrio, Deheuvels et van de Geer [20], Deheuvels [63]).

Le Chapitre 2 illustre l'utilisation combinée de méthodes d'apprentissage supervisé et non supervisé, puisque, dans cette application industrielle concrète, le clustering est destiné à améliorer la précision d'une étape de régression. Dans ce chapitre, nous utilisons une projection sur des bases, effectuons le clustering sur les projections et étudions du point de vue de la distorsion les propriétés de convergence des centres obtenus. La dimension de projection d étant supposée fixée, une extension pourrait consister à développer une méthode permettant dans cette situation de choisir la dimension de projection automatiquement.

Complément logique aux deux précédents chapitres, le Chapitre 3 aborde l'importante question du choix du nombre de groupes. Elle est traitée dans le contexte de cette partie, qui ne permet pas une analyse aussi fine que le cadre des modèles de mélange. Nous avons vu que le problème est en partie mal posé du fait de l'absence de cible pertinente, ce qui rend difficile l'appréciation théorique de la qualité d'une solution. Néanmoins, notre approche constitue un premier pas dans cette direction, avec des résultats pratiques plutôt satisfaisants.

Dans la seconde partie de la thèse, après avoir proposé une synthèse sur ces beaux objets mathématiques que sont les courbes principales, mettant en évidence la diversité des points de vue possibles et des applications, nous nous sommes attachés, pour la définition reposant sur la minimisation d'un critère de moindres carrés, à élaborer une méthode permettant le choix d'une classe de courbes convenable. Nous avons traité ce problème sous l'angle de la sélection de modèle, en ajoutant au critère empirique considéré une fonction de pénalité, et avons obtenu des garanties théoriques pour l'estimateur représentant la classe de courbes qui correspond à la minimisation de ce critère pénalisé.

Deux approches différentes ont été développées, la première consistant à partir d'un modèle gaussien et donnant un résultat relatif aux points échantillonnés sur la courbe, et la seconde, proposant une inégalité portant sur la courbe dans un cadre borné. Pour chacune d'entre elles, nous entrevoyons une extension qui pourrait se révéler fructueuse. Dans la première situation, il s'agit de faire l'hypothèse que les points sont distribués le long de la courbe suivant une loi uniforme et de considérer le problème dans le contexte de l'estimation de densité, tandis que le second cas pourrait gagner en précision par le biais d'arguments de concentration permettant de contrôler le comportement local d'un processus empirique (Massart et Nédélec [142]). En outre, puisque l'on sait calculer l'entropie métrique de tels

objets, une autre piste consisterait à exploiter le fait que les courbes de $\mathbb{R}^d$ de longueur bornée par ℓ correspondent aux fonctions lipschitziennes à valeurs dans $\mathbb{R}^d$ ayant ℓ pour constante de Lipschitz.

D'autre part, si nous nous sommes ici concentrés sur le point de vue de Kégl *et al.* [117] et Sandilya et Kulkarni [166], il y a fort à parier que travailler dans le même sens sur d'autres définitions de courbes principales conduirait à d'intéressants résultats.

Enfin, bien que l'heuristique de pente, mise en œuvre en pratique pour calibrer les pénalités, n'ait pas constitué le sujet central de cette thèse, certaines questions soulevées peuvent susciter un vif intérêt. En particulier, réfléchir au cas des nappes correspondant à plusieurs paramètres serait une direction de recherche possible dans ce domaine.

Annexes

A. Moyennes de Rademacher

Cette annexe rappelle la définition et les principales propriétés des moyennes de Rademacher, qui sont des objets fort utiles en apprentissage statistique (Bartlett, Boucheron et Lugosi [22], Koltchinskii [122], Boucheron, Bousquet et Lugosi [39]).

A.1. Définition

Soient $A \subset \mathbb{R}^n$ un ensemble borné constitué de vecteurs $\mathbf{a} = (a_1, \ldots, a_n)$ et $\varepsilon_1, \ldots, \varepsilon_n$ des variables aléatoires de Rademacher indépendantes, c'est-à-dire des variables aléatoires indépendantes à valeurs dans $\{-1, 1\}$ telles que $\mathbb{P}\{\varepsilon_i = -1\} = \mathbb{P}\{\varepsilon_i = 1\} = \frac{1}{2}$. La moyenne de Rademacher associée à A est définie par

$$R_n(A) = \mathbb{E}\left[\sup_{\mathbf{a} \in A} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i a_i\right].$$

A.2. Propriétés

- Si A est un ensemble symétrique, au sens où $\mathbf{a} \in A$ implique $-\mathbf{a} \in A$, on a

$$R_n(A) = \mathbb{E}\left[\sup_{\mathbf{a} \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \varepsilon_i a_i \right|\right].$$

- Soient A, B des parties symétriques bornées de $\mathbb{R}^n$, et c une constante. Notons $A+B = \{\mathbf{a}+\mathbf{b} : \mathbf{a} \in A, \mathbf{b} \in B\}$ et $cA = \{c\mathbf{a} : \mathbf{a} \in A\}$. Les propriétés suivantes sont alors vérifiées :
  1. $R_n(A \cup B) \leq R_n(A) + R_n(B)$,
  2. $R_n(cA) = |c| R_n(A)$,
  3. $R_n(A + B) \leq R_n(A) + R_n(B)$.
- Soit $\phi : \mathbb{R} \to \mathbb{R}$ une fonction L_ϕ-lipschitzienne telle que $\phi(0) = 0$. On définit $\phi \circ A$ comme l'ensemble des vecteurs $(\phi(a_1), \ldots, \phi(a_n)) \in \mathbb{R}^n$ avec $\mathbf{a} = (a_1, \ldots, a_n) \in A$. Alors,

$$R_n(\phi \circ A) \leq L_\phi R_n(A).$$

  Il s'agit du principe de contraction, dû à Ledoux et Talagrand [126].

A.3. Type d'un espace de Banach

Soit $(\varepsilon_i)_{i\geq 1}$ une suite de variables indépendantes de Rademacher. Un espace de Banach E est dit de type p, $1 \leq p \leq 2$, s'il existe une constante T_p telle que, pour toute suite finie $a_1, \dots a_n$ d'éléments de E,

$$\left(\mathbb{E}\left[\left\|\sum_{i=1}^{n} \varepsilon_i a_i\right\|^p\right]\right)^{1/p} \leq T_p \left(\sum_{i=1}^{n} \|a_i\|^p\right)^{1/p}.$$

Pour davantage de détails sur cette notion, on pourra se reporter à Ledoux et Talagrand [126, Chapitre 9].

A.4. Moyennes de Rademacher d'une classe de fonctions

Pour un ensemble d'observations $X_1, \dots, X_n$, la moyenne de Rademacher relative à une classe $\mathcal{G}$ de fonctions à valeurs réelles est donnée par

$$R_n(\mathcal{G}) = \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g(X_i)\right].$$

B. Quelques rappels de sélection de modèle

Dans cette annexe, nous rappelons les principes de la sélection de modèle par pénalisation (Birgé et Massart [35], Barron, Birgé, Massart [21], Birgé et Massart [36]). La première section introduit le cadre général de la sélection de modèle. Ensuite sont présentés deux théorèmes, utilisés dans le Chapitre 3 de la première partie et le Chapitre 2 de la seconde partie. Le sujet de la dernière section est la méthode de calibration des pénalités appelée heuristique de pente introduite par Birgé et Massart [37] (voir également Arlot et Massart [10]). Pour un exposé complet sur la théorie de la sélection de modèle, le lecteur pourra se reporter à la monographie de Massart [141].

B.1. Cadre de la sélection de modèle

B.1.1. Minimisation de contraste empirique

Considérons un vecteur aléatoire $\xi^{(n)} = (\xi_1, \dots, \xi_n)$, dont la loi dépend d'une quantité inconnue s, appartenant à un certain ensemble $\mathcal{S}$.

Nous nous donnons alors un critère empirique γ_n, basé sur les observations $\xi_1, \dots, \xi_n$, tel qu'il existe $s \in \mathcal{S}$ réalisant le minimum de la fonction $t \mapsto \mathbb{E}[\gamma_n(t)]$. Un tel critère est appelé contraste empirique pour l'estimation de s. En particulier, soit γ_n un contraste empirique qui s'écrit, pour une fonction γ convenable, sous la forme

$$\gamma_n(t) = \frac{1}{n} \sum_{i=1}^{n} \gamma(t, \xi_i).$$

Si S est un sous-ensemble de $\mathcal{S}$, appelé modèle, un estimateur du minimum de contraste $\hat{s}$ de s est défini comme un minimiseur de γ_n sur l'ensemble S. Au contraste empirique γ_n, nous pouvons associer la fonction de perte

$$\ell(s, t) = \mathbb{E}[\gamma_n(t)] - \mathbb{E}[\gamma_n(s)].$$

Remarquons que $\ell(s, t) \geq 0$ pour tout t, par définition de s.

Dans le cadre de la régression, par exemple, supposons que nous observons des copies $(X_1, Y_1), \dots, (X_n, Y_n)$ d'un couple de variables aléatoires (X, Y), où X est à valeurs dans un certain espace mesurable $\mathcal{X}$ et Y dans $[0, 1]$. Plus précisément,

$$Y_i = s(X_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

où les variables $\varepsilon_1, \dots, \varepsilon_n$ sont indépendantes, de même loi et centrées. La quantité à estimer est la fonction de régression $s(x) = \mathbb{E}[Y|X = x]$. L'ensemble $\mathcal{S}$ est l'ensemble de toutes les fonctions mesurables de $\mathcal{X}$ dans $[0, 1]$. Dans ce cas, le contraste empirique correspondant à

$$\gamma(t, (x, y)) = (y - t(x))^2$$

convient, car s minimise $\mathbb{E}[(Y - t(X))^2]$ sur toutes les fonctions mesurables à valeurs dans $[0, 1]$. La fonction de perte associée est alors

$$\ell(s, t) = \mathbb{E}[(t(X) - s(X))^2].$$

En effet, en utilisant le fait que $\mathbb{E}[Y - s(X)|X] = 0$,

$$\begin{aligned}
\mathbb{E}[(Y - t(X))^2] - \mathbb{E}[(Y - s(X))^2] &= \mathbb{E}[t(X)^2 - s(X)^2 + 2\langle Y, s(X) - t(X)\rangle] \\
&= \mathbb{E}\Big[\mathbb{E}[t(X)^2 - s(X)^2 + 2\langle Y, s(X) - t(X)\rangle|X]\Big] \\
&= \mathbb{E}[t(X)^2 - s(X)^2 + 2\langle s(X), s(X) - t(X)\rangle] \\
&= \mathbb{E}[(t(X) - s(X))^2].
\end{aligned}$$

Le problème qui se pose est celui du choix d'un modèle S approprié. Un petit modèle S est un bon choix si l'on a de bonnes raisons de penser que s est effectivement très proche de ce modèle, mais peut s'avérer complètement faux sinon. Pour que le vrai s ne soit pas trop éloigné du modèle S retenu, on aurait envie de prendre S le plus grand possible, voire $S = \mathcal{S}$, mais alors l'estimateur obtenu peut être mauvais, même lorsque s appartient à ce modèle. En effet, si nous reprenons l'exemple de la régression, nous ne pouvons pas espérer obtenir un résultat satisfaisant en autorisant toutes les fonctions continues sur $[0, 1]$.

Soit $\{S_m\}_{m\in\mathcal{M}}$ une collection de modèles au plus dénombrable. A chaque modèle S_m, on associe l'estimateur du minimum de contraste $\hat{s}_m$. Il s'agit de trouver le meilleur estimateur de la collection $\{\hat{s}_m\}_{m\in\mathcal{M}}$. On voudrait idéalement utiliser $m(s)$, minimisant le risque $\mathbb{E}[\ell(s, \hat{s}_m)]$ en $m \in \mathcal{M}$. L'élément $\hat{s}_{m(s)}$, correspondant au modèle $S_{m(s)}$, est appelé un oracle (Donoho et Johnstone [73]). Puisque le risque dépend du paramètre inconnu s, $m(s)$ en dépend aussi, et ainsi l'oracle ne peut être un estimateur de s. Cependant, $\hat{s}_{m(s)}$ sert de référence et l'on peut évaluer la performance d'un estimateur $\hat{s}_m$ en comparant son risque avec celui d'un oracle. L'objectif est ainsi de sélectionner un estimateur $\hat{s}_{\hat{m}}$ dont le risque se rapproche le plus possible de celui de l'oracle. Pour ce faire, une méthode reposant sur la minimisation d'un critère pénalisé peut être utilisée.

B.1.2. Sélection de modèle par pénalisation

La sélection de modèle par pénalisation consiste à considérer une fonction de pénalité pen : $\mathcal{M} \to \mathbb{R}^+$ destinée à pénaliser les modèles de trop grande complexité. On choisit ensuite $\hat{m}$ minimisant sur $\mathcal{M}$ le critère

$$\gamma_n(\hat{s}_m) + \text{pen}(m)$$

. On peut alors définir le modèle $S_{\hat{m}}$ et l'estimateur associé $\tilde{s} = \hat{s}_{\hat{m}}$. Le problème essentiel est ici le choix d'une pénalité pen(m) convenable.

Les premiers critères pénalisés ont été proposés dans les années 1970 par Akaike [2], Mallows [139] et Schwarz [167]. En notant D_m le nombre de paramètres du modèle S_m, le critère AIC d'Akaike et le critère BIC de Schwarz, utilisés dans le cadre de l'estimation de densité, correspondent à la log-vraisemblance pénalisée par D_m/n et $D_m \ln n/n$ respectivement. Pour la régression des moindres carrés pénalisée avec variance σ^2 connue, le C_p de Mallows consiste à prendre une pénalité en $2D_m\sigma^2/n$. On peut remarquer que ces pénalités sont proportionnelles au nombre de paramètres du modèle. Ces méthodes présentent certains inconvénients. Le C_p de Mallows suppose que $\|\hat{s}_m\|^2$ reste proche de son espérance $\|s_m\|^2 + \sigma^2/n$ uniformément en $m \in \mathcal{M}$, mais rien ne le garantit. Les critères AIC et BIC, quant à eux, reposent sur des théorèmes limites. Ils sont asymptotiques, dans le sens où la taille n de l'échantillon est censée tendre vers l'infini tandis que la collection de modèles est fixée, alors que dans diverses situations, il est pertinent de laisser varier la taille et la liste des modèles en fonction de n.

La théorie de sélection de modèle développée dans les années 1990 par Birgé et Massart [35] est basée sur des inégalités de concentration, qui permettent d'obtenir des résultats non asymptotiques et de justifier ou corriger l'heuristique de Mallows.

B.2. Deux théorèmes de sélection de modèle

Dans cette section sont présentés deux résultats qui reposent sur cette théorie non-asymptotique de sélection de modèle. Ces théorèmes, qui s'emploient dans des contextes différents, sont utilisés dans le Chapitre 3 de la première partie et le Chapitre 2 de la seconde partie.

Pour simplifier l'exposition, les deux théorèmes sont énoncés pour de vrais estimateurs de minimum de contraste. Observons qu'il est possible de considérer des minimiseurs $\{\hat{s}_m\}_{m\in\mathcal{M}}$ de contraste empirique « approchés », vérifiant pour tout $t \in S_m$,

$$\gamma_n(\hat{s}_m) \leq \gamma_n(t) + \rho,$$

où $\rho \geq 0$.

B.2.1. Théorème de sélection de modèle gaussien non linéaire

Le premier des deux résultats est un théorème de sélection de modèle gaussienne adapté au cas où les modèles considérés ne sont pas nécessairement linéaires. Sous la forme donnée ci-dessous, ce résultat est une conséquence de Massart [141, Théorème 4.18] et du lemme de *pealing* (voir [141, Lemme 4.23]).

Nous nous plaçons dans $\mathbb{R}^n$ muni du produit scalaire défini par

$$\langle \mathbf{u}, \mathbf{v} \rangle = \frac{1}{n} \sum_{i=1}^{n} u_i v_i.$$

La norme associée est notée $\|\cdot\|$. Un vecteur gaussien $\mathbf{Y}$ est observé, donné par

$$Y_i = s_i + \sigma \varepsilon_i, \quad i = 1, \ldots, n,$$

où les ε_i sont des variables aléatoires indépendantes, de loi normale centrée réduite. Notons $\mathbf{s} = (s_1, \ldots, s_n)$ et $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$.

Théorème B.2.1. *Soit $\{S_m\}_{m \in \mathcal{M}}$ une collection au plus dénombrable de sous-ensembles de $\mathbb{R}^n$. Pour tout $m \in \mathcal{M}$, φ_m désigne une fonction continue croissante à valeurs positives, telle que $u \mapsto \frac{\varphi_m(u)}{u}$ soit décroissante, et vérifiant*

$$\mathbb{E}\left[\sup_{\mathbf{t} \in S_m, \|\mathbf{t} - \mathbf{t}'\| \leq u} \sqrt{n} \langle \boldsymbol{\varepsilon}, \mathbf{t} - \mathbf{t}' \rangle \right] \leq \frac{\varphi_m(u)}{8}$$

pour tout $\mathbf{t}' \in S_m$ et tout $u > 0$. Soit $D_m > 0$ tel que

$$\varphi_m\left(\tau_m \sigma \sqrt{\frac{D_m}{n}} \right) = \frac{\sigma D_m}{\sqrt{n}},$$

où $\tau_m = 1$ si S_m est convexe et fermé et $\tau_m = 2$ sinon, et soit $\{x_m\}_{m \in \mathcal{M}}$ une famille de poids telle que

$$\sum_{m \in \mathcal{M}} e^{-x_m} = \Sigma < \infty.$$

Prenons alors

$$\mathrm{pen}(m) \geq \eta \frac{\sigma^2}{n} (\sqrt{D_m} + \sqrt{2x_m})^2,$$

où $\eta > 1$ est une constante. Si pour tout $m \in \mathcal{M}$, $\hat{\mathbf{s}}_m$ est un minimiseur de $\|\mathbf{Y} - \mathbf{t}\|^2$ sur S_m, alors, presque sûrement, il existe un minimiseur $\hat{m}$ de

$$\|\mathbf{Y} - \hat{\mathbf{s}}_m\|^2 + \mathrm{pen}(m)$$

sur $\mathcal{M}$. De plus, si l'on note $\tilde{\mathbf{s}} = \hat{\mathbf{s}}_{\hat{m}}$,

$$\mathbb{E}\left[\|\tilde{\mathbf{s}} - \mathbf{s}\|^2\right] \leq c(\eta)\left[\inf_{m\in\mathcal{M}}(d^2(\mathbf{s}, S_m) + \text{pen}(m)) + \frac{\sigma^2}{n}(\Sigma + 1)\right],$$

où la constante $c(\eta)$ ne dépend que de η et $d(\mathbf{s}, S_m) = \inf_{\mathbf{t}\in S_m} \|\mathbf{t} - \mathbf{s}\|$.

Le critère de Dudley [77] montre qu'il suffit de choisir

$$\varphi_m(u) = \kappa \int_0^u \sqrt{\mathcal{H}(S_m, \|\cdot\|, \varepsilon)}\mathrm{d}\varepsilon,$$

où κ est une constante absolue et $\mathcal{H}$ désigne l'entropie métrique (voir par exemple van der Vaart et Wellner [184]). Nous pouvons remarquer que cette fonction est effectivement croissante, et que $u \mapsto \frac{\varphi(u)}{u}$ est décroissante puisque sa dérivée est égale à $\frac{\kappa}{u}\left[\sqrt{\mathcal{H}(S_m, \|\cdot\|, u)} - \frac{1}{u}\int_0^u \sqrt{\mathcal{H}(S_m, \|\cdot\|, \varepsilon)}\mathrm{d}\varepsilon\right]$.

Calculer exactement l'intégrale $\int_0^u \sqrt{\mathcal{H}(S_m, \|\cdot\|, \varepsilon)}\mathrm{d}\varepsilon$ est souvent difficile, mais la preuve de Massart [141, Théorème 4.18] montre qu'il est aussi possible de considérer une fonction φ_m vérifiant, pour tout u,

$$\varphi_m(u) \geq \kappa \int_0^u \sqrt{\mathcal{H}(S_m, \|\cdot\|, \varepsilon)}\mathrm{d}\varepsilon.$$

B.2.2. Un théorème général de sélection de modèle

Le théorème général de sélection de modèle énoncé ici s'applique dans le cadre de l'apprentissage statistique (voir Massart [141, Chapitre 8]).

Théorème B.2.2. *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes à valeurs dans un espace mesurable Ξ, de même loi dépendant d'un paramètre inconnu $s \in \mathcal{S}$. On note $\gamma : \mathcal{S} \times \Xi \to \mathbb{R}$ une fonction de contraste telle que, pour tout $t \in \mathcal{S}$, $0 \leq \gamma(t, .) \leq 1$ et γ_n le contraste empirique défini par*

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^n \gamma(t, X_i).$$

Soit une collection dénombrable $\{S_m\}_{m\in\mathcal{M}}$ de parties dénombrables de $\mathcal{S}$ et une famille d'estimateurs de minimum de contraste associés $\{\hat{s}_m\}_{m\in\mathcal{M}}$. On considère une collection $\{x_m\}_{m\in\mathcal{M}}$ de poids positifs tels que

$$\sum_{m\in\mathcal{M}} e^{-x_m} = \Sigma < +\infty$$

et une fonction de pénalité $\mathrm{pen} : \mathcal{M} \to \mathbb{R}^+$ *(dépendant éventuellement des données). On désigne par* $\tilde{s} = \hat{s}_{\hat{m}}$ *un minimiseur du critère*

$$\gamma_n(\hat{s}_m) + \mathrm{pen}(m)$$

et par $\bar{\gamma}_n$ *le processus empirique centré*

$$\bar{\gamma}_n(t) = \gamma_n(t) - \mathbb{E}[\gamma(t, X_1)],\ t \in \mathcal{S}.$$

Alors, si pour une constante $C \geq 0$*, pour tout* $m \in \mathcal{M}$ *et tout* $z > 0$*,*

$$\mathrm{pen}(m) \geq \mathbb{E}\Big[\sup_{t \in S_m} (-\bar{\gamma}_n(t))\Big] + \sqrt{\frac{x_m}{2n}} - C\sqrt{\frac{z}{2n}}$$

avec probabilité au moins égale à $1 - \exp(-x_m - z)$*, on a*

$$\mathbb{E}[\ell(s, \tilde{s})] \leq \inf_{m \in \mathcal{M}} \Big(\ell(s, S_m) + \mathbb{E}[\mathrm{pen}(m)]\Big) + \Sigma(1 + C)\sqrt{\frac{\pi}{2n}}$$

où $\ell(s,t) = \mathbb{E}[\gamma(t, X_1) - \gamma(s, X_1)]$ *et* $\ell(s, S_m) = \inf_{t \in S_m} \ell(s,t)$*.*

B.3. Rappel sur l'heuristique de pente

Dans cette section, nous exposons brièvement le raisonnement qui conduit à l'heuristique de pente.

Nous considérons des observations $\xi_1, \ldots, \xi_n$, indépendantes et de même loi qu'une variable aléatoire générique ξ, et supposons que cette loi dépend d'un paramètre inconnu s à estimer. Soit $\{S_m\}_{m \in \mathcal{M}}$ une collection dénombrable de modèles et $\{\hat{s}_m\}_{m \in \mathcal{M}}$ une famille associée d'estimateurs de minimum de contraste de s. Nous cherchons une bonne fonction de pénalité telle que la minimisation du critère

$$\mathrm{crit}(m) = \gamma_n(\hat{s}_m) + \mathrm{pen}(m)$$

conduise à un estimateur $\hat{s}_{\hat{m}}$ satisfaisant. Tout d'abord, remarquons que pour une pénalité égale à $\ell(s, \hat{s}_m) - \gamma_n(\hat{s}_m)$, minimiser $\mathrm{crit}(m)$ revient à minimiser la perte $\ell(s, \hat{s}_m) = \mathbb{E}[\gamma(\hat{s}_m, \xi)] - \mathbb{E}[\gamma(s, \xi)]$, c'est-à-dire à sélectionner un oracle. Si pour tout m, s_m désigne un minimiseur de $\mathbb{E}[\gamma(t, \xi)]$ sur S_m,

$$\begin{aligned}\ell(s, \hat{s}_m) - \gamma_n(\hat{s}_m) = {} & \mathbb{E}[\gamma(\hat{s}_m, \xi)] - \mathbb{E}[\gamma(s_m, \xi)] + \mathbb{E}[\gamma(s_m, \xi)] - \mathbb{E}[\gamma(s, \xi)] \\ & + \gamma_n(s_m) - \gamma_n(\hat{s}_m) + \gamma_n(s) - \gamma_n(s_m) - \gamma_n(s). \qquad \text{(B.1)}\end{aligned}$$

Observons que $-\gamma_n(s)$ ne dépend pas de m, et considérons la pénalité

$$\mathrm{pen}^\star(m) = v_m + \hat{v}_m + \delta_n(s_m),$$

où $v_m = \mathbb{E}[\gamma(\hat{s}_m, \xi)] - \mathbb{E}[\gamma(s_m, \xi)]$ est un terme de variance, $\hat{v}_m = \gamma_n(s_m) - \gamma_n(\hat{s}_m)$ un terme de variance empirique et

$$\delta_n(s_m) = \mathbb{E}[\gamma(s_m, \xi)] - \mathbb{E}[\gamma(s, \xi)] - [\gamma_n(s_m) - \gamma_n(s)]$$

est la différence entre un terme de biais et son équivalent empirique. Pour tout $m \in \mathcal{M}$, l'égalité (B.1) peut se récrire

$$\ell(s, \hat{s}_m) + \gamma_n(s) = \gamma_n(\hat{s}_m) + \mathrm{pen}^\star(m).$$

Soient $\mathrm{pen} : \mathcal{M} \to \mathbb{R}^+$ une certaine fonction de pénalité et $\hat{m}$ le minimiseur du critère $\mathrm{crit}(m)$ correspondant. Par définition de $\hat{m}$, pour tout $m \in \mathcal{M}$,

$$\gamma_n(\hat{s}_{\hat{m}}) + \mathrm{pen}(\hat{m}) \leq \gamma_n(\hat{s}_m) + \mathrm{pen}(m).$$

Il vient

$$\begin{aligned}
\ell(s, \hat{s}_{\hat{m}}) &= \gamma_n(\hat{s}_{\hat{m}}) + \mathrm{pen}^\star(\hat{m}) - \gamma_n(s) \\
&\leq \gamma_n(\hat{s}_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + \mathrm{pen}^\star(\hat{m}) - \gamma_n(s) \\
&= \ell(s, \hat{s}_m) + \mathrm{pen}(m) - \mathrm{pen}^\star(m) - \mathrm{pen}(\hat{m}) + \mathrm{pen}^\star(\hat{m}).
\end{aligned}$$

Cette inégalité étant valable pour tout $m \in \mathcal{M}$,

$$\ell(s, \hat{s}_{\hat{m}}) + \mathrm{pen}(\hat{m}) - \mathrm{pen}^\star(\hat{m}) \leq \inf_{m \in \mathcal{M}} \{\ell(s, \hat{s}_m) + \mathrm{pen}(m) - \mathrm{pen}^\star(m)\},$$

ce qui indique que pour établir une inégalité oracle, la pénalité choisie doit être de l'ordre de $\mathrm{pen}^\star(m)$ pour tout m. Puisque $\mathrm{pen}^\star(m)$ dépend de s, il faut approcher cette pénalité idéale.

Pour ce faire, constatons en premier lieu que poser $\mathrm{pen}(m) = \hat{v}_m$ implique $\mathrm{crit}(m) = \gamma_n(s_m)$, de sorte que minimiser ce critère revient à minimiser le biais sans tenir compte de la variance. Dans ce cas, des modèles trop complexes risquent d'être sélectionnés. Il s'agit en fait d'une pénalité minimale, comme on le voit en posant $\mathrm{pen}(m) = \lambda \hat{v}_m$, ce qui entraîne $\mathrm{crit}(m) = (1-\lambda)\gamma_n(\hat{s}_m) + \lambda\gamma_n(s_m)$. Pour $\lambda < 1$, le modèle obtenu sera trop complexe car le critère décroît lorsque la complexité des modèles augmente, mais si $\lambda > 1$, pour des modèles assez complexes, qui présentent des biais similaires, le critère augmente avec la complexité, ce qui laisse espérer qu'un modèle correct pourra être sélectionné.

Ensuite, en supposant que les termes de variance v_m et $\hat{v}_m$ sont suffisamment proches l'un de l'autre, et en contrôlant $\delta_n(s_m)$ par un argument de concentration, $\mathrm{pen}^\star(m)$ peut être approchée par $2\hat{v}_m$, c'est-à-dire deux fois la pénalité minimale.

Finalement, si nous connaissons à une constante multiplicative près la forme d'une pénalité, en écrivant

$$2\hat{v}_m = 2[\gamma_n(s_m) - \gamma_n(\hat{s}_m)] = 2[\gamma_n(s_m) - \gamma_n(s)] + 2[\gamma_n(s) - \gamma_n(\hat{s}_m)],$$

nous constatons que, le terme de biais empirique restant relativement stable pour les modèles les plus complexes, le tracé de $-2\gamma_n(\hat{s}_m)$ en fonction de la forme de la pénalité permet d'évaluer le coefficient multiplicatif cherché.

C. Courbes paramétrées

Cette annexe rappelle quelques définitions et propriétés utiles relatives aux courbes paramétrées de $\mathbb{R}^d$. Pour une présentation plus complète, le lecteur est invité à se reporter au livre d'Alexandrov et Reshetnyak [7].

C.1. Définition

Soient $I \subset \mathbb{R}$ un intervalle et $\mathbf{f}$ l'arc paramétré défini par

$$\begin{aligned} \mathbf{f} : I &\to \mathbb{R}^d \\ t &\mapsto (f_1(t), \dots, f_d(t)), \end{aligned}$$

où $f_1, \dots, f_d$ sont des fonctions de I dans $\mathbb{R}$. On dit que $\mathbf{f}$ est de classe C^k lorsque les fonctions coordonnées $f_1, \dots, f_d$ sont de classe C^k.

C.2. Longueur et courbure

La définition suivante exprime que la longueur d'une courbe paramétrée est la borne supérieure des longueurs des lignes polygonales inscrites.

Définition C.2.1 (Longueur)**.** *La longueur d'une courbe* $\mathbf{f} : I \to \mathbb{R}^d$ *sur un intervalle* $[\alpha, \beta] \subset I$ *est donnée par*

$$\mathscr{L}(\mathbf{f}, \alpha, \beta) = \sup \sum_{j=1}^{m} \|\mathbf{f}(t_j) - \mathbf{f}(t_{j-1})\|,$$

où la borne supérieure est prise sur toutes les subdivisions $\alpha = t_0 < t_1 < \cdots < t_m = \beta$, $m \geq 1$.

Observons que, dans cette définition, la courbe $\mathbf{f}$ n'est pas supposée différentiable. La notion de courbure intégrale se définit de manière analogue.

Définition C.2.2 (Courbure intégrale)**.** *La courbure intégrale d'une courbe* $\mathbf{f} : I \to \mathbb{R}^d$ *sur* $[\alpha, \beta] \subset I$ *est définie par*

$$\mathscr{K}(\mathbf{f}, \alpha, \beta) = \sup \sum_{j=1}^{m-1} \widehat{\mathbf{f}(t_j)},$$

où $\widehat{\mathbf{f}(t_j)}$ *désigne l'angle entre les vecteurs* $\overrightarrow{\mathbf{f}(t_{j-1})\mathbf{f}(t_j)}$ *et* $\overrightarrow{\mathbf{f}(t_j)\mathbf{f}(t_{j+1})}$*, la borne supérieure étant prise sur toutes les subdivisions* $\alpha = t_0 < t_1 < \cdots < t_m = \beta$*,* $m \geq 1$.

Si nous considérons à présent des courbes de classe C^2, longueur et courbure peuvent s'exprimer en fonction des dérivées d'ordre 1 et 2.

Définition C.2.3 (Longueur)**.** *La longueur d'arc d'une courbe* $\mathbf{f}$ *de* α *à* β *est donnée par*

$$\mathscr{L}(\mathbf{f}, \alpha, \beta) = \int_\alpha^\beta \|\mathbf{f}'(t)\| \mathrm{d}t.$$

Différentes paramétrisations peuvent conduire à la même courbe. Plus précisément, si l'on applique une transformation monotone à t, le résultat ne change pas. Le paramètre naturel est l'abscisse curviligne.

Définition C.2.4 (Abscisse curviligne)**.** *Fixons* $t_0 \in I$*. L'abscisse curviligne* s *d'origine* t_0 *(et orientée dans le sens des* t *croissants) de la courbe* $\mathbf{f}$ *est définie par*

$$s(t) = \int_{t_0}^t \|\mathbf{f}'(u)\| \mathrm{d}u.$$

Remarquons que si $\|\mathbf{f}'(t)\| = 1$, on a $\mathscr{L}(\mathbf{f}, \alpha, \beta) = \beta - \alpha$. Or, si $\|\mathbf{f}'(t)\| > 0$, il est toujours possible de reparamétriser la courbe $\mathbf{f}$ de manière à avoir $\|\mathbf{f}'(t)\| = 1$. Cette paramétrisation particulière est appelée paramétrisation normale, paramétrisation par la longueur d'arc, ou encore paramétrisation par l'abscisse curviligne.

Définition C.2.5 (Courbure)**.** *La courbure de la courbe* $\mathbf{f}$ *à l'instant* t *est définie par*

$$k(t) = \|\mathbf{f}''(t)\|.$$

L'inverse de la courbure

$$r(t) = \frac{1}{k(t)} = \frac{1}{\|\mathbf{f}''(t)\|}$$

est appelé rayon de courbure de $\mathbf{f}$ *à* t*. La courbure intégrale de* $\mathbf{f}$ *est alors définie comme l'intégrale de la courbure, c'est-à-dire*

$$\mathscr{K}(\mathbf{f}, \alpha, \beta) = \int_\alpha^\beta \|\mathbf{f}''(t)\| \mathrm{d}t.$$

D. Quantization and clustering with Bregman divergences*

Abstract

This paper deals with the problem of quantization of a random variable X taking values in a separable and reflexive Banach space, and with the related question of clustering independent random observations distributed as X. To this end, we use a quantization scheme with a class of distortion measures called Bregman divergences, and provide conditions ensuring the existence of an optimal quantizer and an empirically optimal quantizer. Rates of convergence are also discussed.

D.1. Introduction

Bregman divergences are a broad class of dissimilarity measures indexed by strictly convex functions. Introduced in 1967 by Bregman [9], these proximity functions are useful in a wide range of areas, among which are statistical learning and data mining (Banerjee, Merugu, Dhillon and Ghosh [4], Cesa-Bianchi and Lugosi [11]), computational geometry (Nielsen, Boissonnat and Nock [27]), natural sciences, speech processing and information theory (Gray, Buzo, Gray and Matsuyama [19]). A lot of well-known proximity measures such as squared Euclidean, Mahalanobis, Kullback-Leibler and L^2 distances are particular cases of Bregman divergences. In $\mathbb{R}^d$, a Bregman divergence d_ϕ has the form

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle,$$

where $\langle\cdot,\cdot\rangle$ denotes the standard inner product, and $\nabla\phi(y)$ the gradient of ϕ at y. For example, taking $\phi(x) = \|x\|_2^2$ gives back the squared Euclidean distance. The same definition is valid in Hilbert spaces, and it even generalizes to Banach spaces by setting

$$d_\phi(x, y) = \phi(x) - \phi(y) - D_y\phi(x - y),$$

*Cette annexe reprend le Chapitre 1 de la première partie, sous sa version publiée dans la revue *Journal of Multivariate Analysis*.

with $D_y\phi$ the Fréchet derivative of ϕ at y (Alber and Butnariu [1], Frigyik, Srivastava and Gupta [16]; see also Jones and Byrne [20] and Csiszár [12]). Note that a Bregman divergence is not necessarily a true metric, since it may be asymmetric or fail to satisfy the triangle inequality. However, Bregman divergences fulfill an interesting projection property which generalizes the Hilbert projection on a closed convex set, as shown in Bregman [9] for the finite-dimensional setting and Alber and Butnariu [1] for the functional case. Recently, Banerjee, Merugu, Dhillon and Ghosh [4] have established a bijection between finite-dimensional Bregman divergences and exponential families, and shown that the standard k-means clustering algorithm (Lloyd [25]) generalizes to these divergences.

Following the approach of Banerjee et al. [4], we propose in the present paper to use this class of proximity measures for quantization and clustering purposes. Quantization, also called lossy data compression in information theory, is the problem of replacing data by an efficient and compact representation which allows one to reconstruct the original observations with a certain accuracy. More formally, for a fixed integer $k \geq 1$, a random variable X with distribution μ, taking values in a set $\mathcal{X}$, will be represented by a so-called k-quantizer $q(X)$. Here q is a Borel measurable mapping from $\mathcal{X}$ to a finite subset of $\mathcal{X}$ with at most k elements. The error committed when representing X by $q(X)$ is given by the distortion

$$W(\mu, q) = \mathbb{E}d(X, q(X)),$$

where $\mathbb{E}$ denotes expectation with respect to the distribution μ and $d(\cdot, \cdot)$ is called the distortion measure. For more information on quantization, we refer the reader to Gersho and Gray [17], Graf and Luschgy [18] and Linder [24]. In practice, the distribution μ is unknown, and $W(\mu, q)$ is replaced by the empirical criterion

$$W(\mu_n, q) = \frac{1}{n}\sum_{i=1}^{n} d(X_i, q(X_i)),$$

where $X_1, \ldots, X_n$ are independent random observations with distribution μ, and μ_n denotes the empirical measure associated with $X_1, \ldots, X_n$, i.e.,

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

for any Borel subset A of $\mathcal{X}$. In this context, the problem is called clustering and it consists in grouping data items in meaningful classes by minimizing $W(\mu_n, q)$ over all possible k-quantizers. In short, the goal is to find a data-based quantizer q_n such that the clustering risk $W(\mu, q_n)$ becomes "close" to the optimal risk $\inf_q W(\mu, q)$ as the size of the data set grows.

To date, most of the results pertaining to the clustering problem have been reported in the finite-dimensional case, that is when $\mathcal{X} = \mathbb{R}^d$ ($d \geq 1$) endowed with the Euclidean metric. However, in many applied problems, the data items are in the form of random functions rather than standard vectors, and this casts the problem into the general class of functional data clustering. Besides, Bregman divergences represent a natural tool for measuring proximity between infinite-dimensional objects, such as curves or even probability measures. For a comprehensive introduction to the topic of functional data analysis, see the book of Ramsay and Silverman [29]. In this functional statistics context, Biau, Devroye and Lugosi [7] investigate clustering with Hilbert norms and Laloë explores in [22] quantization and clustering with L^1 norms in Banach spaces.

In the present contribution, we go one step further and consider the problem of quantization and clustering when $d(\cdot, \cdot)$ is a general Bregman divergence $d_\phi(\cdot, \cdot)$ defined on a reflexive and separable Banach space E. Our approach extends and completes the results presented in [4], which focuses on the finite-dimensional setting and adopts a more algorithmic-oriented point of view. The paper is organized as follows. In Section 2, we set up notation and assumptions, and recall the relevant definitions. In Section 3, we provide conditions ensuring the existence of a minimizer q^* of the distortion $W(\mu, q)$ and its empirical counterpart q_n^*. Then, in Section 4, we focus on the convergence of the distortion and prove almost sure and L^1 convergence of $W(\mu, q_n^*)$ towards $W(\mu, q^*)$. Rates of convergence which do not depend on the dimension of E are also obtained, using Rademacher averages as complexity measures. For the sake of clarity, proofs are postponed to Section D.5.

D.2. Context and assumptions

In this section, we formally define Bregman divergences, quantization and k-means clustering. We first need some notation and assumptions. Throughout the paper, $(E, \|\cdot\|)$ will denote a separable and reflexive Banach space, and $\mathcal{C}$ will be a measurable convex subset of E. Whenever E is a Hilbert space, $\langle \cdot, \cdot \rangle$ will stand for its inner product. Recall that the relative interior of a convex set $\mathcal{C}$, denoted hereafter by $ri(\mathcal{C})$, is its interior with respect to the affine hull. Finally, we will write $\partial\mathcal{C}$ for the complement of $ri(\mathcal{C})$ in its closure $\overline{\mathcal{C}}$.

We are now in a position to state the general definition of a Bregman divergence in E (Alber and Butnariu [1], Frigyik, Srivastava and Gupta [16]).

Definition D.2.1. *Let $\mathcal{C}$ be a convex subset of E, and let $\phi : \mathcal{C} \to \mathbb{R}$ be strictly convex and twice continuously differentiable on $ri(\mathcal{C})$. The Bregman divergence*

associated with ϕ is defined by

$$d_\phi(x, y) = \phi(x) - \phi(y) - D_y\phi(x - y),$$

where $D_y\phi$ denotes the Fréchet derivative of ϕ at y.

In particular, when E is a Hilbert space, it reduces to

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle.$$

Although Bregman divergences are not true metrics, they satisfy some interesting properties, such as non-negativity and separation, convexity in the first argument and linearity. For a complete description and proofs of these basic properties, the reader is referred to Bregman [9], Nielsen, Boissonnat and Nock [27] and Frigyik, Srivastava and Gupta [15]. Table 1 collects the most common examples of Bregman divergences.

Now, let X be a random variable with distribution μ, taking values in $\mathcal{C}$. Throughout the paper, we make the following assumptions:

1. $\mathbb{E}\|X\| < +\infty$.
2. $\mathbb{E}X \in ri(\mathcal{C})$.
3. $\mathbb{E}|\phi(X)| < +\infty$ and, for all $c \in ri(\mathcal{C})$, $\mathbb{E}|D_c\phi(X)| < +\infty$. This implies in particular that $\mathbb{E}d_\phi(X, c) < +\infty$ for all c.

Let $k \geq 1$. As already mentioned in the introduction, a k-quantizer is a Borel measurable mapping $q : \mathcal{C} \subset E \to \mathbf{c}$, where $\mathbf{c} = \{c_1, \ldots, c_\ell\}$, $\ell \leq k$, is a subset of $ri(\mathcal{C})$ called its codebook. In the sequel, the elements of $\mathbf{c}$ will also be named the centers associated to q. Every $x \in \mathcal{C}$ is represented by a unique $\hat{x} = q(x) \in \mathbf{c}$ and q induces a partition of $\mathcal{C}$ in cells $S_1, \ldots, S_\ell$. Each cell S_j is made of the elements of $\mathcal{C}$ whose image by q is c_j. Every k-quantizer is characterized by its codebook $\mathbf{c} = \{c_1, \ldots, c_\ell\}$ and its partition cells $S_1, \ldots, S_\ell$.

The error committed when representing X by $q(X)$ is assessed by the distortion

$$W(\mu, q) = \mathbb{E}d_\phi(X, q(X)) = \int_{\mathcal{C}} d_\phi(x, q(x))\mathrm{d}\mu(x). \tag{D.1}$$

Let

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

where $\mathcal{Q}_k$ is the set of all k-quantizers. To get a representation that is as accurate as possible, we look for an optimal quantizer, i.e., a quantizer q^* satisfying

$$W(\mu, q^*) = W^*(\mu).$$

Bregman divergence	E	$\mathcal{C}$
Squared loss	$\mathbb{R}$	$\mathbb{R}$
Exponential loss	$\mathbb{R}$	$\mathbb{R}$
Norm-like	$\mathbb{R}$	$\mathbb{R}^+$
I-divergence (dim 1)	$\mathbb{R}$	$\mathbb{R}^+$
Logistic loss	$\mathbb{R}$	$[0,1]$
Itakura-Saito (dim 1)	$\mathbb{R}$	$(0,+\infty)$
Squared Euclidean distance	$\mathbb{R}^d$	$\mathbb{R}^d$
Mahalanobis distance	$\mathbb{R}^d$	$\mathbb{R}^d$
Kullback-Leibler (discrete)	$\mathbb{R}^d$	$(d-1)-$simplex
I-divergence (discrete)	$\mathbb{R}^d$	$(\mathbb{R}^+)^d$
Squared L^2 norm	$L^2(I,m)$	$L^2(I,m)$
Kullback-Leibler (continuous)	$L^2([0,1],dt)$	$\{x \in C^0([0,1]), \int_0^1 x(t)\mathrm{d}t = 1\}$
I-divergence (continuous)	$L^2([0,1],dt)$	$\{x \in C^0([0,1]), x \geq 0\}$
Itakura-Saito (continuous)	$L^2_{2\pi}(dt)$	$\{x \in C^0_{2\pi}, x > 0\}$

Bregman divergence	$\phi(x)$	$d_\phi(x,y)$
Squared loss	x^2	$(x-y)^2$
Exponential loss	e^x	$e^x - e^y - (x-y)e^y$
Norm-like	x^α	$x^\alpha + (\alpha-1)y^\alpha - \alpha x y^{\alpha-1}$
I-divergence (dim 1)	$x \ln x$	$x \ln \frac{x}{y} - (x-y)$
Logistic loss	$x \ln x + (1-x)\ln(1-x)$	$x \ln \frac{x}{y} + (1-x)\ln\left(\frac{1-x}{1-y}\right)$
Itakura-Saito (dim 1)	$-\ln x$	$\frac{x}{y} - \ln \frac{x}{y} - 1$
Squared Euclidean distance	$\|x\|_2^2$	$\|x-y\|_2^2$
Mahalanobis distance	${}^t x A x$	${}^t(x-y)A(x-y)$
Kullback-Leibler (discrete)	$\sum_{\ell=1}^d x_\ell \ln x_\ell$	$\sum_{\ell=1}^d x_\ell \ln \frac{x_\ell}{y_\ell}$
I-divergence (discrete)	$\sum_{\ell=1}^d x_\ell \ln x_\ell$	$\sum_{\ell=1}^d x_\ell \ln \frac{x_\ell}{y_\ell} - \sum_{\ell=1}^d (x_\ell - y_\ell)$
Squared L^2 norm	$\int_I x^2(t)\mathrm{d}m(t)$	$\|x-y\|_{L^2}^2$
Kullback-Leibler (continuous)	$\int_0^1 x(t)\ln x(t)\mathrm{d}t$	$\int_0^1 x(t)\ln\frac{x(t)}{y(t)}\mathrm{d}t$
I-divergence (continuous)	$\int_0^1 x(t)\ln x(t)\mathrm{d}t$	$\int_0^1 x(t)\ln\frac{x(t)}{y(t)} + y(t) - x(t)\mathrm{d}t$
Itakura-Saito (continuous)	$-\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln(x(\theta))\mathrm{d}\theta$	$-\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\ln\frac{x(\theta)}{y(\theta)} - \frac{x(\theta)}{y(\theta)} + 1\right)\mathrm{d}\theta$

TABLE D.1.: Some examples of Bregman divergences. The matrix A is supposed to be positive definite. The notation $L^2(I,m)$ stands for the set of square integrable functions on an interval $I \subset \mathbb{R}$, with respect to the positive measure m, $L^2_{2\pi}(dt)$ for the set of 2π-periodic square integrable functions, $C^0([0,1])$ denotes the set of continuous functions on $[0,1]$, and $C^0_{2\pi}$ the set of 2π-periodic continuous functions.

In a statistical context, we only have at hand independent random observations $X_1, \ldots, X_n$ with distribution μ. The empirical distortion associated with $X_1, \cdots, X_n$ is given by

$$W(\mu_n, q) = \frac{1}{n} \sum_{i=1}^{n} d_\phi(X_i, q(X_i)), \tag{D.2}$$

where μ_n is the empirical measure. Observe that this is just the distortion (D.1) calculated with μ_n instead of μ. Clustering data into k groups means looking for an optimal quantizer q_n^* with respect to the empirical distortion (D.2).

The codebook and partition characterize a quantizer. As in the Euclidean case, it is easy to show that among all quantizers with same codebook, the best one (with respect to the distortion) is the nearest neighbor quantizer, whose partition $S_1, \ldots, S_\ell$ is the Voronoi partition, i.e.,

$$S_1 = \{x \in \mathcal{C}, d_\phi(x, c_1) \leq d_\phi(x, c_p), p = 1, \ldots, \ell\}$$

and for $j = 2, \ldots, \ell$,

$$S_j = \{x \in \mathcal{C}, d_\phi(x, c_j) \leq d_\phi(x, c_p), p = 1, \ldots, \ell\} \setminus \bigcup_{m=1}^{j-1} S_m$$

(see Linder [24]). If an optimal quantizer exists, it is necessarily a nearest neighbor quantizer. Hence, in the sequel, we will always consider nearest neighbor quantizers. Conversely, given a partition $\{S_j\}_{j=1}^{\ell}$, with $\mu(S_j) > 0$ and $\mathbb{E}[X|X \in S_j] \in ri(\mathcal{C})$ for $j = 1, \ldots, \ell$, the best quantizer is obtained by setting

$$c_j \in \arg\min_{c \in ri(\mathcal{C})} \mathbb{E}[d_\phi(X, c)|X \in S_j] \quad \text{for } j = 1, \ldots, \ell.$$

The next proposition, proved in Section D.5, extends a result of Banerjee, Guo and Wang [3] to the case of functional Bregman divergences.

Proposition D.2.1. *Let d_ϕ be a Bregman divergence. If S is a Borel subset of $\mathcal{C}$ with $\mu(S) > 0$ and $\mathbb{E}[X|X \in S] \in ri(\mathcal{C})$, the function*

$$c \mapsto \mathbb{E}[d_\phi(X, c)|X \in S]$$

reaches its infimum at a unique element of $ri(\mathcal{C})$, namely $\mathbb{E}[X|X \in S]$.

Thus, for every Bregman divergence, the minimizer is the conditional expectation, just like for the squared Euclidean distance. Observe that it is the median instead of the expectation when the distortion measure is an L^1 norm.

Observe that the combination of Proposition D.2.1 and the optimality of the Voronoi partition is of computational interest. Indeed, even for the squared Euclidean

distance, minimizing the empirical distortion is generally a computationally hard problem, the complexity of an exact algorithm being exponential in the dimension of the space. In practice, a k-means type algorithm converging to local minima yields approximate solutions, and this adapts to general Bregman divergences. More precisely, given an initial codebook, which is made for instance of data items chosen at random, the algorithm proceeds by alternating between two steps. The first one consists in computing the Voronoi partition corresponding to the current centers. Then, during the second step, the new codebook is obtained by computing the mean of the data points falling in each cluster, according to Proposition D.2.1. For further information on k-means algorithms with Bregman divergences, see Banerjee et al. [4].

D.3. Existence of an optimal quantizer

In this section, we look for conditions ensuring the existence of an optimal quantizer q^*, i.e., a q^* such that $W(\mu, q^*) = W^*(\mu)$. Since a nearest neighbor quantizer is characterized by its codebook $\mathbf{c} = (c_1, \ldots, c_k)$, we may rewrite the distortion as

$$W(\mu, \mathbf{c}) = \mathbb{E} \min_{j=1,\ldots,k} d_\phi(X, c_j)$$

and look for an optimal codebook $\mathbf{c}^*$.

The existence of a minimum rests upon a compactness argument. We distinguish the finite-dimensional case (Theorem D.3.1) from the general case (Theorem D.3.2). In finite dimension, we prove the result by exploiting an idea of Sabin and Gray [31] based on Alexandroff one-point compactification (see, e.g., Dudley [14]).

Theorem D.3.1 (**Finite-dimensional case**). *Assume that the convex set $\mathcal{C}$ lies in a finite-dimensional affine space and that the following statements hold:*

1. *For all $x \in \mathcal{C}$, the function $y \mapsto d_\phi(x, y)$ is lower semi-continuous on $ri(\mathcal{C})$.*
2. *For all $(x, y) \in \mathcal{C} \times ri(\mathcal{C})$, $d_\phi(x, y) \leq \liminf_{z \to \tilde{z} \in \partial\mathcal{C}} d_\phi(x, z)$ for all $\tilde{z} \in \partial\mathcal{C}$.*
3. *For all $(x, y) \in \mathcal{C} \times ri(\mathcal{C})$, $d_\phi(x, y) \leq \liminf_{\|z\| \to +\infty} d_\phi(x, z)$.*

Then, there exists an optimal codebook $\mathbf{c}^*$, i.e.,

$$W(\mu, \mathbf{c}^*) = W^*(\mu).$$

Requirement 1 is not restrictive since $y \mapsto d_\phi(x, y)$ is continuous for most well-known Bregman divergences. Observe that ϕ and $y \mapsto D_y\phi$ are continuous on $ri(\mathcal{C})$, so that condition 1 could be replaced by lower semi-continuity of $y \mapsto D_y\phi(y)$. Roughly speaking, requirements 2 and 3 prevent a possible minimizer from running

to infinity. Note that condition 3 is void whenever $\mathcal{C}$ is bounded. In this case, $\overline{\mathcal{C}}$ is compact and the existence of an optimal codebook can easily be shown without resorting to Alexandroff compactification.

When E is potentially infinite dimensional and $\mathcal{C}$ is any convex subset of E, things are not so simple, since Alexandroff compactification only applies to locally compact spaces. As we know that E is locally compact if and only if it is finite dimensional (see for instance Dudley [14]), this tool is not suited to the infinite-dimensional case. However, since E is reflexive, a closed and bounded convex subset of E is compact for the weak topology $\sigma(E, E')$, that is the coarsest topology on E making all continuous linear forms on E continuous. Moreover, every weakly lower semi-continuous function reaches its minimum on a weakly compact set. Thus, if we know in advance that $\mathbf{c}^*$ is to be searched for in a closed and bounded convex set, an argument of continuity suffices to show the existence of $\mathbf{c}^*$. In the sequel, $\mathcal{C}_R \subset ri(\mathcal{C})$ will denote a closed and bounded convex set of diameter $2R$. For example, $\mathcal{C}_R = B(0, R) = \{x \in E, \|x\| \leq R\}$ the closed ball of center 0 and radius R. A key fact is that $X \in \mathcal{C}_R$ implies that $\mathbf{c}^* \in \mathcal{C}_R$ if it exists, by Bregman projection (Alber and Butnariu [1]).

For further details about weak convergence and lower semi-continuous and convex functions, the reader is referred to Brezis [10] and Rockafellar [30].

Theorem D.3.2 (**General case**). *Suppose that there exists $R > 0$ such that $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$, and that for all $x \in \mathcal{C}$, $y \mapsto d_\phi(x, y)$ is weakly lower semi-continuous on $\mathcal{C}_R$. Then, there exists an optimal quantizer.*

Example D.3.1 Convex functions which are lower semi-continuous for the norm are examples of weakly lower semi-continuous functions (see, e.g., [10]).

Observe that since the weak topology coincides with the norm topology in finite dimension, the term "weakly" in Theorem D.3.2 can be dropped whenever E is finite dimensional.

In fact, if we only have $\mathcal{C}_R \cap ri(\mathcal{C}) \neq \emptyset$ instead of $\mathcal{C}_R \subset ri(\mathcal{C})$, but ϕ is of Legendre type (see Rockafellar [30], and for the infinite-dimensional definition, Bauschke, Borwein and Combettes [6]), it remains possible to use Bregman projection to obtain the same result.

In the particular case where $d_\phi(\cdot, \cdot)$ is the squared distance induced by the inner product of a Hilbert space, it can be shown (see Section D.5) that it is sufficient to look for an optimal quantizer on a ball. Hence, Theorem D.3.2 admits the following corollary.

Corollary D.3.1. *Let E be a Hilbert space. If $\phi(\cdot) = \|\cdot\|^2$, there exists an optimal quantizer corresponding to the Bregman divergence $d_\phi(\cdot,\cdot)$.*

In the last part of this section, we turn to the existence of an empirically optimal quantizer. In other words, we will look for a minimizer $\mathbf{c}_n^*$ of the empirical distortion

$$W(\mu_n, \mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\dots,k} d_\phi(X_i, c_j).$$

Since the support of the empirical measure μ_n contains at most n points, it is included in a closed ball B_R. Thus, Theorem D.3.2 implies the following result.

Corollary D.3.2. *Assume that for all $x \in \mathcal{C}$, $y \mapsto d_\phi(x,y)$ is weakly lower semi-continuous. Then, there exists an empirically optimal quantizer.*

As above, the term "weakly" may be omitted when E is finite dimensional.

D.4. Convergence

D.4.1. Convergence of the distortion

Suppose that there exists an optimal codebook $\mathbf{c}_n^*$ that achieves the minimum of the empirical distortion $W(\mu_n, \mathbf{c})$. We turn our attention to the "true" distortion $W(\mu, \mathbf{c})$ for $\mathbf{c} = \mathbf{c}_n^*$ and would like to know whether this quantity gets close to the minimal distortion $W^*(\mu)$ as the number n of observations becomes large.

Assuming that $\mathbf{c}^*$ exists,

$$\begin{aligned} W(\mu, \mathbf{c}_n^*) - W^*(\mu) &= W(\mu, \mathbf{c}_n^*) - W(\mu, \mathbf{c}^*) \\ &= W(\mu, \mathbf{c}_n^*) - W(\mu_n, \mathbf{c}_n^*) + W(\mu_n, \mathbf{c}_n^*) - W(\mu, \mathbf{c}^*) \\ &\leq W(\mu, \mathbf{c}_n^*) - W(\mu_n, \mathbf{c}_n^*) + W(\mu_n, \mathbf{c}^*) - W(\mu, \mathbf{c}^*) \\ &\leq 2 \sup_{\mathbf{c} \in ri(\mathcal{C})^k} |W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c})|. \end{aligned}$$

Thus, if we intend to show that $W(\mu, \mathbf{c}_n^*)$ converges to $W^*(\mu)$ as n tends to infinity, it will be enough to prove that $\sup_{\mathbf{c} \in ri(\mathcal{C})^k} |W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c})|$ vanishes as n tends to infinity.

As in the previous section, we distinguish the finite-dimensional case (Theorem 4.1) from the general case (Theorem 4.2).

Theorem D.4.1 (**Finite-dimensional case**). *Assume that $\mathcal{C}$ lies in a finite-dimensional affine space and that the following statements hold:*

1. *The Bregman divergence $d_\phi(\cdot,\cdot)$ is continuous.*
2. *For all $x \in \mathcal{C}$, $\tilde{z} \in \partial\mathcal{C}$, $\lim_{z\to\tilde{z}\in\partial\mathcal{C}} d_\phi(x,z) = +\infty$.*
3. *For all $x \in \mathcal{C}$, $\lim_{\|z\|\to+\infty} d_\phi(x,z) = +\infty$.*
4. *For all $x \in \mathcal{C}$, the function $y \mapsto d_\phi(x,y)$ is convex on $ri(\mathcal{C})$.*

Then, if $\mathbf{c}_n^*$ is a minimizer of the empirical distortion,

$$\lim_{n\to+\infty} W(\mu,\mathbf{c}_n^*) = W^*(\mu) \quad a.s.$$

Note that the existence of $\mathbf{c}_n^*$ (and $\mathbf{c}^*$) is guaranteed under these assumptions.

In view of the definition of ϕ, the requirement 1 could be replaced by the continuity of $(x,y) \mapsto D_y\phi(x-y)$. Condition 4 is not necessarily satisfied for each Bregman divergence. For instance, the Itakura-Saito divergence $d_\phi(x,y) = \frac{x}{y} - \ln\frac{x}{y} - 1$ is not convex in the second argument.

As for the existence of an optimal quantizer, the infinite-dimensional setting requires further hypotheses, as expressed in the following theorem:

Theorem D.4.2 (**General case**). *Assume that for all $x \in \mathcal{C}$, $y \mapsto d_\phi(x,y)$ is weakly lower semi-continuous, so that there exists a minimizer $\mathbf{c}_n^*$ of the empirical distortion. If there exists $R > 0$ such that $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$, and $M = M(\phi,R) \geq 0$ such that, for all $c \in \mathcal{C}_R$, $\|D_c\phi\| \leq M$, then*

$$\lim_{n\to+\infty} W(\mu,\mathbf{c}_n^*) = W^*(\mu) \quad a.s.$$

and

$$\lim_{n\to+\infty} \mathbb{E}W(\mu,\mathbf{c}_n^*) = W^*(\mu).$$

Let us point out that the convergence results

$$\lim_{n\to+\infty} W(\mu,\mathbf{c}_n^*) = W^*(\mu) \quad a.s.$$

and

$$\lim_{n\to+\infty} \mathbb{E}W(\mu,\mathbf{c}_n^*) = W^*(\mu)$$

always hold when $\phi(\cdot) = \|\cdot\|^2$ (Biau, Devroye and Lugosi [7]).

Let us now discuss some examples.

Example D.4.1 Here $E = \mathbb{R}$, $\mathcal{C} = \mathbb{R}^+$ and $d_\phi(x, y) = x \ln \frac{x}{y} - (x - y)$. Let $x \in \mathcal{C}$. The map $y \mapsto x \ln \frac{x}{y} - (x - y)$ is continuous and convex on $ri(\mathcal{C}) = (0, +\infty)$ (its second derivative is $\frac{x}{y^2} \geq 0$) and tends to $+\infty$ as y tends to 0 or $+\infty$. Thus there exists a quantizer whose codebook achieves the minimum of the distortion $W(\mu, \mathbf{c})$ (Theorem D.3.1) as well as an empirically optimal quantizer (Corollary D.3.2). Moreover, if $\mathbf{c}_n^*$ is a minimizer of the empirical distortion, almost sure convergence of $W(\mu, \mathbf{c}_n^*)$ to $W^*(\mu)$ is ensured (Theorem D.4.1).

Example D.4.2 Let $E = \mathcal{C} = \mathbb{R}$ and $\phi(x) = e^x$, which yields $d_\phi(x, y) = e^x - e^y - (x - y)e^y$. The function $y \mapsto e^x - e^y - (x - y)e^y$ is continuous on $\mathbb{R}$. If $\mathbb{P}\{|X| \leq R\} = 1$, there exists an optimal quantizer (Theorem D.3.2), and since $\phi'(x) = e^x \leq e^R$ on $[-R, R]$, $W(\mu, \mathbf{c}_n^*)$ converges almost surely and in L^1 to $W^*(\mu)$ (Theorem D.4.2).

Example D.4.3 When $d_\phi(\cdot, \cdot)$ is the squared Euclidean distance, the existence of an optimal quantizer, almost sure and L^1 convergence of the distortion are guaranteed.

Example D.4.4 Here, $E = \mathbb{R}^d$, $\mathcal{C}$ is the $(d-1)$-simplex and $d_\phi(p, q) = \sum_{\ell=1}^d p_\ell \ln \frac{p_\ell}{q_\ell}$. The function $q = (q_1, \ldots, q_d) \mapsto \sum_{\ell=1}^d p_\ell \ln \frac{p_\ell}{q_\ell}$ is continuous and convex on $ri(\mathcal{C}) = \{(p_1, \ldots, p_d) \in (0, +\infty)^d, \sum_{\ell=1}^d p_\ell = 1\}$ and tends to $+\infty$ as one of the q_ℓ's tends to 0. Thus, there exists an optimal quantizer and we have almost sure convergence of the distortion.

Example D.4.5 Let $E = \mathcal{C} = L^2([0, 1], dt)$, and $d_\phi(x, y) = \int_0^1 (x(t) - y(t))^2 \mathrm{d}t$. This is a Hilbert norm; thus the existence of a minimizer of the distortion and the convergence are guaranteed.

Example D.4.6 Let $E = L^2([0, 1], dt)$ and let $\mathcal{C}$ be the set of all continuous non-negative elements of E. Here $d_\phi(p, q) = \int_0^1 [p(t) \ln \frac{p(t)}{q(t)} + q(t) - p(t)] \mathrm{d}t$. The map $q \mapsto d_\phi(p, q)$ is continuous and convex and therefore weakly semi-continuous. Assume that $\mathbb{P}\{r \leq \|X\| \leq R\} = 1$ $(r > 0)$. Then, there exists an optimal quantizer. Moreover, we have almost sure and L^1 convergence of the distortion.

D.4.2. Rates of convergence

The previous section indicates that $W(\mu, \mathbf{c}_n^*)$ gets close to the minimal distortion when the sample size grows. However, it gives no information about the rates of convergence. To address this question, let us first observe that minimizing

$$W(\mu, \mathbf{c}) = \mathbb{E} \min_{j=1,\ldots,k} d_\phi(X, c_j) = \mathbb{E} \min_{j=1,\ldots,k} \Big(\phi(X) - \phi(c_j) - D_{c_j}\phi(X - c_j)\Big)$$

is equivalent to minimizing the quantity

$$\overline{W}(\mu, \mathbf{c}) = \mathbb{E} \min_{j=1,\dots,k} \left(-\phi(c_j) - D_{c_j}\phi(X - c_j) \right).$$

Similarly, to $W(\mu_n, \mathbf{c})$, we associate

$$\overline{W}(\mu_n, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\dots,k} \left(-\phi(c_j) - D_{c_j}\phi(X_i - c_j) \right).$$

Since

$$W(\mu, \mathbf{c}_n^*) - \inf_{\mathbf{c} \in ri(\mathcal{C})^k} W(\mu, \mathbf{c}) = \overline{W}(\mu, \mathbf{c}_n^*) - \inf_{\mathbf{c} \in ri(\mathcal{C})^k} \overline{W}(\mu, \mathbf{c})$$

and

$$\begin{aligned} &\mathbb{E}\overline{W}(\mu, \mathbf{c}_n^*) - \inf_{\mathbf{c} \in ri(\mathcal{C})^k} \overline{W}(\mu, \mathbf{c}) \\ &\quad \leq \mathbb{E} \sup_{\mathbf{c} \in ri(\mathcal{C})^k} \left(\overline{W}(\mu_n, \mathbf{c}) - \overline{W}(\mu, \mathbf{c}) \right) + \mathbb{E} \sup_{\mathbf{c} \in ri(\mathcal{C})^k} \left(\overline{W}(\mu, \mathbf{c}) - \overline{W}(\mu_n, \mathbf{c}) \right) \end{aligned} \tag{D.3}$$

(see Lemma 8.2 in Devroye, Györfi and Lugosi [13]), we are done if we can find upper bounds for the uniform deviation

$$\mathbb{E} \sup_{\mathbf{c} \in ri(\mathcal{C})^k} \left(\overline{W}(\mu_n, \mathbf{c}) - \overline{W}(\mu, \mathbf{c}) \right).$$

(The second term of the right-hand side of (D.3) can indeed be bounded by an upper bound of the first term.) The next theorem may be proved by resorting to the Rademacher averages as a complexity measure for a function class (see, e.g., Bartlett, Boucheron, and Lugosi [5]).

Theorem D.4.3. *For $\mathcal{C}_R \subset ri(\mathcal{C})$, the following inequality holds:*

$$\begin{aligned} &\mathbb{E} \sup_{\mathbf{c} \in \mathcal{C}_R^k} \left(\overline{W}(\mu_n, \mathbf{c}) - \overline{W}(\mu, \mathbf{c}) \right) \\ &\qquad \leq \frac{2k}{\sqrt{n}} \Big(\sup_{c \in \mathcal{C}_R} | -\phi(c) + D_c\phi(c)| + \sup_{c \in \mathcal{C}_R} \|D_c\phi\| (\mathbb{E}\|X\|^2)^{1/2} \Big). \end{aligned}$$

Corollary D.4.1. *Suppose that for all $x \in \mathcal{C}$, $y \mapsto d_\phi(x, y)$ is weakly lower semi-continuous, which ensures the existence of an optimal codebook $\mathbf{c}_n^*$. Assume that there exists $R > 0$ such that $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$. If $| -\phi(c) + D_c\phi(c)|$ and $\|D_c\phi\|$ are uniformly bounded on $\mathcal{C}_R$ by $M_1 = M_1(\phi, R) \geq 0$ and $M_2 = M_2(\phi, R) \geq 0$ respectively, then*

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \leq \frac{4k}{\sqrt{n}} \left(M_1 + M_2 (\mathbb{E}\|X\|^2)^{1/2} \right),$$

and thus

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \leq \frac{4k}{\sqrt{n}} \left(M_1 + M_2 R \right).$$

Note that Corollary D.4.1 yields dimension-free upper bounds. This is worth pointing out since E is allowed to be high (or even infinite) dimensional.

Example D.4.7 In this example, we give the bounds obtained for some usual Bregman divergences. We assume throughout that there exists $R > 0$ such that $\mathbb{P}\{\|X\| \leq R\} = 1$.

1. Squared loss. For $\phi(x) = x^2$,

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \leq \frac{4k}{\sqrt{n}} \left(R^2 + 2R(\mathbb{E}|X|^2)^{1/2} \right),$$

   and then

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \leq \frac{12kR^2}{\sqrt{n}}.$$

2. Exponential loss. For $\phi(x) = e^x$,

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \leq \frac{4k(2R-1)e^R}{\sqrt{n}}.$$

3. Squared Euclidean distance. For the squared Euclidean norm $\phi(x) = \|x\|^2$,

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \leq \frac{12kR^2}{\sqrt{n}}.$$

4. Mahalanobis distance. For $\phi(x) = {}^t xAx$ with A positive definite,

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \leq \frac{12k\|A\|R^2}{\sqrt{n}}.$$

5. Squared L^2 distance. When ϕ is a squared L^2 norm,

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \leq \frac{12kR^2}{\sqrt{n}}.$$

Remark D.4.1. Some Bregman divergences, typically Kullback-Leibler, involve a logarithm, which prevents $||D_c\phi||$ from being uniformly bounded on a ball B_R. In order to circumvent this difficulty, a possible solution is to consider a class of elements of E satisfying the following assumption:
– In dimension 1, $0 < r \leq x \leq R < +\infty \quad a.s.$
– In dimension d ($2 \leq d \leq +\infty$), when the logarithm appears in a sum or an integral, $\sum_{\ell=1}^{d} \ln^2(x_\ell) \leq M(R)$ or $\int \ln^2(x(t))\mathrm{d}t \leq M(R)$.

Several conditions of this type can be found in the literature on Kullback-Leibler divergence. For instance, Jordan, Nguyen and Wainwright [21], who develop an estimation method for the Kullback-Leibler divergence, require an envelope condition or boundedness from above and below.

As an illustration, let $d_\phi(x,y) = \int_0^1 x(t) \ln \frac{x(t)}{y(t)} \mathrm{d}t$. Suppose that $\mathbb{P}\{\|X\| \leq R\} = 1$ for some $R > 0$ and that $\int_0^1 \ln^2(X(t))\mathrm{d}t \leq R^2$. Assuming that the codebooks belong to the same function class as X, we obtain

$$\mathbb{E}W(\mu, \mathbf{c}_n^*) - W^*(\mu) \leq \frac{2kR}{\sqrt{n}}(1+R).$$

D.5. Proofs

D.5.1. Proof of Proposition D.2.1

The proof of Banerjee, Guo and Wang [3] may be adapted to the infinite-dimensional case. We will check that $\mathbb{E}[X|X \in S]$ minimizes $\mathbb{E}[d_\phi(X,c)|X \in S]$ and that it is the only element of $ri(\mathcal{C})$ with this property. For $c \in ri(\mathcal{C})$,

$$\begin{aligned}
&\mathbb{E}[d_\phi(X,c)|X \in S] - \mathbb{E}[d_\phi(X, \mathbb{E}[X|X \in S])|X \in S] \\
&\quad = \mathbb{E}[\phi(X) - \phi(c) - D_c\phi(X-c) - \phi(X) + \phi(\mathbb{E}[X|X \in S]) \\
&\qquad + D_{\mathbb{E}[X|X\in S]}\phi(X - \mathbb{E}[X|X \in S])|X \in S] \\
&\quad = \phi(\mathbb{E}[X|X \in S]) - \phi(c) - D_c\phi(\mathbb{E}[X|X \in S] - c) \\
&\quad = d_\phi(\mathbb{E}[X|X \in S], c).
\end{aligned}$$

Indeed, expectation and differentiation intertwine, since the derivative is a continuous linear form (see, e.g., Proposition 1.1.6 in Arendt, Batty, Hieber and Neubrander [2]). However,

$$d_\phi(\mathbb{E}[X|X \in S], c) \geq 0$$

and $d_\phi(\mathbb{E}[X|X \in S], c) = 0$ if and only if $c = \mathbb{E}[X|X \in S]$. Hence,

$$\mathbb{E}[d_\phi(X,c)|X \in S] \geq \mathbb{E}[d_\phi(X, \mathbb{E}[X|X \in S])|X \in S],$$

and equality holds if and only if $c = \mathbb{E}[X|X \in S]$. Thus, $\mathbb{E}[X|X \in S]$ is the unique minimizer of the function $c \mapsto \mathbb{E}[d_\phi(X,c)|X \in S]$ on $ri(\mathcal{C})$.

D.5.2. Proof of Theorem D.3.1

Setting $d_\phi(x, \tilde{z}) = \liminf_{z \to \tilde{z} \in \partial\mathcal{C}} d_\phi(x,z)$ for all $x \in \mathcal{C}$ and $\tilde{z} \in \partial\mathcal{C}$, $y \mapsto d_\phi(x,y)$ extends to a lower semi-continuous function $\overline{\mathcal{C}} \to [0, +\infty]$. We compactify $\overline{\mathcal{C}}$ by adding a point at infinity ω. Let $\tilde{\mathcal{C}} = \overline{\mathcal{C}} \cup \{\omega\}$ denote the Alexandroff compactification

of $\overline{\mathcal{C}}$ (for details about the Alexandroff one-point compactification, see for instance Dudley [14]). By Tychonoff's theorem, (see, e.g., Dudley [14]) the product $\tilde{\mathcal{C}}^k$ is also compact. We set for all $x \in \mathcal{C}$, $d_\phi(x,\omega) = \liminf_{\|z\|\to+\infty} d_\phi(x,z)$. According to the assumptions, the function $y \mapsto d_\phi(x,y)$ from $\tilde{\mathcal{C}}$ to $[0,+\infty]$ is lower semi-continuous, that is the level set $\{c \in \tilde{\mathcal{C}}, d_\phi(x,c) \leq \lambda\}$ is closed for all $\lambda \in \mathbb{R}$. Since $\{\mathbf{c} \in \tilde{\mathcal{C}}^k, \min_{j=1,\ldots,k} d_\phi(x,c_j) \leq t\} = \bigcup_{j=1}^k \{\mathbf{c} \in \tilde{\mathcal{C}}^k, d_\phi(x,c_j) \leq t\}$, the level sets of $\mathbf{c} \mapsto \min_{j=1,\ldots,k} d_\phi(x,c_j)$ are also closed, i.e., this function is lower semi-continuous. Hence, for $\mathbf{c} \in \tilde{\mathcal{C}}^k$,

$$\begin{aligned}
\liminf_{\mathbf{c}'\to\mathbf{c}} W(\mu,\mathbf{c}') &= \liminf_{\mathbf{c}'\to\mathbf{c}} \int_{\mathcal{C}} \min_{j=1,\ldots,k} d_\phi(x,c'_j)\mathrm{d}\mu(x) \\
&\geq \int_{\mathcal{C}} \liminf_{\mathbf{c}'\to\mathbf{c}} \min_{j=1,\ldots,k} d_\phi(x,c'_j)\mathrm{d}\mu(x) \\
&\qquad \text{(by Fatou's Lemma)} \\
&\geq \int_{\mathcal{C}} \min_{j=1,\ldots,k} d_\phi(x,c_j)\mathrm{d}\mu(x) \\
&= W(\mu,\mathbf{c}).
\end{aligned}$$

Thus, $\mathbf{c} \mapsto W(\mu,\mathbf{c})$ is lower semi-continuous on the compact $\tilde{\mathcal{C}}^k$, and it reaches its minimum at some codebook $\mathbf{c}^*$. By conditions 2 and 3, we can assume that $\mathbf{c}^* \in ri(\mathcal{C})^k$: if not, we replace the coordinates which belong to $\partial\mathcal{C}$ or equal ω by elements of $ri(\mathcal{C})$. Therefore, there exists an optimal codebook $\mathbf{c}^*$, and the result is proved.

D.5.3. Proof of Theorem D.3.2

Since $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$, it suffices to look for a minimizer $\mathbf{c}^*$ of the distortion over $\mathcal{C}_R^k$. Indeed,

$$\forall c \in ri(\mathcal{C}), d_\phi(X,c) \geq d_\phi(X,\overline{c})$$

with $\overline{c}$ the Bregman projection [1] of c on $\mathcal{C}_R$. Thus, for any codebook $\mathbf{c} = (c_1,\ldots,c_k)$, if $\overline{\mathbf{c}} = (\overline{c}_1,\ldots,\overline{c}_k)$, $\mathbb{E}\min_{j=1,\ldots,k} d_\phi(X,c_j) \geq \mathbb{E}\min_{j=1,\ldots,k} d_\phi(X,\overline{c}_j)$, i.e., $W(\mu,\mathbf{c}) \geq W(\mu,\overline{\mathbf{c}})$. This shows that projecting any center on the closed and bounded convex set $\mathcal{C}_R$ can only reduce the distortion. Since E is reflexive, $\mathcal{C}_R$ is weakly compact by Kakutani's Theorem (see [10] for instance), and so is $\mathcal{C}_R^k$. Let us show that $W(\mu,\cdot)$ is weakly lower semi-continuous. The function $y \mapsto d_\phi(x,y)$ is weakly lower semi-continuous for all $x \in \mathcal{C}$. This means that for all $\lambda \in \mathbb{R}$, the level sets $\{c \in \mathcal{C}_R, d_\phi(x,c) \leq \lambda\}$ are weakly closed. As $\{\mathbf{c} \in \mathcal{C}_R^k, \min_{j=1,\ldots,k} d_\phi(x,c_j) \leq t\} = \bigcup_{j=1}^k \{\mathbf{c} \in \mathcal{C}_R^k, d_\phi(x,c_j) \leq t\}$, the level sets of $\mathbf{c} \mapsto \min_{j=1,\ldots,k} d_\phi(x,c_j)$ are weakly closed as well, and this function is weakly lower semi-continuous. If $\mathbf{c}'$ converges

weakly to $\mathbf{c}$,

$$\begin{aligned}\liminf_{\mathbf{c}'\to\mathbf{c}} W(\mu,\mathbf{c}') &= \liminf_{\mathbf{c}'\to\mathbf{c}} \int \min_{j=1,\dots,k} d_\phi(x,c'_j)\mathrm{d}\mu(x)\\ &\geq \int \liminf_{\mathbf{c}'\to\mathbf{c}} \min_{j=1,\dots,k} d_\phi(x,c'_j)\mathrm{d}\mu(x)\\ &\qquad \text{(Fatou's Lemma)}\\ &\geq \int \min_{j=1,\dots,k} d_\phi(x,c_j)\mathrm{d}\mu(x) = W(\mu,\mathbf{c})\end{aligned}$$

since $\mathbf{c}\mapsto \min_{j=1,\dots,k} d_\phi(x,c_j)$ is weakly lower semi-continuous. Thus $W(\mu,\cdot)$ is weakly lower semi-continuous on a weakly compact set, which implies that it reaches its minimum, i.e., there exists $\mathbf{c}^*\in\mathcal{C}_R^k$, such that $W(\mu,\mathbf{c}^*)=W^*(\mu)$.

D.5.4. Proof of Corollary D.3.1

The result follows from Theorem D.3.2 and from the following lemma whose proof is close to the first part of the proof of Theorem 1 in Linder [24] (see also Pollard [28]).

Lemma D.5.1. *Let d_ϕ be a Bregman divergence. Assume that the second derivative of $\phi : E\to\mathbb{R}$ is uniformly strongly positive, i.e., there exists $m=m(\phi)>0$ such that for all c, $D_c^2\phi(x,x)\geq m\|x\|^2$, and that there exists $M=M(\phi)$ such that for all c, $\|D_c^2\phi\|\leq M$. Then,*

$$\inf_{\mathbf{c}\in E^k} W(\mu,\mathbf{c}) = \inf_{\mathbf{c}\in B_R^k} W(\mu,\mathbf{c})$$

for some $R>0$.

Proof of Lemma D.5.1. By Taylor's formula, there exists z belonging to the open segment xy, such that

$$\phi(x)=\phi(y)+D_y\phi(x-y)+\frac{1}{2}D_z^2\phi(x-y,x-y).$$

Thus,

$$d_\phi(x,y)=\frac{1}{2}D_z^2\phi(x-y,x-y),$$

which implies, by assumption,

$$\frac{m}{2}\|x-y\|^2\leq d_\phi(x,y)\leq\frac{M}{2}\|x-y\|^2.$$

For an integer $\ell\geq 1$ and $\mathbf{c}^\ell=(c_1,\dots,c_\ell)$, let

$$w_\ell(\mathbf{c}^\ell)=\mathbb{E}\min_{j=1,\dots,\ell} d_\phi(x,c_j).$$

Let $W_\ell^*(\mu)$ denote the optimal distortion with respect to ℓ-quantizers. Since Corollary D.3.1 corresponds to Proposition D.2.1 when $k = 1$, we suppose $k \geq 2$. Moreover, we can assume that the support of μ contains at least k points (otherwise, we would not look for a k-quantizer), so $W_k^*(\mu) < W_{k-1}^*(\mu)$. Let $\varepsilon > 0$ be such that

$$\varepsilon < \frac{1}{2}(W_{k-1}^*(\mu) - W_k^*(\mu)) \tag{D.4}$$

and let $0 < r_1 < r_2$ such that

$$\frac{m}{2}(r_2 - r_1)^2 \mu(B_{r_1}) > W_k^*(\mu) + \varepsilon \tag{D.5}$$

and

$$2M \int_{B_{2r_2}^c} \|x\|^2 \mathrm{d}\mu(x) < \varepsilon. \tag{D.6}$$

We choose a codebook $\mathbf{c}^k = (c_1, \ldots, c_k)$ such that

$$w_k(\mathbf{c}^k) < W_k^*(\mu) + \varepsilon.$$

This implies

$$w_k(\mathbf{c}^k) < W_{k-1}^*(\mu) - \varepsilon,$$

which ensures that $c_1, \ldots, c_k$ are distinct. Assume that these elements are sorted in increasing order, that is $\|c_1\| \leq \cdots \leq \|c_k\|$. Then, $\|c_1\| \leq r_2$. To see this, suppose that $\|c_1\| > r_2$. This means that $\|c_j\| > r_2$ for all j. Thus, for $x \in B_{r_1}$,

$$\begin{aligned}
\min_{j=1,\ldots,k} d_\phi(x, c_j) &\geq \frac{m}{2} \min_{j=1,\ldots,k} \|x - c_j\|^2 \\
&\geq \frac{m}{2} \min_{j=1,\ldots,k} (\|c_j\| - \|x\|)^2 \\
&\geq \frac{m}{2}(r_2 - r_1)^2.
\end{aligned}$$

Hence, $W_k^*(\mu) + \varepsilon > \frac{m}{2}(r_2 - r_1)^2 \mu(B_{r_1})$, contradicting inequality (D.5). We will now show that for all j, $\|c_j\| \leq C r_2$ where $C = 2 + 3\sqrt{\frac{M}{m}} > 0$. Suppose that $\|c_k\| > C r_2$. For $x \in B_{2r_2}$,

$$d_\phi(x, c_1) \leq \frac{M}{2}(\|x\| + \|c_1\|)^2 \leq \frac{9}{2} M r_2^2$$

and

$$d_\phi(x, c_k) \geq \frac{m}{2}(\|c_k\| - \|x\|)^2 > \frac{m}{2}(C r_2 - 2r_2)^2 = \frac{9}{2} M r_2^2,$$

and thus

$$d_\phi(x, c_1) \leq d_\phi(x, c_k).$$

For $x \in B^c_{2r_2}$,

$$d_\phi(x, c_1) \leq \frac{M}{2}(\|x\| + \|c_1\|)^2 \leq 2M\|x\|^2.$$

Then

$$d_\phi(x, c_1) \leq d_\phi(x, c_k) + 2M\|x\|^2 \mathbf{1}_{\{x \in B^c_{2r_2}\}}. \tag{D.7}$$

Let $\mathbf{c}^{k-1} = (c_1, \ldots, c_{k-1})$ and let $\{S_j\}_{j=1}^k$ denote the Voronoi partition associated with the components of $\mathbf{c}^k$. We obtain

$$\begin{aligned} w_{k-1}(\mathbf{c}^{k-1}) &= \sum_{j=1}^{k} \int_{S_j} \min_{j=1,\ldots,k-1} d_\phi(x, c_j) \mathrm{d}\mu(x) \\ &\leq \sum_{j=1}^{k-1} \int_{S_j} d_\phi(x, c_j) \mathrm{d}\mu(x) + \int_{S_k} d_\phi(x, c_1) \mathrm{d}\mu(x) \\ &\leq \sum_{j=1}^{k} \int_{S_j} d_\phi(x, c_j) \mathrm{d}\mu(x) + 2M \int_{B^c_{2r_2}} \|x\|^2 \mathrm{d}\mu(x). \end{aligned}$$

The last statement follows from inequality (D.7). Then, by inequalities (D.6) and (D.4),

$$\begin{aligned} w_{k-1}(\mathbf{c}^{k-1}) &\leq w_k(\mathbf{c}^k) + \varepsilon \\ &\leq W_k^*(\mu) + 2\varepsilon \\ &< W_{k-1}^*(\mu). \end{aligned}$$

This contradicts the definition of $W_{k-1}^*(\mu)$. Thus, $w_k(\mathbf{c}^k) < W_k^*(\mu) + \varepsilon$ implies $(c_1, \ldots, c_k) \in (B_{Cr_2})^k$, and finally, setting $R = Cr_2$,

$$W_k^*(\mu) = \inf_{\mathbf{c}^k \in B_R^k} w_k(\mathbf{c}^k).$$

□

D.5.5. Proof of Theorem D.4.1

As mentioned earlier, to prove Theorem D.4.1, it is enough to show that $W(\mu_n, \cdot)$ converges uniformly to $W(\mu, \cdot)$ almost surely. The method is inspired from Sabin and Gray [31] again. As in the proof of Theorem D.3.1, we define the Bregman divergence $d_\phi(\cdot, \cdot)$ on $\mathcal{C} \times \tilde{\mathcal{C}}$ with $\tilde{\mathcal{C}}$ the Alexandroff compactification of $\overline{\mathcal{C}}$. The assumptions imply that the extended function $d_\phi(\cdot, \cdot)$ is continuous. Continuous convergence on a compact set is equivalent to uniform convergence (see, e.g., Theorem 3.1.9 in Lojasiewicz [26]). Hence, as $\tilde{\mathcal{C}}^k$ is compact, it suffices to show that if $\{\mathbf{c}_n\}_{n \geq 0}$ is a sequence of points in $\tilde{\mathcal{C}}^k$ converging to $\mathbf{c}$, then

$$\lim_{n \to +\infty} W(\mu_n, \mathbf{c}_n) = W(\mu, \mathbf{c}) \quad a.s.$$

By a theorem of Varadarajan (Theorem 11.4.1 in Dudley [14] for example), almost surely, the empirical measure μ_n converges weakly to μ. Since E is a separable Banach space, by Skorohod's Representation Theorem (see, e.g., Theorem 11.7.2 in [14]), there exist random variables Y and Y_n, defined on the same probability space, such that Y has distribution μ, Y_n has distribution μ_n, and Y_n converges to Y almost surely. Since the extended function $d_\phi(\cdot,\cdot)$ is continuous, $\min_{j=1,\dots,k} d_\phi(x_n, c_{nj})$ converges to $\min_{j=1,\dots,k} d_\phi(x, c_j)$ as $(x_n, \mathbf{c}_n)$ converges to $(x, \mathbf{c})$. Therefore, as $\mathbf{c}_n$ converges to $\mathbf{c}$, $\min_{j=1,\dots,k} d_\phi(Y_n, c_{nj})$ converges almost surely (and thus in distribution) to $\min_{j=1,\dots,k} d_\phi(Y, c_j)$. Moreover, for all c, $d_\phi(Y_n, c)$ converges to $d_\phi(Y, c)$ almost surely, thus also in distribution.

If for all $j = 1, \dots, k$, $c_j = \omega$ or $c_j \in \partial\mathcal{C}$, then $W(\mu, \mathbf{c}) = +\infty$. Besides, by Fatou's Lemma,

$$\liminf_{n\to+\infty} W(\mu_n, \mathbf{c}_n) = \liminf_{n\to+\infty} \mathbb{E} \min_{j=1,\dots,k} d_\phi(Y_n, c_{nj}) \geq \mathbb{E} \min_{j=1,\dots,k} d_\phi(Y, c_j) = W(\mu, \mathbf{c}).$$

Thus, $\lim_{n\to+\infty} W(\mu_n, \mathbf{c}_n) = +\infty = W(\mu, \mathbf{c})$.

Otherwise, let c_m be an element of $\mathbf{c}$ belonging to $ri(\mathcal{C})$. There exists in $ri(\mathcal{C})$ a regular convex polyhedron centered at c_m, containing the c_{nm}'s for large enough n (for example, an s-dimensional hypercube centered at c_m, where s denotes the dimension of the affine subspace spanned by $ri(\mathcal{C})$). Let $\mathcal{V}$ denote the finite set of its vertices. As the function $y \mapsto d_\phi(x, y)$ is assumed to be convex, for large enough n,

$$\min_{j=1,\dots,k} d_\phi(x, c_{nj}) \leq d_\phi(x, c_{nm}) \leq \sum_{v\in\mathcal{V}} d_\phi(x, v). \tag{D.8}$$

By the strong law of large numbers, almost surely, for all $v \in \mathcal{V}$,

$$\mathbb{E}d_\phi(Y_n, v) = \int d_\phi(x, v)\mathrm{d}\mu_n(x) = \frac{1}{n}\sum_{i=1}^{n} d_\phi(X_i, v)$$

tends, as $n \to +\infty$, to

$$\mathbb{E}d_\phi(X, v) = \mathbb{E}d_\phi(Y, v).$$

According to Theorem 3.6 in [8], for any $v \in \mathcal{V}$, $d_\phi(Y_n, v)$ is uniformly integrable. This implies by (D.8) that the variables $\min_{j=1,\dots,k} d_\phi(Y_n, c_{nj})$ are uniformly integrable as well. Therefore, by Theorem 3.5 in [8],

$$W(\mu_n, \mathbf{c}_n) = \mathbb{E} \min_{j=1,\dots,k} d_\phi(Y_n, c_{nj})$$

tends almost surely to $\mathbb{E}\min_{j=1,\dots,k} d_\phi(Y, c_j) = W(\mu, \mathbf{c})$, as desired.

D.5.6. Proof of Theorem D.4.2

Since $\mathbb{P}\{X \in \mathcal{C}_R\} = 1$, the centers stay in the closed and bounded convex set $\mathcal{C}_R$ as the proof of Theorem D.3.2 shows. Let Y and Y_n be the random variables with distribution μ and μ_n respectively, given by Skorohod's Theorem. Then, for all $\mathbf{c}$,

$$\begin{aligned} W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c}) &= \mathbb{E} \min_{j=1,\dots,k} d_\phi(Y_n, c_j) - \mathbb{E} \min_{j=1,\dots,k} d_\phi(Y, c_j) \\ &= \mathbb{E} \min_{j=1,\dots,k} (\phi(Y_n) - \phi(c_j) - D_{c_j}\phi(Y_n - c_j)) \\ &\quad - \mathbb{E} \min_{j=1,\dots,k} (\phi(Y) - \phi(c_j) - D_{c_j}\phi(Y - c_j)) \\ &\leq \mathbb{E}\phi(Y_n) - \mathbb{E}\phi(Y) + \mathbb{E}(- \min_{j=1,\dots,k} D_{c_j}\phi(Y_n - Y)) \\ &\leq \mathbb{E}\phi(Y_n) - \mathbb{E}\phi(Y) + M\mathbb{E}\|Y_n - Y\|. \end{aligned}$$

Yet, one has

$$\mathbb{E}\phi(Y_n) = \int \phi(x)\mu_n(\mathrm{d}x) = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i).$$

Thus, by the strong law of large numbers, $\mathbb{E}\phi(Y_n)$ converges to $\mathbb{E}\phi(X) = \mathbb{E}\phi(Y)$ almost surely. By the triangle inequality, $\|Y\| + \|Y_n\| - \|Y_n - Y\| \geq 0$. By Fatou's Lemma,

$$\liminf_{n\to+\infty} \mathbb{E}(\|Y\| + \|Y_n\| - \|Y_n - Y\|) \geq \mathbb{E} \lim_{n\to+\infty} (\|Y\| + \|Y_n\| - \|Y_n - Y\|) = 2\mathbb{E}\|Y\|.$$

Moreover, the law of large numbers implies that $\mathbb{E}\|Y_n\|$ converges to $\mathbb{E}\|Y\|$ almost surely. Thus,

$$\mathbb{E}\|Y_n - Y\| \underset{n\to+\infty}{\longrightarrow} 0 \quad a.s.$$

Hence, almost surely,

$$\sup_{\mathbf{c}\in\mathcal{C}_R^k} |W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c})| \underset{n\to+\infty}{\longrightarrow} 0$$

This completes the proof of the first statement.

We now turn to the second assertion. The following inequality (see [13]) shows that it suffices to prove that $\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k} (W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c}))$ vanishes as n tends to infinity:

$$\begin{aligned} &\mathbb{E}W(\mu, \mathbf{c}_n^*) - \inf_{\mathbf{c}\in\mathcal{C}_R^k} W(\mu, \mathbf{c}) \\ &\quad \leq \mathbb{E} \sup_{\mathbf{c}\in\mathcal{C}_R^k} (W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c})) + \mathbb{E} \sup_{\mathbf{c}\in\mathcal{C}_R^k} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c})) . \end{aligned}$$

As stated above,

$$W(\mu_n, \mathbf{c}) - W(\mu, \mathbf{c}) \leq \mathbb{E}\phi(Y_n) - \mathbb{E}\phi(Y) + M\mathbb{E}\|Y_n - Y\|,$$

for all $\mathbf{c} \in \mathcal{C}_R^k$. Moreover,

$$\mathbb{E}\phi(Y_n) - \mathbb{E}\phi(Y) = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i) - \mathbb{E}\phi(X),$$

and taking expectation with respect to the X_i's, we have

$$\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} \phi(X_i) - \mathbb{E}\phi(X)\Big) = 0.$$

It remains to show that the expectation (with respect to the X_i's) of $\mathbb{E}\|Y_n - Y\|$ tends to 0 as n tends to infinity. This can be done by a slight adaptation of the proof of Lemma 4.2. in Biau, Devroye and Lugosi [7].

D.5.7. Proof of Theorem D.4.3

We first recall the definition and some useful properties of the Rademacher averages. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables, that is independent random variables taking values in $\{-1, 1\}$ such that $\mathbb{P}\{\varepsilon_i = -1\} = \mathbb{P}\{\varepsilon_i = 1\} = \frac{1}{2}$, independent of $X_1, \ldots, X_n$. For a class $\mathcal{G}$ of functions from E to $\mathbb{R}$, the Rademacher averages of $\mathcal{G}$ are defined by

$$R_n(\mathcal{G}) = \mathbb{E}\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i g(X_i).$$

We will use the following properties:

1. For $a \in \mathbb{R}$, $R_n(a\mathcal{G}) = |a|R_n(\mathcal{G})$, where $a\mathcal{G} = \{ag, g \in \mathcal{G}\}$.
2. $R_n(|\mathcal{G}|) \leq R_n(\mathcal{G})$, where $|\mathcal{G}| = \{|g|, g \in \mathcal{G}\}$. (This property follows from the contraction principle of Ledoux and Talagrand [23].)
3. $R_n(\mathcal{G}_1 + \mathcal{G}_2) \leq R_n(\mathcal{G}_1) + R_n(\mathcal{G}_2)$, where $\mathcal{G}_1 + \mathcal{G}_2 = \{g_1 + g_2, (g_1, g_2) \in \mathcal{G}_1 \times \mathcal{G}_2)\}$.

Theorem D.4.3 is a consequence of the following lemma.

Lemma D.5.2. *For $c \in \mathcal{C}_R$, let ℓ_c denote the real-valued function defined by*

$$\ell_c(x) = -\phi(c) - D_c\phi(x - c), \quad x \in \mathcal{C}.$$

Then,

(i)

$$\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\Big(\overline{W}(\mu_n,\mathbf{c})-\overline{W}(\mu,\mathbf{c})\Big)\le 2\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\min_{j=1,\dots,k}\ell_{c_j}(X_i).$$

(ii)

$$\begin{aligned}&\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\min_{j=1,\dots,k}\ell_{c_j}(X_i)\\&\quad\le k\Big(\mathbb{E}\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_i D_c\phi(X_i)+\frac{1}{\sqrt{n}}\sup_{c\in\mathcal{C}_R}|-\phi(c)+D_c\phi(c)|\Big).\end{aligned}$$

(iii)

$$\mathbb{E}\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_i D_c\phi(X_i)\le\frac{1}{\sqrt{n}}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|(\mathbb{E}\|X\|^2)^{1/2}.$$

Proof of Lemma D.5.2. (i) Let $X_1',\dots,X_n'$ be an independent copy of $X_1,\dots,X_n$, independent of the Rademacher variables $\varepsilon_1,\dots,\varepsilon_n$. We have

$$\begin{aligned}&\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\Big(\overline{W}(\mu_n,\mathbf{c})-\overline{W}(\mu,\mathbf{c})\Big)\\&\quad=\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\Big(\frac{1}{n}\sum_{i=1}^n\min_{j=1,\dots,k}\ell_{c_j}(X_i)-\mathbb{E}\min_{j=1,\dots,k}\ell_{c_j}(X)\Big)\\&\quad=\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\Big(\frac{1}{n}\sum_{i=1}^n\min_{j=1,\dots,k}\ell_{c_j}(X_i)-\mathbb{E}\frac{1}{n}\sum_{i=1}^n\min_{j=1,\dots,k}\ell_{c_j}(X_i')\Big)\\&\quad=\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n\min_{j=1,\dots,k}\ell_{c_j}(X_i)-\frac{1}{n}\sum_{i=1}^n\min_{j=1,\dots,k}\ell_{c_j}(X_i')\Big|X_1,\dots,X_n\Big).\end{aligned}$$

By Jensen's inequality,

$$\begin{aligned}&\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\Big(\overline{W}(\mu_n,\mathbf{c})-\overline{W}(\mu,\mathbf{c})\Big)\\&\quad\le\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\Big(\min_{j=1,\dots,k}\ell_{c_j}(X_i)-\min_{j=1,\dots,k}\ell_{c_j}(X_i')\Big)\\&\quad=\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\Big(\min_{j=1,\dots,k}\ell_{c_j}(X_i)-\min_{j=1,\dots,k}\ell_{c_j}(X_i')\Big)\\&\quad\le\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\min_{j=1,\dots,k}\ell_{c_j}(X_i)+\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n(-\varepsilon_i)\min_{j=1,\dots,k}\ell_{c_j}(X_i')\\&\quad=2\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\min_{j=1,\dots,k}\ell_{c_j}(X_i).\end{aligned}$$

(ii) To obtain the inequality (ii), we argue by induction on k. For $k = 1$,

$$\begin{aligned}
&\mathbb{E}\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\ell_c(X_i)\\
&=\mathbb{E}\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_i(-\phi(c)-D_c\phi(X_i-c))\\
&\le\mathbb{E}\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_iD_c\phi(X_i)+\mathbb{E}\sup_{c\in\mathcal{C}_R}\frac{|-\phi(c)+D_c\phi(c)|}{n}\left|\sum_{i=1}^n\varepsilon_i\right|\\
&\le\mathbb{E}\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_iD_c\phi(X_i)+\frac{1}{n}\sup_{c\in\mathcal{C}_R}|-\phi(c)+D_c\phi(c)|\left(\mathbb{E}\Big(\sum_{i=1}^n\varepsilon_i\Big)^2\right)^{1/2}\\
&=\mathbb{E}\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_iD_c\phi(X_i)+\frac{1}{\sqrt{n}}\sup_{c\in\mathcal{C}_R}|-\phi(c)+D_c\phi(c)|,
\end{aligned}$$

using the fact that the ε_i's are independent. Assume that statement (ii) is true for $k-1$, and let us show that it is true for k. Let $\mathbf{c}^{k-1} = (c_1, \ldots, c_{k-1})$.

$$\begin{aligned}
&\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\min_{j=1,\ldots,k}\ell_{c_j}(X_i)\\
&=\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\min\Big(\ell_{c_k}(X_i),\min_{j=1,\ldots,k-1}\ell_{c_j}(X_i)\Big)\\
&=\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{2n}\sum_{i=1}^n\varepsilon_i\Big(\ell_{c_k}(X_i)+\min_{j=1,\ldots,k-1}\ell_{c_j}(X_i)-|\ell_{c_k}(X_i)-\min_{j=1,\ldots,k-1}\ell_{c_j}(X_i)|\Big),
\end{aligned}$$

since $\min(a,b) = (a+b)/2 - |a-b|/2$. By properties of Rademacher averages,

$$\begin{aligned}
&\mathbb{E}\sup_{\mathbf{c}\in\mathcal{C}_R^k}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\min_{j=1,\ldots,k}\ell_{c_j}(X_i)\\
&\le\mathbb{E}\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\ell_c(X_i)+\mathbb{E}\sup_{\mathbf{c}^{k-1}\in(\mathcal{C}_R)^{k-1}}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\min_{j=1,\ldots,k-1}\ell_{c_j}(X_i)\\
&=\mathbb{E}\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_iD_c\phi(X_i)+\frac{1}{\sqrt{n}}\sup_{c\in\mathcal{C}_R}|-\phi(c)+D_c\phi(c)|+\\
&\qquad(k-1)\Big(\mathbb{E}\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_iD_c\phi(X_i)+\frac{1}{\sqrt{n}}\sup_{c\in\mathcal{C}_R}|-\phi(c)+D_c\phi(c)|\Big)\\
&=k\Big(\mathbb{E}\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_iD_c\phi(X_i)+\frac{1}{\sqrt{n}}\sup_{c\in\mathcal{C}_R}|-\phi(c)+D_c\phi(c)|\Big),
\end{aligned}$$

which is the desired bound for k.

(iii) We have

$$\begin{aligned}\mathbb{E}\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_i D_c\phi(X_i) &= \mathbb{E}\sup_{c\in\mathcal{C}_R}\frac{1}{n}D_c\phi\left(\sum_{i=1}^n\varepsilon_i X_i\right)\\ &\leq \frac{1}{n}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|\mathbb{E}\left\|\sum_{i=1}^n\varepsilon_i X_i\right\|\\ &\leq \frac{1}{n}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|\left(\mathbb{E}\left\|\sum_{i=1}^n\varepsilon_i X_i\right\|^2\right)^{1/2}\\ &\leq \frac{1}{n}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|\left(\mathbb{E}\left(\sum_{i=1}^n\|X_i\|\right)^2\right)^{1/2}.\end{aligned}$$

Using the fact that the X_i's are independent and identically distributed,

$$\begin{aligned}\mathbb{E}\sup_{c\in\mathcal{C}_R}\frac{1}{n}\sum_{i=1}^n\varepsilon_i D_c\phi(X_i) &\leq \frac{1}{n}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|\left(\mathbb{E}\sum_{i=1}^n\|X_i\|^2\right)^{1/2}\\ &= \frac{1}{n}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|(n\mathbb{E}\|X\|^2)^{1/2}\\ &= \frac{1}{\sqrt{n}}\sup_{c\in\mathcal{C}_R}\|D_c\phi\|(\mathbb{E}\|X\|^2)^{1/2}.\end{aligned}$$

□

References

[1] Y. Alber and D. Butnariu. Convergence of Bregman projection methods for solving consistent convex feasibility problems in reflexive Banach spaces. *Journal of Optimization Theory and Applications*, 92:33–61, 1997.

[2] W. Arendt, J. K. Batty, M. Hieber, and F. Neubrander. *Vector-valued Laplace Transforms and Cauchy Problems*. Monographs in Mathematics. Birkhäuser, Basel, 2001.

[3] A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51, 2005.

[4] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[5] P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2001.

[6] H. H. Bauschke, J. M. Borwein, and P. L. Combettes. Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces. *Communications in Contemporary Mathematics*, 3:615–647, 2001.

[7] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54:781–790, 2008.

[8] P. Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, 1999.

[9] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[10] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer, New York, 2010.

[11] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, 2006.

[12] I. Csiszár. Generalized projections for non-negative functions. *Acta Mathematica Hungarica*, 68:161–185, 1995.

[13] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Applications of Mathematics. Springer, New York, 1996.

[14] R. M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2002.

[15] B. A. Frigyik, S. Srivastava, and M. R. Gupta. An introduction to functional derivatives. Technical Report UWEETR-2008-0001, Department of Electrical Engineering, University of Washington, Seattle, 2008.

[16] B. A. Frigyik, S. Srivastava, and M. R. Gupta. Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 54:5130–5139, 2008.

[17] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Norwell, 1992.

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

[19] R. M. Gray, A. Buzo, A. H. Gray, and Y. Matsuyama. Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28:367–376, 1980.

[20] L. Jones and C. Byrne. General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis. *IEEE Transactions on Information Theory*, 36, 1990.

[21] M. I. Jordan, X. Nguyen, and M. J. Wainwright. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56:5847–5861, 2010.

[22] T. Laloë. L1-quantization and clustering in Banach spaces. *Mathematical Methods of Statistics*, 19:136–150, 2010.

[23] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer, Berlin, Heidelberg, 1991.

[24] T. Linder. Learning-theoretic methods in vector quantization. In L. Györfi, editor, *Principles of Nonparametric Learning*. Springer-Verlag, Wien, 2002.

[25] S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[26] S. Lojasiewicz. *An Introduction to the Theory of Real Functions*. John Wiley and Sons, New York, 1988.

[27] F. Nielsen, J.D. Boissonnat, and R Nock. Bregman Voronoi diagrams: properties, algorithms and applications. Technical Report 6154, INRIA, 2007.

[28] D. Pollard. Quantization and the method of k-means. *IEEE Transactions on Information Theory*, 28, 1982.

[29] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, 2006.

[30] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, New Jersey, 1970.

[31] M. J. Sabin and R. M. Gray. Global convergence and empirical consistency of the Generalized Lloyd Algorithm. *IEEE Transactions on Information Theory*, 32:148–155, 1986.

E. On the number of groups in clustering*

Abstract

This paper deals with the choice of the number k of groups in clustering. The approach consists in selecting k by minimizing a criterion defined by adding a penalty function to the empirical distortion. A result relying on model selection tools yields a penalty shape and ensures a control of the distortion of the codebook obtained by minimization of this penalized criterion. The method is then illustrated on simulated and real-life data.

E.1. Introduction

Clustering consists in dividing data into a finite number of relevant classes, so that items in the same group are as similar as possible, and items in different groups are as dissimilar as possible (Duda, Hart and Stork [15]). This unsupervised learning technique has been widely used for statistical data analysis in a variety of areas. For $k \geq 1$, the so-called k-means clustering method allows to partition a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with values in $(\mathbb{R}^d, \|\cdot\|)$ in k groups by minimizing the empirical distortion

$$W(\mu_n, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|\mathbf{X}_i - c_j\|^2$$

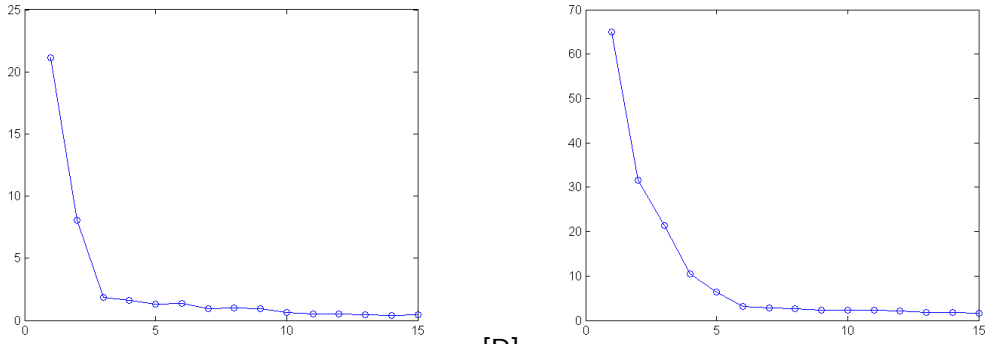
over all possible codebooks $\mathbf{c} = (c_1, \ldots, c_k) \in (\mathbb{R}^d)^k$. An essential problem is that of selecting the right number k of clusters. Indeed, if in some situations the choice of k may be motivated by the applications, it is in general unknown. Consider the empirical distortion

$$W(\mu_n, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{c \in \mathbf{c}} \|\mathbf{X}_i - c\|^2,$$

defined for every $\mathbf{c}$ with a number k of components such that $1 \leq k \leq n$. It seems natural to use this quantity depending on the observations in order to choose k.

*Cette annexe reprend le Chapitre 3 de la première partie.

However, the empirical distortion is decreasing in k (see Figure E.1). Indeed, the more groups there are, the nearer each observation is from the center of one of the groups. Minimizing the distortion directly would lead to choose k as large as possible, which does not present much interest (think, for example, of the situation where $k = n$, and each single observation builds a cluster). Moreover, Hastie, Tibshirani and Friedman [20] observe that it is not enough to evaluate the criterion on an independent test data set, since numerous centers would anyway densely fill the whole data space so that each observation is very close to one of them. Thus, it is not possible to select k by cross validation like in supervised learning. Nevertheless, the empirical distortion tends to decrease more strongly when increasing k leads to separate two existing classes in the data structure, than when partitioning one group into artificial subgroups.



[A] [B]

FIGURE E.1.: Graph of the empirical distortion as a function of k for two examples: [A] 3 clusters and [B] 6 clusters.

Most methods proposed in the literature for choosing k involve the empirical distortion. A presentation of various procedures can be found in Milligan and Cooper [29] and Hardy [18], while Gordon compares in [17] the performances of the five best rules exposed in [29]. These methods can be divided in two main types, global or local. Global procedures consist in performing clustering for different values of k and then retaining the value minimizing or maximizing some function of k. In local procedures, it must be decided at each step whether a cluster should be partitioned (or two groups merged into a single one).

Calinski and Harabasz [13] propose to choose the value of k maximizing an index based on the quotient

$$\frac{B(k)/(k-1)}{W(k)/(n-k)}.$$

Here, to underline the dependency in k, $W(k)$ denotes the empirical distortion or

within sum of squares and $B(k) = \sum_{j=1}^{k} \|c_j - \bar{c}\|^2$ the between sum of squares, where $\bar{c}$ is the mean of the data. The method by Krzanowski and Lai [24] consists in maximizing $W(k)k^{2/d}$, or more precisely the related quantity

$$\Big|\frac{\mathrm{DIFF}(k)}{\mathrm{DIFF}(k+1)}\Big|,$$

where

$$\mathrm{DIFF}(k) = W(k-1)(k-1)^{2/d} - W(k)k^{2/d},$$

whereas in Hartigan's rule [19], a new cluster is added as long as the quantity

$$H(k) = \left(\frac{W(k)}{W(k+1)} - 1\right)(n-k-1)$$

is sufficiently large. The *Silhouette* statistic of Kaufman and Rousseeuw [21] is given by

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))},$$

where $a(i)$ is the average distance between $\mathbf{X}_i$ and the observations belonging to the same cluster as $\mathbf{X}_i$, and $b(i)$ is the average distance between $\mathbf{X}_i$ and the observations in the nearest cluster (i.e. the cluster minimizing $b(i)$). An observation $\mathbf{X}_i$ is well clustered when $s(i)$ is large, and Kaufman and Rousseeuw [21] suggest to choose the value of k maximizing the average of $s(i)$ for $i = 1, \ldots, n$. The *Gap Statistic* of Tibshirani, Walther and Hastie [37] compares the evolution of the logarithm of the distortion for the considered clustering problem with the function obtained for uniformly distributed observations. Kim, Park and Park [23] develop an index which allows to select $\hat{k}$ by combining two functions of opposite monotonicity presenting a jump around the optimal value of k, whereas Sugar and James [36] propose to apply to the empirical distortion a transformation $w \mapsto w^{-p}, p > 0$. Observe that there also exist methods based on the stability of the partitions obtained. In this case, $\hat{k}$ is selected thanks to a criterion which has been determined using the clustering results obtained when considering several subsamples of the data set (see, e.g., Levine and Domany [25] and Ben-Hur, Elisseeff and Guyon [9]). The relation between the number k and cluster stability has been investigated from a theoretical point of view in Shamir and Tishby [33, 34], Ben-David, Luxburg, and Pál [8], Ben-David, Pál, and Simon [6] and Ben-David and Luxburg [7].

The method proposed in this paper to evaluate k is based on the empirical distortion and relies on the model selection theory introduced by Birgé and Massart [11] and Barron, Birgé, Massart [3].

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent random vectors, with the same distribution as a generic random vector $\mathbf{X}$ taking its values in $\mathbb{R}^d$ endowed with its standard Euclidean norm $\|\cdot\|$. Assume that, for some $R > 0$,

$$\mathbb{P}\{\|\mathbf{X}\| \leq R\} = 1. \tag{E.1}$$

The rest of the paper is organized as follows. In Section 2, we obtain a penalty shape and show an inequality ensuring a control of the risk of the estimator obtained by minimization of the corresponding penalized criterion. Section 3 presents some simulations and real data experiments illustrating the practical implementation of the proposed approach. The proof of the main result is postponed to Section 4 for the sake of clarity.

E.2. The choice of k

Recall that, for every k, the minimization of the empirical distortion

$$W(\mu_n, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{c \in \mathbf{c}} \|\mathbf{X}_i - c\|^2$$

yields some codebook with k components. The aim is to select the best one over all possible values of k.

More formally, for all k, $1 \leq k \leq n$, let S_k denote the (countable) set of all $(c_1, \ldots, c_k) \in \mathcal{Q}^k$, where $\mathcal{Q}$ is some grid over $\mathbb{R}^d$. Observe that, in practice, an algorithm can only provide centers belonging to such a grid. For every k, let $\hat{\mathbf{c}}_k$ be a minimizer of the criterion $W(\mu_n, \mathbf{c})$ over S_k, i.e. $\hat{\mathbf{c}}_k \in \arg\min_{\mathbf{c} \in S_k} W(\mu_n, \mathbf{c})$. To determine the best codebook in $\{\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_n\}$, we will search for $\hat{k}$, minimizing a criterion of the type

$$\mathrm{crit}(k) = W(\mu_n, \hat{\mathbf{c}}_k) + \mathrm{pen}(k),$$

where $\mathrm{pen} : \{1, \ldots, n\} \to \mathbb{R}^+$ is a penalty function, whose role is to avoid the choice of a too large k. For such a function $\mathrm{pen}(k)$, let $\tilde{\mathbf{c}} = \hat{\mathbf{c}}_{\hat{k}}$.

The particularity of the problem is that there is no relevant aim $\mathbf{c}^*$ here. Note that, contrary to mixture models, we have no information about the distribution of $\mathbf{X}$. If we set $\mathbf{c}^* \in \arg\min W(\mu, \mathbf{c})$, where the arg min is over all vectors $\mathbf{c}$ having a number of components at most n, the number of components of $\mathbf{c}^*$ will anyway be maximal, that is equal to n. In other words, there is no variance term in this situation, so that minimizing the risk of an estimator amounts to minimize its bias.

In a way, the penalty added to the empirical distortion can be seen as a "fictive" variance term.

In order to design a penalty, we may adapt Theorem 8.1 in Massart [27]. The proof of Theorem E.2.1 below relies on the following upper bound established by Linder [26] (see also Bartlett, Linder and Lugosi [4]):

$$\mathbb{E}\left[\sup_{\mathbf{c}\in S_k}(W(\mu,\mathbf{c})-W(\mu_n,\mathbf{c}))\right]\leq A(d,R)\sqrt{\frac{k}{n}},$$

where $A(d,R)$ only depends on the dimension d and the radius R of the ball appearing in the so-called peak power constraint (E.1).

Théorème E.2.1. *Consider a family of nonnegative weights* $\{x_k\}_{1\leq k\leq n}$ *such that*

$$\sum_{k=1}^{n} e^{-x_k}=\Sigma.$$

If for every $1\leq k\leq n$,

$$\mathrm{pen}(k)\geq A(d,R)\sqrt{\frac{k}{n}}+4R^2\sqrt{\frac{x_k}{2n}}, \tag{E.2}$$

then

$$\mathbb{E}\left[W(\mu,\tilde{\mathbf{c}})\right]\leq \inf_{1\leq k\leq n}\Big(W(\mu,S_k)+\mathrm{pen}(k)\Big)+R^2\Sigma\sqrt{\frac{2\pi}{n}}, \tag{E.3}$$

where $W(\mu,S_k)=\inf_{\mathbf{c}\in S_k}W(\mu,\mathbf{c})$.

Some comments are in order here.

Theorem E.2.1 suggests a penalty function $\mathrm{pen}(k)$ tending to 0 at the rate $1/\sqrt{n}$ and provides an upper bound for the expectation of the distortion at $\tilde{\mathbf{c}}$, the set of centers obtained by minimizing the penalized criterion $\mathrm{crit}(k)=W(\mu_n,\hat{\mathbf{c}}_k)+\mathrm{pen}(k)$. Inequality (E.3) ensures that if the penalty function is large enough, the expectation of the distortion at $\tilde{\mathbf{c}}$ remains relatively low, close to the smallest value of the distortion over k, up to a vanishing term.

The term of the order $\sqrt{k/n}$ in the penalty reflects the complexity of the models since a model here is more complex when k is larger. Let us point out that the proof of Theorem 3 in Linder [26] shows that we can set

$$A(d,R)=96R^2\sqrt{d}.$$

However, this value of the constant $A(d, R)$, which is an upper bound, is of limited interest in practice, all the more so since it involves the radius R. In fact, Theorem E.2.1 gives the shape of the penalty rather than an exact penalty function.

Consider the weights $\{x_k\}_{1\leq k\leq n}$. The larger they are, the smaller Σ is. Nevertheless, they should not be too large since they also appear in the penalty. In the Gaussian linear model selection framework, where each model S_m, $m \in \mathcal{M}$, has dimension D_m, a possible choice when there is no redundancy in the models dimension consists in taking x_m proportional to D_m (see Massart [27], Section 4.2.1). By analogy, the weights may here be taken proportional to k. Since the cardinality of the collection of models is at most n, another possibility would be to set $x_k = \ln n$ for every k, which does not change the penalty shape and leads to $\Sigma = 1$. Consequently, to a first approximation, the penalty is proportional to $\sqrt{k/n}$.

Then, let $\rho \geq 0$ (typically, $\rho = n^{-2}$). If for all k, $\hat{\mathbf{c}}_k$ is an approximate minimizer of the empirical risk $W(\mu_n, \mathbf{c})$, in the sense that for all $\mathbf{c} \in S_k$,

$$W(\mu_n, \hat{\mathbf{c}}_k) \leq W(\mu_n, \mathbf{c}) + \rho,$$

Theorem E.2.1 remains true provided one adds ρ in the right-hand term of the inequality.

Finally, observe that Theorem E.2.1 adapts to the case of observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ taking their values in a separable Hilbert space $(\mathcal{H}, \|\cdot\|)$, replacing the first term in the right-hand side in inequality (E.2) by a term proportional to $k/\sqrt{n}$. Indeed, there exists some constant $a(R)$ such that

$$\mathbb{E}\left[\sup_{\mathbf{c}\in S_k} \left(W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c})\right)\right] \leq a(R)\frac{k}{\sqrt{n}}$$

(Biau, Devroye and Lugosi [10]). More generally, the Euclidean (or Hilbert) norm can be replaced by another distortion measure, as soon as there exists an upper bound for the expected maximal deviation

$$\mathbb{E}\left[\sup_{\mathbf{c}\in S_k} \left(W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c})\right)\right].$$

E.3. Experimental results

In the present section, we propose to illustrate the choice of the number k of clusters using the penalized criterion suggested by Theorem E.2.1 on some simulated and real data examples. As mentioned earlier, we can assume to a first approximation

that the penalty shape given by Theorem E.2.1 is $c\sqrt{k/n}$, where c is a constant which has to be determined in practice. To this end, we will use the so-called slope heuristics, introduced by Birgé and Massart [12] and further developed by Arlot and Massart [2]. This method precisely allows calibrating a penalty known up to a multiplicative constant. An essential condition for applying the method is that the empirical contrast is a decreasing function of the complexity of the models and the penalty shape an increasing function. This condition is clearly satisfied in our clustering framework. Nevertheless, because of the lack of variance and aim mentioned above, our problem does not seem really suited to the use of the slope heuristics. However, this method turns out to perform quite well in the examples below, though there is not much theoretical justification in this clustering context. Two techniques based on the slope heuristics may be employed in order to calibrate a penalty. The dimension jump method consists in identifying an abrupt jump in the models complexity, whereas the other possibility is to observe that the empirical contrast is proportional to the penalty shape for complex models and compute the slope of this line. Both methods have been implemented in MATLAB by Baudry, Maugis and Michel [5] as an interface called CAPUSHE (CAlibrating Penalty Using Slope HEuristics).

From a computational point of view, the k-means algorithm used in all examples is initialized by taking as the unique center for $k = 1$ the mean of all observations. Then, at the kth step, a new center chosen uniformly at random among the observations is added to the $k - 1$ centers resulting from the previous step—the algorithm is therefore random. This procedure is repeated 50 times and the set of centers yielding the lower distortion is kept. Let us point out that there is an abundant literature about the interesting problem of the initialization of the k-means algorithm and that the strategy could be replaced by a more robust one (see, e.g., Pena, Lozano and Larranaga [31], Su and Dy [35], Khan and Ahmad [22], Perim, Wandekokem and Varejão [32], Al-Shboul and Myaeng [1]).

E.3.1. Simulated data

E.3.1.1. Some different numbers of groups and dimensions

In a first series of simulations, we tried to recover the number of clusters for 5 types of samples, which are different by dimension d and underlying number k of groups. In these examples, the right number of clusters is found in general. We observe that the direct slope estimation and the dimension jump method perform approximately similarly.

G1 A single group. We begin with a situation where it is not relevant to cluster the data and consider 200 points distributed uniformly in the unit hypercube in dimension 10.

G2 3 groups in dimension 2. The observations were sampled following a normal bivariate distribution with variance the identity matrix and build 3 groups, centered at $(0,0),(0,6)$ and $(5,-3)$ respectively, each containing 30 observations (see Figure E.2). An example of CAPUSHE output for this data set is visible in Figure E.3.

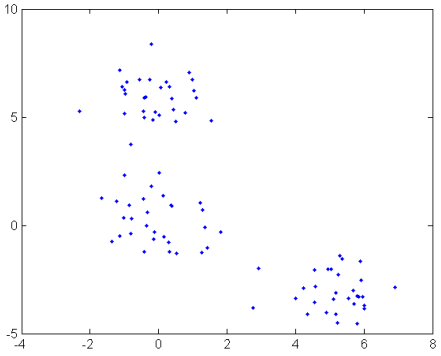


FIGURE E.2.: Groups of 30 observations following a normal distribution centered at $(0,0),(0,6)$ and $(5,-3)$.

[A] [B]

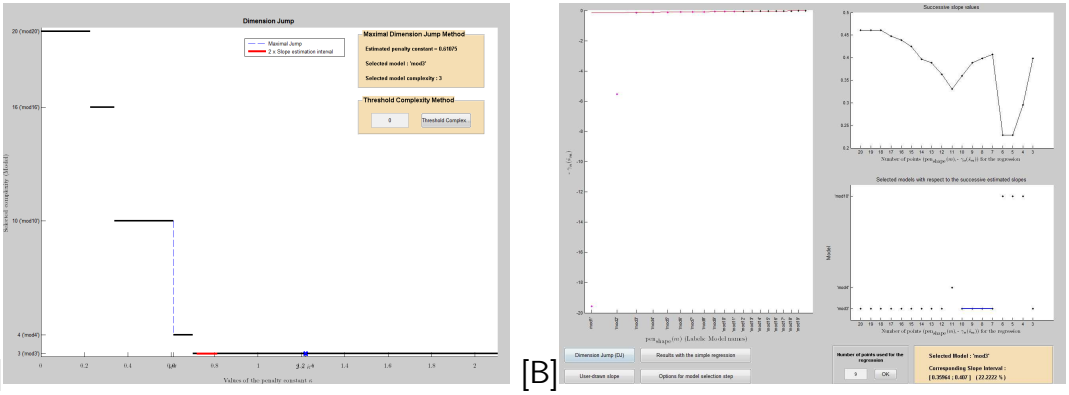


FIGURE E.3.: Output of CAPUSHE (n=90, d=2, k=3). [A] Dimension jump: Selected $\hat{k}$ versus penalty constant values. [B] Slope estimation: **Left**: Graph of the criterion $-W(\mu_n, \hat{\mathbf{c}}_k)$ as a function of $\sqrt{k/n}$. **Upper right**: Successive estimated slope values versus the number of points used for the slope estimation. **Bottom right**: Selected values of $\hat{k}$ versus the number of points used for the slope estimation.

G3 4 groups in dimension 3. Next, we used 4 groups of observations following

a normal distribution in dimension 3 with variance the identity matrix. These groups, depicted in Figure E.4, are centered at $(0,0,0), (3,5,-1), (-5,0,0)$, and $(6,6,6)$.

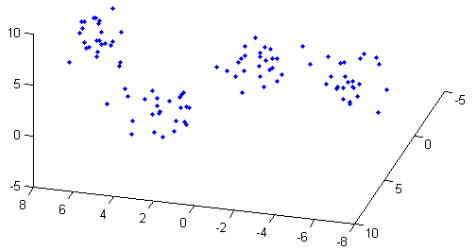


FIGURE E.4.: Groups of 25 observations following a normal distribution centered at $(0,0,0), (3,5,-1), (-5,0,0)$ and $(6,6,6)$.

G4 5 groups in dimension 4. This data set is made of 5 normal groups in dimension 4. The 5 groups are centered at $(0,0,0,0), (3,5,-1,0), (-5,0,0,0)$, $(1,1,6,-2)$ and $(1,-3,-2,5)$ respectively.

G5 4 groups in dimension 10. Finally, we have simulated, still using the normal distribution, 4 groups of data in dimension 10. For each of them, the 10 components of the mean vector were chosen uniformly at random between 0 and 10.

Table E.1 shows for the 5 simulated data sets the number $\hat{k}$ of clusters obtained with the slope estimation method by averaging over 20 repeated trials.

Data set	G1	G2	G3	G4	G5
Number of groups $\hat{k}$	1.05	3.2	4.1	5	4.05

TABLE E.1.: Number of clusters given by the algorithm based on the slope heuristics (average over 20 repeated trials).

E.3.1.2. More or less separate groups

In this subsection, two different configurations corresponding to 4 groups in dimension 3 are studied (see Figure E.5). In the first example (Figure E.5 [A]), the clusters are not as well separated as in the second (Figure E.5 [B]). We compared

the results of the algorithm based on the slope method and the *Gap Statistic* of Tibshirani et al. [37] in both situations.

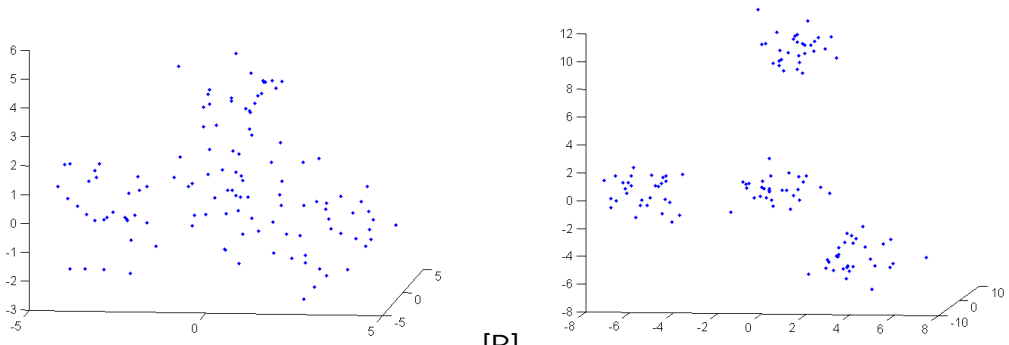


FIGURE E.5.: More or less separate groups: 4 groups of 30 observations following a normal distribution. [A] Groups centered at $(0,0,0), (0,2,3), (3,0,-1)$ and $(-3,-1,0)$. [B] Groups centered at $(0,0,0), (0,6,10), (3,0,-5)$ and $(-6,-3,0)$.

For the least separate groups, which are centered at $(0,0,0)$, $(0,2,3)$, $(3,0,-1)$ and $(-3,-1,0)$, the method selecting $\hat{k}$ by means of the slope heuristics yields $\hat{k}=4$ a little more than half of the time. The other values given by the algorithm are 3, 5, 6 and 7. The value 3 is obtained rarely, whereas the *Gap Statistic* founds $\hat{k}=3$ almost every time. For 10 realizations of such groups, Table E.2 shows, for both methods, the average value of $\hat{k}$ over 20 trials. The fact that the *Gap Statistic* does not perform very well here suggests that these clusters are two close from each other to estimate $\hat{k}$ accurately.

Slope method	4.25	5.2	5.2	4.6	4.6	4.5	4.55	4.55	4.6	4.15
Gap Statistic	3	4	3	3	3	3.1	3	3	3	3.2

TABLE E.2.: Number of clusters given by the algorithm based on the slope method and the *Gap Statistic*, for 10 realizations of Gaussian groups centered at $(0,0,0)$, $(0,2,3)$, $(3,0,-1)$ and $(-3,-1,0)$ (average over 20 repeated trials).

However, when the groups are well separated, the algorithm relying on the slope method seems to perform well, and the results are very similar to those obtained with the *Gap Statistic*: both methods almost always recover the expected result. For the well separate normal clusters, centered at $(0,0,0), (0,6,10), (3,0,-5)$ and

$(-6, -3, 0)$, visible in Figure E.5 [B], we obtain for example the CAPUSHE outputs shown in Figure E.6.

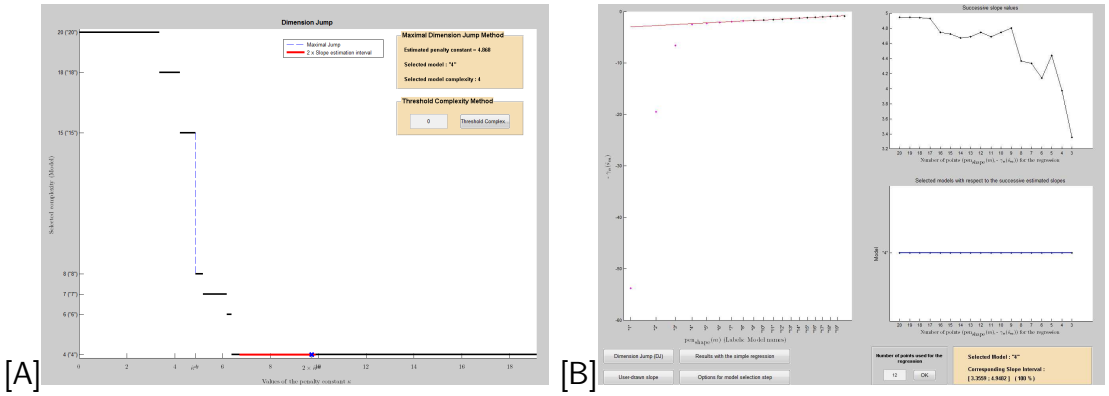


FIGURE E.6.: Output of CAPUSHE. [A] Dimension jump. [B] Slope estimation.

E.3.2. Real-life data

E.3.2.1. Zoo

The data, available from the UCI Machine Learning Repository [16], contains information about different animal species (such as wolf, herring, chicken, crab...). For each animal, 16 features have been registered: hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domestic, catsize (Booleans), number of legs (integer in $\{0, 2, 4, 5, 6, 8\}$). We consider a sample made of 92 data items building 5 underlining groups: mammal, fish, invertebrate, bird, insect. Most of the time, the output of the algorithm is $\hat{k} = 5$, the other values obtained being 4 and 6. The average number $\hat{k}$ over 20 trials was 5.05.

E.3.2.2. Dyslexia

Here, the data arises from a study about dyslexia, carried out at the Laboratoire de Sciences Cognitives et Psycholinguistique located in Paris in the Département d'Etudes Cognitives (DEC) of Ecole Normale Supérieure. In order to better understand this disability affecting a person's fluency or accuracy in reading, speaking, and spelling, several hypotheses have to be tested by comparing the performance of dyslexic and non-dyslexic adults (`http://www.ehess.fr/lscp/persons/ramus/fr/phonodysfr.html`). People aged from 18 to 31 took part to experiments based RAN tests (Rapid Automatized Naming, see, e.g., Denkla and Rudel [14]), which consist in naming rapidly digits, colors and objects, and on listening to a list of words and "non-words". Here, "non-word" means syllables put together that do

not build an existing word. For instance, “distu”, “malani” and “sonper” are non-words. For each of the 57 considered people, we have results such as response time, number of errors and response accuracy rate.

On this problem, the most often selected value of $\hat{k}$ was 4, and the average over 20 runs 3.9. The algorithm did not often choose $\hat{k} = 2$, which would correspond to the two classes, dyslexic people and control group. Nevertheless, the result can be explained by some “false positive” and “false negative” results in the study. Indeed, a few dyslexic individuals answered quite accurately and rapidly compared to the other dyslexic people, whereas some people not affected by dyslexia were slower and made more errors than expected.

E.3.2.3. Tasmanian Abalone

Abalone, also called ear-shell or sea ear, is a kind of sea snail. This marine gastropod mollusk in the family Haliotidae and the genus Haliotis presents several shell layers (“rings”), which can be used to learn its age, an important task to study the biology and ecology of a species. In fact, the number of “rings” plus 1.5 gives the age in years. More precisely, to determine the age of abalone, the shell is cut through the cone, stained, and the biologist counts the number of rings through a microscope. To avoid this time-consuming task, it can be of interest to predict the age of abalone from other physical measurements, which are easier to obtain. Here, we used a data set originating from the Marine Resources Division of the Marine Research Laboratories, Taroona, Department of Primary Industry and Fisheries, Tasmania (Nash, Sellers, Talbot, Cawthorn, and Ford [30]) and available from the UCI Machine Learning Repository [16]. The data contains information relative to 4177 abalones, labeled “female”, “male” or “infant”. For each of them, seven representative features have been measured: length (longest shell measurement), diameter (perpendicular to length), height, whole weight, shucked weight, viscera weight (after bleeding), shell weight (after being dried).

Considering a subsample of 1303 female abalones, with number of rings ranging from 5 to 23, we intend to recover the number of age groups from the physical measurements. The algorithm yields a number $\hat{k}$ of groups between 16 and 22 and the average value of $\hat{k}$ over 20 trials is 18.

E.4. Proof of Theorem E.2.1

Theorem E.2.1 is an adaptation of Theorem 8.1 in Massart [27]. For the proof, we will need the following lemma, which is a consequence of McDiarmid’s inequality [28] (see Massart [27, Theorem 5.3]).

Lemme E.4.1. *If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent random variables and $\mathcal{G}$ is a finite or countable class of real-valued functions such that $a \leq g \leq b$ for all $g \in \mathcal{G}$, then if $Z = \sup_{g\in\mathcal{G}} \sum_{i=1}^{n} (g(\mathbf{X}_i) - \mathbb{E}[g(\mathbf{X}_i)])$, we have, for every $\varepsilon \geq 0$,*

$$\mathbb{P}\left\{Z - \mathbb{E}[Z] \geq \varepsilon\right\} \leq \exp\left(-\frac{2\varepsilon^2}{n(b-a)^2}\right).$$

Proof of the theorem. Observe that, by the definition of $\tilde{\mathbf{c}}$,

$$W(\mu_n, \tilde{\mathbf{c}}) + \mathrm{pen}(\hat{k}) \leq W(\mu_n, \mathbf{c}_k) + \mathrm{pen}(k)$$

for all $k, 1 \leq k \leq n$ and $\mathbf{c}_k \in S_k$. Thus,

$$W(\mu_n, \tilde{\mathbf{c}}) - W(\mu_n, \mathbf{c}_k) \leq \mathrm{pen}(k) - \mathrm{pen}(\hat{k}),$$

which leads to

$$W(\mu, \tilde{\mathbf{c}}) \leq W(\mu_n, \mathbf{c}_k) + W(\mu, \tilde{\mathbf{c}}) - W(\mu_n, \tilde{\mathbf{c}}) + \mathrm{pen}(k) - \mathrm{pen}(\hat{k}). \tag{E.4}$$

Consider nonnegative weights $\{x_k\}_{1\leq k\leq n}$ such that

$$\sum_{k=1}^{n} e^{-x_k} = \Sigma,$$

and let $z > 0$. Applying Lemma E.4.1, we obtain, for all $k', 1 \leq k' \leq n$ and all $\varepsilon \geq 0$,

$$\mathbb{P}\left\{ \sup_{\mathbf{c}\in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c})) \geq \mathbb{E}\left[\sup_{\mathbf{c}\in S'_k} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c}))\right] + \varepsilon \right\} \leq \exp\left(-\frac{n\varepsilon^2}{8R^4}\right).$$

It follows that for every $k', 1 \leq k' \leq n$,

$$\mathbb{P}\left\{ \sup_{\mathbf{c}\in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c})) \geq \mathbb{E}\left[\sup_{\mathbf{c}\in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c}))\right] + 4R^2\sqrt{\frac{x_{k'} + z}{2n}} \right\} \leq e^{-x_{k'} - z}.$$

Setting $E_{k'} = \mathbb{E}\left[\sup_{\mathbf{c}\in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c}))\right]$, we have, for all $k', 1 \leq k' \leq n$,

$$\sup_{\mathbf{c}\in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c})) \leq E_{k'} + 4R^2\sqrt{\frac{x_{k'} + z}{2n}},$$

except on a set with probability not larger than Σe^{-z}. By inequality (E.4), we thus get

$$\begin{aligned} W(\mu, \tilde{\mathbf{c}}) &\leq W(\mu_n, \mathbf{c}_k) + E_{\hat{k}} + 4R^2\sqrt{\frac{x_{\hat{k}} + z}{2n}} - \mathrm{pen}(\hat{k}) + \mathrm{pen}(k) \\ &\leq W(\mu_n, \mathbf{c}_k) + E_{\hat{k}} + 4R^2\sqrt{\frac{x_{\hat{k}}}{2n}} - \mathrm{pen}(\hat{k}) + \mathrm{pen}(k) + 4R^2\sqrt{\frac{z}{2n}}, \end{aligned}$$

except on a set of probability not larger than Σe^{-z}. Next, according to Linder [26, Theorem 3], there exists a constant $A(d, R) > 0$ such that

$$E_{k'} \leq A(d, R)\sqrt{\frac{k'}{n}}.$$

Thus, if for all $k', 1 \leq k' \leq n$,

$$\mathrm{pen}(k') \geq A(d, R)\sqrt{\frac{k'}{n}} + 4R^2\sqrt{\frac{x_{k'}}{2n}},$$

then

$$W(\mu, \tilde{\mathbf{c}}) \leq W(\mu_n, \mathbf{c}_k) + \mathrm{pen}(k) + 4R^2\sqrt{\frac{z}{2n}},$$

except on a set of probability not larger than Σe^{-z}. This may be rewritten

$$\mathbb{P}\left\{(4R^2)^{-1}\sqrt{2n}[W(\mu, \tilde{\mathbf{c}}) - W(\mu_n, \mathbf{c}_k) + \mathrm{pen}(k)] \geq \sqrt{z}\right\} \leq \Sigma e^{-z},$$

or, setting $z = u^2$,

$$\mathbb{P}\left\{(4R^2)^{-1}\sqrt{2n}[W(\mu, \tilde{\mathbf{c}}) - W(\mu_n, \mathbf{c}_k) + \mathrm{pen}(k)] \geq u\right\} \leq \Sigma e^{-u^2}.$$

Recalling that $\int_0^{+\infty} e^{-u^2}\mathrm{d}u = \frac{\sqrt{\pi}}{2}$ and letting $g_+ = \max(g, 0)$, we get

$$\mathbb{E}\left[(W(\mu, \tilde{\mathbf{c}}) - W(\mu_n, \mathbf{c}_k) + \mathrm{pen}(k))_+\right] \leq R^2\Sigma\sqrt{\frac{2\pi}{n}}.$$

Since $\mathbb{E}[W(\mu_n, \mathbf{c}_k)] = W(\mu, \mathbf{c}_k)$,

$$\mathbb{E}\left[W(\mu, \tilde{\mathbf{c}})\right] \leq W(\mu, \mathbf{c}_k) + \mathrm{pen}(k) + R^2\Sigma\sqrt{\frac{2\pi}{n}}.$$

As this is true for every k,

$$\mathbb{E}\left[W(\mu, \tilde{\mathbf{c}})\right] \leq \inf_{1\leq k\leq n}(W(\mu, S_k) + \mathrm{pen}(k)) + R^2\Sigma\sqrt{\frac{2\pi}{n}},$$

where $W(\mu, S_k) = \inf_{\mathbf{c}\in S_k} W(\mu, \mathbf{c})$. This completes the proof of the theorem.

References

[1] B. Al-Shboul and S.-H. Myaeng. Initializing k-means using genetic algorithms. *World Academy of Science, Engineering and Technology*, 54:114–118, 2009.

[2] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279, 2009.

[3] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

[4] P. L. Bartlett, T. Linder, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, 44:1802–1813, 1998.

[5] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 2011. In press. Available at `http://hal.archives-ouvertes.fr/docs/00/46/16/39/PDF/RR-7223.pdf`.

[6] S. Ben-David, D. Pál, and H. U. Simon. Stability of k-means clustering. In N. Bshouty and C. Gentile, editors, *Proceedings of the 20th Annual Conference on Learning Theory (COLT 2007)*, pages 20–34. Springer, 2007.

[7] S. Ben-David and U. von Luxburg. Relating clustering stability to properties of cluster boundaries. In R. A. Servedio and T. Zhang, editors, *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pages 379–390, Madison, 2008. Omnipress.

[8] S. Ben-David, U. von Luxburg, and D. Pál. A sober look on clustering stability. In G. Lugosi and H. U. Simon, editors, *Proceedings of the 19th Annual Conference on Learning Theory (COLT 2006)*, pages 5–19, Berlin, 2006. Springer.

[9] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In *Proceedings of the 7th Pacific Symposium on Biocomputing*, volume 7, pages 6–17, 2002.

[10] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54:781–790, 2008.

[11] L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 55–87. Springer, New York, 1997.

[12] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138:33–73, 2007.

[13] R. B. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3:1–27, 1974.

[14] M. B. Denckla and R. G. Rudel. Rapid "automatized" naming (R.A.N.): dyslexia differentiated from other learning disabilities. *Neuropsychologia*, 14:471–479, 1976.

[15] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, New York, 2000.

[16] A. Frank and A. Asuncion. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2010. `http://archive.ics.uci.edu/ml`.

[17] A. D. Gordon. *Classification*, volume 82 of *Monographs on Statistics and Applied Probability*. Chapman Hall/CRC, Boca Raton, 1999.

[18] A. Hardy. On the number of clusters. *Computational Statistics and Data Analysis*, 23:83–96, 1996.

[19] J. A. Hartigan. *Clustering Algorithms*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, 1975.

[20] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, 2001.

[21] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Mathematical Statistics. Wiley-Interscience, Hoboken, 1990.

[22] S. S. Khan and A. Ahmad. Cluster center initialization algorithm for k-means clustering. *Pattern Recognition Letters*, 25:1293–1302, 2004.

[23] D. J. Kim, Y. W. Park, and D. J. Park. A novel validity index for determination of the optimal number of clusters. *IEICE Transactions on Information and System*, E84D:281–285, 2001.

[24] W. J. Krzanowski and Y. T. Lai. A criterion for determining the number of clusters in a data set. *Biometrics*, 44:23–34, 1985.

[25] E. Levine and E. Domany. Resampling method for unsupervised estimation of cluster validity. *Journal of Neural Computation*, 13:2573–2593, 2002.

[26] T. Linder. On the training distortion of vector quantizers. *IEEE Transactions on Information Theory*, 46:1617–1623, 2000.

[27] P. Massart. *Concentration Inequalities and Model Selection*. Ecole d'Eté de Probabilités de Saint-Flour XXXIII – 2003, Lecture Notes in Mathematics. Springer, Berlin, Heidelberg, 2007.

[28] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press, Cambridge, 1989.

[29] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–79, 1985.

[30] W. J. Nash, T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B Ford. The population biology of Abalone (Haliotis species) in Tasmania. 1, Blacklip Abalone (H. rubra) from the north coast and islands of Bass Strait. Technical Report 48, Sea Fisheries Division, 1994.

[31] J. M. Pena, J. A. Lozano, and P. Larranaga. An empirical comparison of four initialization methods for the K-means algorithm. *Pattern Recognition Letters*, 20:1027–1040, 1999.

[32] G. T. Perim, E. D. Wandekokem, and F. M. Varejão. K-means initialization methods for improving clustering by simulated annealing. In *Advances in artificial intelligence – Iberamia 2008*, volume 5290, pages 133–142. Springer-Verlag, Berlin, Heidelberg, 2008.

[33] O. Shamir and N. Tishby. Cluster stability for finite samples. In J.C. Platt, D. Koller, Y. Singer, and S. Rowseis, editors, *Advances in Neural Information Processing Systems 20*, pages 1297–1304, Cambridge, 2008. MIT Press.

[34] O. Shamir and N. Tishby. Model selection and stability in k-means clustering. In R. A. Servedio and T. Zhang, editors, *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pages 367–378, Madison, 2008. Omnipress.

[35] T. Su and J. Dy. A deterministic method for initializing k-means clustering. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, volume 141, pages 784–786, 2004.

[36] C. A. Sugar and G. M. James. Finding the number of clusters in a data set: an information theoretic approach. *Journal of the American Statistical Association*, 98:750–763, 2003.

[37] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society*, 63:411–423, 2001.

F. Parameter selection for principal curves*

Abstract

Principal curves are nonlinear generalizations of the notion of first principal component. Roughly, a principal curve is a parameterized curve in $\mathbb{R}^d$ which passes through the "middle" of a data cloud drawn from some unknown probability distribution. Depending on the definition, a principal curve relies on some unknown parameters (number of segments, length, turn...) which have to be properly chosen to recover the shape of the data without interpolating. In the present paper, we consider the principal curve problem from an empirical risk minimization perspective and address the parameter selection issue using the point of view of model selection via penalization. We offer oracle inequalities and implement the proposed approaches to recover the hidden structures in both simulated and real-life data.

F.1. Introduction

F.1.1. Principal curves

Statisticians use various methods in order to sum up information and represent the data by simpler quantities. Among these methods, Principal Component Analysis (PCA) aims at determining the maximal variance axes of a data cloud, as a means to represent the observations in a compact manner revealing as well as possible their variability (see, e.g., Mardia, Kent and Bibby [32]). This technique, initiated at the beginning of the last century by Pearson [35] and Spearman [38], and further developed by Hotelling [26], is certainly one of the most famous and most widely used procedure of multivariate analysis. Whether in the context of dimension reduction or feature extraction, PCA often provides a first important insight in the data structure.

*Article écrit en collaboration avec Gérard Biau. Cette annexe reprend la Section 2.2 du Chapitre 2 de la seconde partie.

However, in a number of situations, it may be of interest to summarize information in a nonlinear manner instead of representing the data by straight lines. This approach leads to the notion of principal curve, which can be thought of as a nonlinear generalization of the first principal component. Roughly, the purpose is to search for a curve passing through the middle of the observations, as illustrated in Figure F.1. Principal curves have a broad range of applications in many different areas, such as physics (Hastie and Stuetzle [25], Friedsam and Oren [22]), character and speech recognition (Kégl and Krzyżak [28], Reinhard and Niranjan [36]), mapping and geology (Brunsdon [10], Stanford and Raftery [39], Banfield and Raftery [4], Einbeck, Tutz and Evers [19, 20]), natural sciences (De'ath [14], Corkeron, Anthony and Martin [13], Einbeck, Tutz and Evers [19]) and medicine (Wong and Chung [42], Caffo, Crainiceanu, Deng and Hendrix [11]).

The definition of a principal curve typically depends of the principal component property one wants to generalize. Most of the time, this definition is first stated for an $\mathbb{R}^d$-valued random variable $\mathbf{X} = (X_1, \ldots, X_d)$ with known distribution, and then adapted to the practical situation where one observes independent draws $\mathbf{X}_1, \ldots, \mathbf{X}_n$ distributed as $\mathbf{X}$.

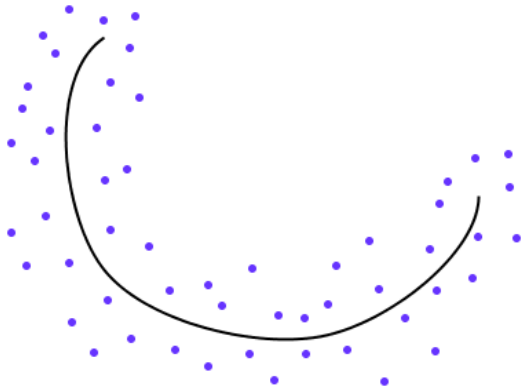


FIGURE F.1.: An example of principal curve.

The original definition of a principal curve goes back to Hastie and Stuetzle [25] and relies on the self-consistency property of principal components. In words, a smooth (infinitely differentiable) parameterized curve $\mathbf{f}(t) = (f_1(t), \ldots, f_d(t))$ is a principal curve for $\mathbf{X}$ if $\mathbf{f}$ does not intersect itself, if it has finite length inside any bounded subset of $\mathbb{R}^d$, and if it is self-consistent. This last requirement means that

$$\mathbf{f}(t) = \mathbb{E}[\mathbf{X} | t_{\mathbf{f}}(\mathbf{X}) = t], \tag{F.1}$$

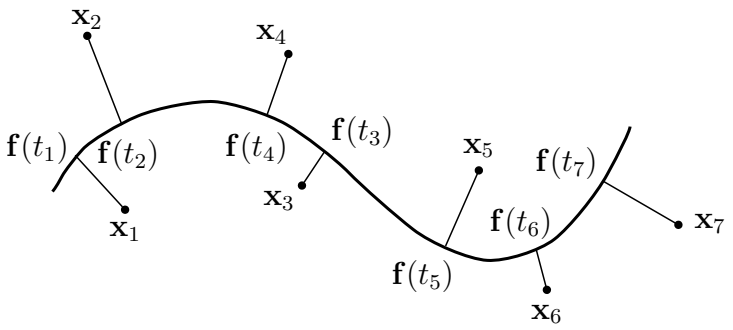


FIGURE F.2.: The projection index $t_{\mathbf{f}}$. For all i, t_i stands for $t_{\mathbf{f}}(\mathbf{x}_i)$.

where the so-called projection index $t_{\mathbf{f}}(\mathbf{x})$ is the largest real number t minimizing the squared Euclidean distance between $\mathbf{x}$ and $\mathbf{f}(t)$, as depicted in Figure F.2. More formally,

$$t_{\mathbf{f}}(\mathbf{x}) = \sup\Big\{t \,:\, \|\mathbf{x} - \mathbf{f}(t)\| = \inf_{t'} \|\mathbf{x} - \mathbf{f}(t')\|\Big\}.$$

The self-consistency property may be interpreted by saying that each point of the curve $\mathbf{f}$ is the mean of the observations projecting on $\mathbf{f}$ around this point. Hastie and Stuetzle discuss in [25] an iterative algorithm, alternating between a projection and a conditional expectation step, which yields an approximate principal curve. As this approach exhibits different types of bias, Banfield and Raftery [4] and Chang and Ghosh [12] propose a modification of the algorithm, whereas Tibshirani, tackling the model bias problem, adopts in [40] a semiparametric strategy and defines principal curves in terms of a mixture model. For more references on principal curves and related points of view, we refer the reader to Verbeek, Vlassis and Kröse [41] (k-segments algorithm), Delicado [15] (principal curves of oriented points), Einbeck, Tutz and Evers [20] (local principal curves) and Genovese, Perone-Pacifico, Verdinelli and Wasserman [23], who recently discussed a closely related approach, called nonparametric filament estimation.

In the present paper, we will adopt the principal curve definition of Kégl, Krzyżak, Linder and Zeger [29], which is slightly different from the original one. The main advantage of this definition, which is recalled in the next paragraph, is that it avoids the implicit conditional expectation requirement (F.1) and, consequently, turns out to be more easily amenable to mathematical analysis.

F.1.2. Constrained principal curves

In the definition of Kégl, Krzyżak, Linder and Zeger [29] (**KKLZ** hereafter), a principal curve of length (at most) L for $\mathbf{X}$ is a parameterized curve minimizing the least-square criterion

$$\Delta(\mathbf{f}) = \mathbb{E}\left[\inf_t \|\mathbf{X} - \mathbf{f}(t)\|^2\right]$$

over a collection $\mathcal{F}_L$ of curves of length not larger than some prespecified positive L. We note that, in this context, a principal curve always exists provided $\mathbb{E}\|\mathbf{X}\|^2 < \infty$, but that it may not necessarily be unique. In practice, as the distribution of $\mathbf{X}$ is unknown, $\Delta(\mathbf{f})$ is replaced by its empirical counterpart

$$\Delta_n(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n} \inf_t \|\mathbf{X}_i - \mathbf{f}(t)\|^2$$

based on a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of independent random variables distributed as $\mathbf{X}$. Considering the minimum $\hat{\mathbf{f}}_{k,n}$ of $\Delta_n(\mathbf{f})$ over the subclass $\mathcal{F}_{k,L} \subset \mathcal{F}_L$ of all polygonal lines $\mathbf{f}_{k,n}$ with k segments and length not larger than L, Kégl, Krzyżak, Linder and Zeger [29] prove that, whenever $\mathbf{X}$ is almost surely bounded, and for the choice $k \propto n^{1/3}$,

$$\Delta(\hat{\mathbf{f}}_{k,n}) - \min_{\mathbf{f}\in\mathcal{F}_L} \Delta(\mathbf{f}) = \mathcal{O}(n^{-1/3}).$$

As the task of finding a polygonal line with k segments and length at most L minimizing $\Delta_n(\mathbf{f})$ is computationally difficult, **KKLZ** propose an approximate iterative algorithm that they call the Polygonal Line Algorithm. This algorithm is initialized using the smallest segment included in the first principal component containing all projected data points. Then, at each step, a vertex—and thus, a segment—is added to the current polygonal line, and the vertices are updated in a cyclic manner during an inner loop alternating between a projection and an optimization step. Performing the projection step is similar to constructing a Voronoi partition, with respect to both the vertices and segments. To optimize a vertex, a local version of $\Delta_n(\mathbf{f})$ is used, involving only the data projecting to this vertex and to the adjacent segments. The criterion is penalized to avoid sharp angles, which in turn amounts to penalizing the length of the curve.

Working out the angle penalty in the Polygonal Line Algorithm, Sandilya and Kulkarni (**SK** hereafter) propose in [37] a closely related definition, by imposing a constraint on the turn (Alexandrov and Reshetnyak [2]) of the curve $\mathbf{f}$. This approach consists in replacing the class $\mathcal{F}_L$ by $\mathcal{F}_K$, where K stands for the maximal turn. Thus, denoting by $\mathcal{F}_{k,K} \subset \mathcal{F}_K$ the subclass of all polygonal lines $\mathbf{f}_{k,n}$ with k

segments and turn not larger than K, **SK** prove that, whenever $\mathbf{X}$ is almost surely bounded, and for the choice $k \propto n^{1/3}$,

$$\Delta(\hat{\mathbf{f}}_{k,n}) - \min_{\mathbf{f} \in \mathcal{F}_K} \Delta(\mathbf{f}) = \mathcal{O}(n^{-1/3}).$$

Whether in the **KKLZ** definition or in the **SK** one, selecting the various smoothness parameters (the number k of segments, the curve length ℓ, the turn κ) is an essential issue, as illustrated in Figure F.3. A good choice of these parameters is critical, since a principal curve obtained with a poor class will be too rough, whereas a class containing too many curves may lead to severe interpolation problems. In practice, the Polygonal Line Algorithm stops when k is larger than a certain threshold, chosen heuristically and tuned after carrying out several experiments. The stopping condition involves the number n of observations and the actual value of the criterion Δ_n. However, to our knowledge, this empirical procedure is not supported by any theoretical argument and leads to variable results, depending on the data set.

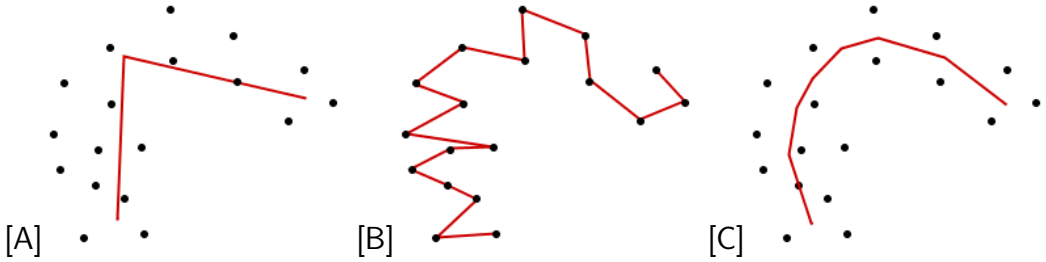


FIGURE F.3.: Principal curves fitted with [A] a too small number k of segments, [B] a too large k and [C] an appropriate one.

As far as we know, the issue of an automatic (i.e., data-dependent) choice of the parameters k, ℓ and κ has not been addressed in the literature. Thus, to fill the gap, we propose in the present contribution to focus on this question both from a theoretical and practical point of view. Our approach will strongly rely on the model selection theory by penalization introduced by Birgé and Massart [8] and Barron, Birgé and Massart [5], as well as on a recent penalty calibration approach proposed by Birgé and Massart [9] and Arlot and Massart [3].

The paper is organized as follows. First, we consider in Section F.2 principal curves with bounded length and show that the polygonal line obtained by minimizing some appropriate penalized criterion satisfies an oracle-type inequality. Section F.3 provides a similar result in the context of principal curves with bounded turn. Our theoretical findings are illustrated on both simulated and real data sets in Section F.4. For the sake of clarity, proofs are collected in Section F.5.

F.2. Principal curves with bounded length

Let $\|\cdot\|$ be the standard Euclidean norm over $\mathbb{R}^d$. A parameterized curve in $\mathbb{R}^d$ is a continuous function

$$\begin{aligned} \mathbf{f} : I &\to \mathbb{R}^d \\ t &\mapsto (f_1, \ldots, f_d), \end{aligned}$$

where $I = [a, b]$ is a closed interval of the real line. The length of $\mathbf{f}$ is defined by

$$\mathscr{L}(\mathbf{f}) = \sup \sum_{j=1}^{m} \|\mathbf{f}(t_j) - \mathbf{f}(t_{j-1})\|,$$

where the supremum is taken over all subdivisions $a = t_0 < t_1 < \cdots < t_m = b$, $m \geq 1$ (see, e.g., Kolmogorov and Fomin [30]). Throughout the document, it is assumed that $\mathbb{E}\|\mathbf{X}\|^2 < \infty$ and that

$$\mathbb{P}\{\mathbf{X} \in \mathcal{C}\} = 1, \tag{F.2}$$

where $\mathcal{C}$ is a convex compact subset of $\mathbb{R}^d$, with diameter δ. By Lemma 1 in Kégl [27], the requirement (F.2) implies that, for any given positive length L, there exists a principal curve for $\mathbf{X}$ with length at most L in $\mathcal{C}$, that is a (non necessarily unique) parameterized curve $\mathbf{f}^\star$ with length not larger than L and support in $\mathcal{C}$ achieving the minimum of $\mathbb{E}\left[\inf_{t\in I} \|\mathbf{X} - \mathbf{f}(t)\|^2\right]$. Consequently, in the sequel, we will restrict ourselves to curves whose support is included in $\mathcal{C}$ and denote by $\mathcal{F}$ the set of all parameterized curves $\mathbf{f} = (f_1, \ldots, f_d)$ belonging to $\mathcal{C}$.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sample of independent random variables distributed as $\mathbf{X}$, and consider the contrast

$$\Delta(\mathbf{f}, \mathbf{x}) = \inf_{t\in I} \|\mathbf{x} - \mathbf{f}(t)\|^2, \quad \mathbf{f} \in \mathcal{F}, \mathbf{x} \in \mathbb{R}^d.$$

The associated empirical risk based on the sample $\mathbf{X}_1, \ldots \mathbf{X}_n$ is defined as

$$\Delta_n(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n} \Delta(\mathbf{f}, \mathbf{X}_i) = \frac{1}{n}\sum_{i=1}^{n} \inf_{t\in I} \|\mathbf{X}_i - \mathbf{f}(t)\|^2.$$

For some prespecified length $L > 0$, we set

$$\mathbf{f}^\star \in \operatorname*{arg\,min}_{\mathbf{f}\in\mathcal{F}, \mathscr{L}(\mathbf{f})\leq L} \mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})].$$

Next, let $\mathcal{L}$ be a countable subset of $]0, L]$ and $\mathcal{Q}$ a grid over $\mathcal{C}$. For every $k \geq 1$ and $\ell \in \mathcal{L}$, the model $\mathcal{F}_{k,\ell}$ is defined as the collection of all polygonal lines with k

segments, with length at most ℓ, and with vertices belonging to $\mathcal{Q}$. We note that each model $\mathcal{F}_{k,\ell}$ as well as the family of models $\{\mathcal{F}_{k,\ell}\}_{k\geq 1,\ell\in\mathcal{L}}$ are countable. For $k \geq 1$ and $\ell \in \mathcal{L}$, let

$$\hat{\mathbf{f}}_{k,\ell} \in \operatorname*{arg\,min}_{\mathbf{f}\in\mathcal{F}_{k,\ell}} \Delta_n(\mathbf{f})$$

be a curve achieving the minimum of the empirical criterion $\Delta_n(\mathbf{f})$ over the polygonal line class $\mathcal{F}_{k,\ell}$.

At this stage of the procedure, we have at hand a family of estimates $\{\hat{\mathbf{f}}_{k,\ell}\}_{k\geq 1,\ell\in\mathcal{L}}$ and our goal is to select the best principal curve $\tilde{\mathbf{f}}$ among this collection. To this aim, we make use of the model selection approach of Barron, Birgé and Massart [5], which allows to assess the adjustment quality by controlling the loss

$$\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) = \mathbb{E}[\Delta(\tilde{\mathbf{f}}, \mathbf{X}) - \Delta(\mathbf{f}^\star, \mathbf{X})]$$

between the target $\mathbf{f}^\star$ and the selected curve $\tilde{\mathbf{f}}$. (For a comprehensive introduction to the area of model selection, the reader is referred to the monograph of Massart [33].) More formally, let $\mathrm{pen} : \mathbb{N}^* \times \mathcal{L} \to \mathbb{R}^+$ be some penalty function and denote by $(\hat{k}, \hat{\ell})$ a pair of minimizers of the criterion

$$\mathrm{crit}(k, \ell) = \Delta_n(\hat{\mathbf{f}}_{k,\ell}) + \mathrm{pen}(k, \ell).$$

In order to obtain the desired principal curve $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k},\hat{\ell}}$, we have to design an adequate penalty $\mathrm{pen}(k, \ell)$. This is done in the following theorem, which is an adaptation of a general model selection result of Massart [33, Theorem 8.1]. However, for the sake of completeness, it is proved in its full length in Section F.5.

Theorem F.2.1. *Consider a family of nonnegative weights* $\{x_{k,\ell}\}_{k\geq 1,\ell\in\mathcal{L}}$ *such that*

$$\sum_{k\geq 1,\ell\in\mathcal{L}} e^{-x_{k,\ell}} = \Sigma < \infty,$$

and a penalty function $\mathrm{pen} : \mathbb{N}^* \times \mathcal{L} \to \mathbb{R}^+$. *Let* $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k},\hat{\ell}}$. *If for all* $(k, \ell) \in \mathbb{N}^* \times \mathcal{L}$,

$$\mathrm{pen}(k, \ell) \geq \mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}} \Big(\mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})] - \Delta_n(\mathbf{f})\Big)\right] + \delta^2\sqrt{\frac{x_{k,\ell}}{2n}},$$

then

$$\mathbb{E}[\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}})] \leq \inf_{k\geq 1,\ell\in\mathcal{L}} \Big[\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\ell}) + \mathrm{pen}(k, \ell)\Big] + \frac{\delta^2\Sigma}{2^{3/2}}\sqrt{\frac{\pi}{n}},$$

where $\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\ell}) = \inf_{\mathbf{f}\in\mathcal{F}_{k,\ell}} \mathcal{D}(\mathbf{f}^\star, \mathbf{f})$.

Theorem F.2.1 offers a nonasymptotic bound, expressing the fact that the expected loss of the final estimate $\tilde{\mathbf{f}}$ is close to the minimal loss over all $k \geq 1$ and $\ell \in \mathcal{L}$, up to a term tending to 0. Thus, in order to apply this theorem to the principal curve problem, we now have to find an upper bound on the quantity

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\Big(\mathbb{E}[\Delta(\mathbf{f},\mathbf{X})]-\Delta_n(\mathbf{f})\Big)\right]. \tag{F.3}$$

This is achieved by Proposition F.2.1 below, which is proved by showing that the expected maximal deviation (F.3) may be bounded by a Rademacher average (see Bartlett, Boucheron, and Lugosi [6] and Koltchinskii [31]) and by resorting to a Dudley integral (Dudley [17]).

Proposition F.2.1. *Let $\mathcal{F}_{k,\ell}$ be the set of all polygonal lines with k segments, length at most ℓ, and vertices in a grid $\mathcal{Q} \subset \mathcal{C}$. Then there exist nonnegative constants $a_0, \ldots, a_3$, depending on the maximal length L, the dimension d and the diameter δ of the convex set $\mathcal{C}$, such that*

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\Big(\mathbb{E}[\Delta(\mathbf{f},\mathbf{X})]-\Delta_n(\mathbf{f})\Big)\right] \leq \frac{1}{\sqrt{n}}\left[a_1\sqrt{k}+a_2\sqrt{\ell}+a_3\frac{\ell}{\sqrt{k}}+a_0\right].$$

Finally, by combining Theorem F.2.1 and Proposition F.2.1, we are in a position to state the main result of this section.

Theorem F.2.2. *Consider a family of nonnegative weights $\{x_{k,\ell}\}_{k\geq 1,\ell\in\mathcal{L}}$ such that*

$$\sum_{k\geq 1,\ell\in\mathcal{L}} e^{-x_{k,\ell}} = \Sigma < \infty,$$

and a penalty function $\mathrm{pen} : \mathbb{N}^* \times \mathcal{L} \to \mathbb{R}^+$. *Let $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k},\hat{\ell}}$. There exist nonnegative constants $c_0, \ldots, c_3$, depending on the maximal length L, the dimension d and the diameter δ of the convex set $\mathcal{C}$, such that, if for all $(k,\ell) \in \mathbb{N}^* \times \mathcal{L}$,*

$$\mathrm{pen}(k,\ell) \geq \frac{1}{\sqrt{n}}\left[c_1\sqrt{k}+c_2\sqrt{\ell}+c_3\frac{\ell}{\sqrt{k}}+c_0\right]+\delta^2\sqrt{\frac{x_{k,\ell}}{2n}},$$

then

$$\mathbb{E}[\mathcal{D}(\mathbf{f}^\star,\tilde{\mathbf{f}})] \leq \inf_{k\geq 1,\ell\in\mathcal{L}}\left[\mathcal{D}(\mathbf{f}^\star,\mathcal{F}_{k,\ell})+\mathrm{pen}(k,\ell)\right]+\frac{\delta^2\Sigma}{2^{3/2}}\sqrt{\frac{\pi}{n}},$$

where $\mathcal{D}(\mathbf{f}^\star,\mathcal{F}_{k,\ell}) = \inf_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\mathcal{D}(\mathbf{f}^\star,\mathbf{f})$.

Some comments are in order.

Firstly, we see that the penalty shape involves a term proportional to $\sqrt{k/n}$ and a term proportional to $\sqrt{\ell/n}$, as well as the quantity $\ell/\sqrt{kn}$. This penalty form, which vanishes at the rate $1/\sqrt{n}$, seems relevant insofar as the number k of segments and the length ℓ of the curves measure the complexity of the models. It is also noteworthy that the term $\ell/\sqrt{kn}$ is natural in some sense, since it reflects the mutual dependence between k and ℓ. In other words, whenever the length ℓ increases, more segments should be allowed in order to keep a certain degree of smoothness.

Observe next that the proof of Proposition F.2.1 provides possible values for the constants $c_0, \ldots, c_3$. However, these values are not very helpful since they are upper bounds which are probably far from being tight. Nevertheless, the proof also reveals that $c_1 = c_1'\delta^2$, $c_2 = c_2'\delta\sqrt{\delta}$, $c_3 = c_3'\delta$ and $c_0 = c_0'\delta^2$, where c_0', c_1', c_2' and c_3' are constants without dimension, so that the penalty is in fact homogeneous to a squared length, just like the criterion $\Delta_n(\mathbf{f})$ is.

Finally, an important practical issue is how to choose the weights $\{x_{k,\ell}\}_{k\geq 1, \ell\in\mathcal{L}}$. These weights should be large enough to ensure the finiteness of Σ, but not too large at the risk of overpenalizing. If the cardinality of the collection of models is not larger than n^2 (this will be the case in all our practical examples), we may set $x_{k,\ell} = 2\ln n$ for every (k,ℓ). This choice does not affect the penalty shape, though modifying the rate, and leads to $\Sigma = 1$ in the risk bound.

Remark F.2.1. Clearly, when the length ℓ of polygonal lines is fixed, and the aim is to select the number k of segments, the dominant term reflecting the complexity of the models in the penalty is $\sqrt{k/n}$. In this particular context, the weights may be taken equal to $\ln n$, or, by analogy with the Gaussian linear model selection framework, proportional to k. Indeed, in this framework, each model S_m, $m \in \mathcal{M}$, has dimension D_m and an interesting choice for x_m is then $x_m = x(D_m)$, where $x(D) = cD + \ln|\{m \in \mathcal{M} : D_m = D\}|$ and $c > 0$. When there is no redundancy in the models dimension, this strategy amounts to choosing x_m proportional to D_m. In our problem, this means setting $x_k = ck$ for every k, where the constant $c > 0$ ensures that $\sum_{k\geq 1} e^{-ck} = \Sigma < \infty$. Thus, in this somewhat restrictive situation, the penalty is of the order $\sqrt{k/n}$.

F.3. Principal curves with bounded turn

As it was already mentioned in the Introduction, Sandilya and Kulkarni [37] (**SK**) suggest an alternative approach for principal curves, based on the control of the

turn. Recall that the turn $\mathscr{K}(\mathbf{f})$ of a curve $\mathbf{f} : I \to \mathbb{R}^d$, $I = [a, b]$, is given by

$$\mathscr{K}(\mathbf{f}) = \sup \sum_{j=1}^{m-1} \widehat{f(t_j)},$$

where $\widehat{f(t_j)}$ denotes the angle between the vectors $\overrightarrow{f(t_{j-1})f(t_j)}$ and $\overrightarrow{f(t_j)f(t_{j+1})}$, and the supremum is taken over all subdivisions $a = t_0 < t_1 < \cdots < t_m = b$, $m \geq 1$ (Alexandrov and Reshetnyak [2]). Thus, the turn of a polygonal line with vertices $v_1, \ldots, v_{k+1}$ is just the sum of the angles at $v_2, \ldots, v_k$ (see Figure F.4 for an illustration).

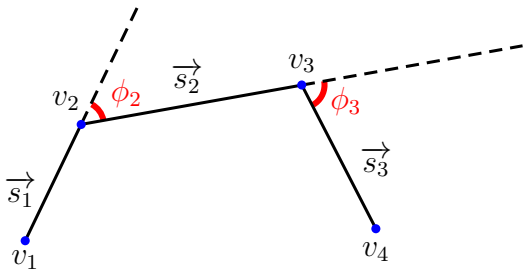


FIGURE F.4.: Denoting by $\overrightarrow{s_j}$ the vector $\overrightarrow{v_j v_{j+1}}$ for all $j = 1, \ldots, k$, the angles involved in the definition of the turn are defined by $\phi_{j+1} = (\overrightarrow{s_j}, \overrightarrow{s_{j+1}})$.

As a logical continuation to Section F.2, we propose in the present section to analyse the **SK** definition from a model selection point of view. To this aim, we use the fact that a curve with bounded turn also has bounded length, as shown in Lemma F.3.1 below.

We still assume that $\mathbb{P}\{\mathbf{X} \in \mathcal{C}\} = 1$, where $\mathcal{C}$ is a convex compact subset of $\mathbb{R}^d$ with diameter δ. By Proposition 1 in **SK**, this requirement ensures the existence of a curve $\mathbf{f}^\star$ with bounded turn minimizing the criterion $\mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})]$. More formally, for some prespecified turn $K \geq 0$, we set

$$\mathbf{f}^\star \in \underset{\mathbf{f} \in \mathcal{F}, \mathscr{K}(\mathbf{f}) \leq K}{\arg\min} \mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})],$$

where $\mathscr{K}(\mathbf{f})$ denotes the turn of $\mathbf{f}$. Proceeding as in Section F.2, we let $\mathcal{K}$ be a countable subset of $[0, K]$ and define a countable collection of models $\{\mathcal{F}_{k,\kappa}\}_{k \geq 1, \kappa \in \mathcal{K}}$ as follows. Each $\mathcal{F}_{k,\kappa}$ consists of polygonal lines with k segments, with turn at most κ, and with vertices belonging to some grid $\mathcal{Q}$ over $\mathcal{C}$. For $k \geq 1$ and $\kappa \in \mathcal{K}$, define

$$\hat{\mathbf{f}}_{k,\kappa} \in \underset{\mathbf{f} \in \mathcal{F}_{k,\kappa}}{\arg\min} \Delta_n(\mathbf{f})$$

to be a polygonal line minimizing the empirical criterion $\Delta_n(\mathbf{f})$ over $\mathcal{F}_{k,\kappa}$. We wish to design an appropriate penalty function pen : $\mathbb{N}^* \times \mathcal{K} \to \mathbb{R}^+$ and minimize the criterion

$$\mathrm{crit}(k,\kappa) = \Delta_n(\hat{\mathbf{f}}_{k,\kappa}) + \mathrm{pen}(k,\kappa)$$

in order to obtain a suitable principal curve. As before, we let $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k},\hat{\kappa}}$, where $(\hat{k},\hat{\kappa})$ is a minimizer of the penalized criterion $\mathrm{crit}(k,\kappa)$, and intend to control the loss $\mathcal{D}(\mathbf{f}^\star,\tilde{\mathbf{f}}) = \mathbb{E}[\Delta(\tilde{\mathbf{f}},\mathbf{X}) - \Delta(\mathbf{f}^\star,\mathbf{X})]$.

To get a result of the form of Theorem F.2.2, we already know that it suffices to find an upper bound on the quantity

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\kappa}} \Big(\mathbb{E}[\Delta(\mathbf{f},\mathbf{X})] - \Delta_n(\mathbf{f})\Big)\right].$$

As a first step towards this direction, we will need the following lemma, which establishes an interesting link between the length of a curve and its turn. For a proof of this result, we refer the reader to Alexandrov and Reshetnyak [2, Chapter 5].

Lemma F.3.1. *Let* $\mathbf{f}$ *be a curve with turn* κ *and let* δ *be the diameter of* $\mathcal{C}$*. Then* $\mathscr{L}(\mathbf{f}) \leq \delta\zeta(\kappa)$*, where the function* ζ *is defined by*

$$\zeta(x) = \begin{cases} \dfrac{1}{\cos(x/2)} & \text{if} \quad 0 \leq x \leq \dfrac{\pi}{2} \\ 2\sin(x/2) & \text{if} \quad \dfrac{\pi}{2} \leq x \leq \dfrac{2\pi}{3} \\ \dfrac{x}{2} - \dfrac{\pi}{3} + \sqrt{3} & \text{if} \quad x \geq \dfrac{2\pi}{3}. \end{cases}$$

The graph of the function ζ is shown in Figure F.5.

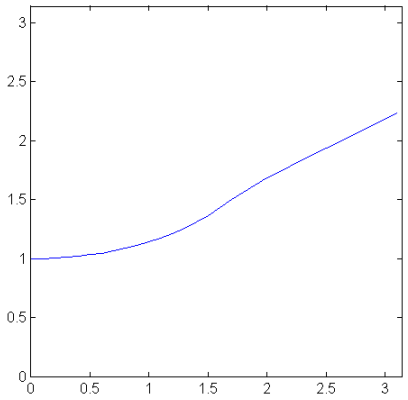


FIGURE F.5.: Graph of the function ζ.

Thanks to this result, the approach developed in Section F.2 adapts to the new context. Proposition F.3.1 below is the counterpart of Proposition F.2.1.

Proposition F.3.1. *Let $\mathcal{F}_{k,\kappa}$ be the set of all polygonal lines with k segments, turn at most κ, and vertices in a grid $\mathcal{Q} \subset \mathcal{C}$, and let δ be the diameter of the convex set $\mathcal{C}$. Then there exist nonnegative constants $a_0, \ldots, a_4$, depending only on the dimension d, such that*

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\kappa}} \Big(\mathbb{E}[\Delta(\mathbf{f},\mathbf{X})] - \Delta_n(\mathbf{f})\Big)\right] \leq \delta^2\left[a_1\sqrt{k} + a_2\sqrt{\zeta(\kappa)} + a_3\frac{\zeta(\kappa)}{\sqrt{k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}<1\}} + a_4\sqrt{k\ln\frac{\zeta(\kappa)}{k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}\geq 1\}} + a_0\right].$$

Putting finally Theorem F.2.1 and Proposition F.3.1 together, we obtain:

Theorem F.3.1. *Consider a family of nonnegative weights $\{x_{k,\kappa}\}_{k\geq 1,\kappa\in\mathcal{K}}$ such that*

$$\sum_{k\geq 1,\kappa\in\mathcal{K}} e^{-x_{k,\kappa}} = \Sigma < \infty,$$

and a penalty function $\mathrm{pen} : \mathbb{N}^* \times \mathcal{K} \to \mathbb{R}^+$. *Let $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k},\hat{\kappa}}$. There exist nonnegative constants $c_0, \ldots, c_3$, depending only on the dimension d, such that, if for all $(k,\kappa) \in \mathbb{N}^* \times \mathcal{K}$,*

$$\mathrm{pen}(k,\kappa) \geq \frac{\delta^2}{\sqrt{n}}\left[c_1\sqrt{k} + c_2\sqrt{\zeta(\kappa)} + c_3\max\left(\frac{\zeta(\kappa)}{\sqrt{k}}, \sqrt{k\ln\frac{\zeta(\kappa)}{k}}\right) + c_0 + \sqrt{\frac{x_{k,\kappa}}{2}}\right],$$

then

$$\mathbb{E}[\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}})] \leq \inf_{k\geq 1,\kappa\in\mathcal{K}}\Big[\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\kappa}) + \mathrm{pen}(k,\kappa)\Big] + \frac{\delta^2\Sigma}{2^{3/2}}\sqrt{\frac{\pi}{n}},$$

where $\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\kappa}) = \inf_{\mathbf{f}\in\mathcal{F}_{k,\kappa}} \mathcal{D}(\mathbf{f}^\star, \mathbf{f})$.

The expression of the penalty shape involves a term of the order $\sqrt{k/n}$—just like in the case of curves with bounded length—, whereas the term depending on the length of the curve $\sqrt{\ell/n}$ is replaced by $\sqrt{\zeta(\kappa)/n}$, which is an increasing function of the turn κ. This is relevant, since the number of segments k and the turn κ characterize the complexity of the models. Moreover, the additive term $\max\left[\zeta(\kappa)/\sqrt{kn}, \sqrt{k\ln\zeta(\kappa)/kn}\right]$ shows that k and κ should be cleverly chosen relatively to each other in order to get a nice principal curve. Roughly, a greater curvature implies more segments.

F.4. Experimental results

This section presents some simulations and real data experiments, carried out with the software MATLAB, to illustrate the model selection procedures suggested by Theorem F.2.2 and Theorem F.3.1. The penalty shapes in these theorems involve constants which have to be practically determined. To this end, a possible route is to use the so-called slope heuristics, introduced by Birgé and Massart [9] and further developed by Arlot and Massart [3]. In short, this calibration method allows to tune a penalty known up to some multiplicative constant. The slope heuristics assumes that the empirical contrast decreases when the complexity of the models increases, which is clearly the case in our principal curve context. The procedure is based on the fact that the graph of the empirical contrast as a function of the penalty shape decreases strongly at the beginning and more slowly later, with a linear trend. At the end, the heuristics specifies that the desired constant is equal to twice the slope of this line.

From an algorithmic perspective, two different strategies were implemented, denoted hereafter by **MS1** and **MS2** (the acronym "**MS**" stands for "model selection").

- The method **MS1**, which is the most closely related to the theory developed in Section F.2 and Section F.3, is based on the simultaneous choice of the number k of segments and the length ℓ of the curve. Precisely, for each number of segments $k = 1, \ldots, 80$ and for a range of values of the length ℓ (the maximal length L and the step depend on the scale of the considered data set), we computed the criterion
$$\Delta_n(\hat{\mathbf{f}}_{k,\ell}) = \frac{1}{n}\sum_{i=1}^{n} \Delta(\hat{\mathbf{f}}_{k,\ell}, \mathbf{X}_i).$$
Then, considering for ease of computation a slightly suboptimal penalty of the form $c_1\sqrt{k}+c_2\ell$, we selected the constants c_1 and c_2 by implementing a bivariate version of the slope heuristics. More precisely, it is assumed that for large values of k and ℓ, the criterion $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ behaves like $c_1\sqrt{k} + c_2\ell$, and the constants $\hat{c}_1$ and $\hat{c}_2$ are chosen thanks to a regression step.
- The second algorithm **MS2** is an adaptation of the Polygonal Line Algorithm proposed by **KKLZ**. In this procedure, the vertices of the principal curve are optimized one after the other in a cyclic manner and the turn is controlled locally, at each step. This is done by means of some local angle penalty, which we set according to **KKLZ** recommendation in [29] and did not try to optimize. Thus, in this second approach, we end up with a penalty of the form $c\sqrt{k/n}$. To calibrate the constant c, we used a MATLAB package called CAPUSHE (CAlibrating Penalties Using the Slope Heuristics), implemented by Baudry, Maugis and Michel in [7].

F.4.1. Simulated data

In this first series of experiments, we considered two-dimensional data distributed with some noise around a reference curve. More formally, observations were generated from the model

$$\mathbf{X} = \mathbf{Y} + \varepsilon,$$

where $\mathbf{Y}$ is uniformly distributed over some planar curve $\mathbf{f}$ and ε is a bivariate Gaussian noise, independent of $\mathbf{Y}$. Even if the generative curve $\mathbf{f}$ is not a principal curve stricto sensu—because of the model bias—, this Gaussian model is considered as a benchmark for simulations in the literature on principal curves.

In a first example, we let $\mathbf{f}$ be a half-circle with radius 1. The noise variance is set to 0.004 and the number n of observations to 100 (see Figure F.6).

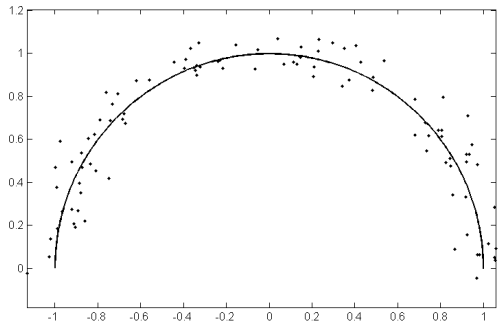


FIGURE F.6.: 100 observations distributed around a half-circle with radius 1.

Recall that the algorithm **MS1** computes the criterion $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ for a table of values of $\sqrt{k}$ and ℓ and selects the best constants according to a bivariate slope heuristics. Figure F.7 shows the surface $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ as a function of $\sqrt{k}$ and ℓ.

Both algorithms were applied to the data set. The resulting selected principal curves are visible in Figure F.8. For comparison purposes, Figures F.9 and F.10 also show some curves obtained by specifying other values for k and ℓ.

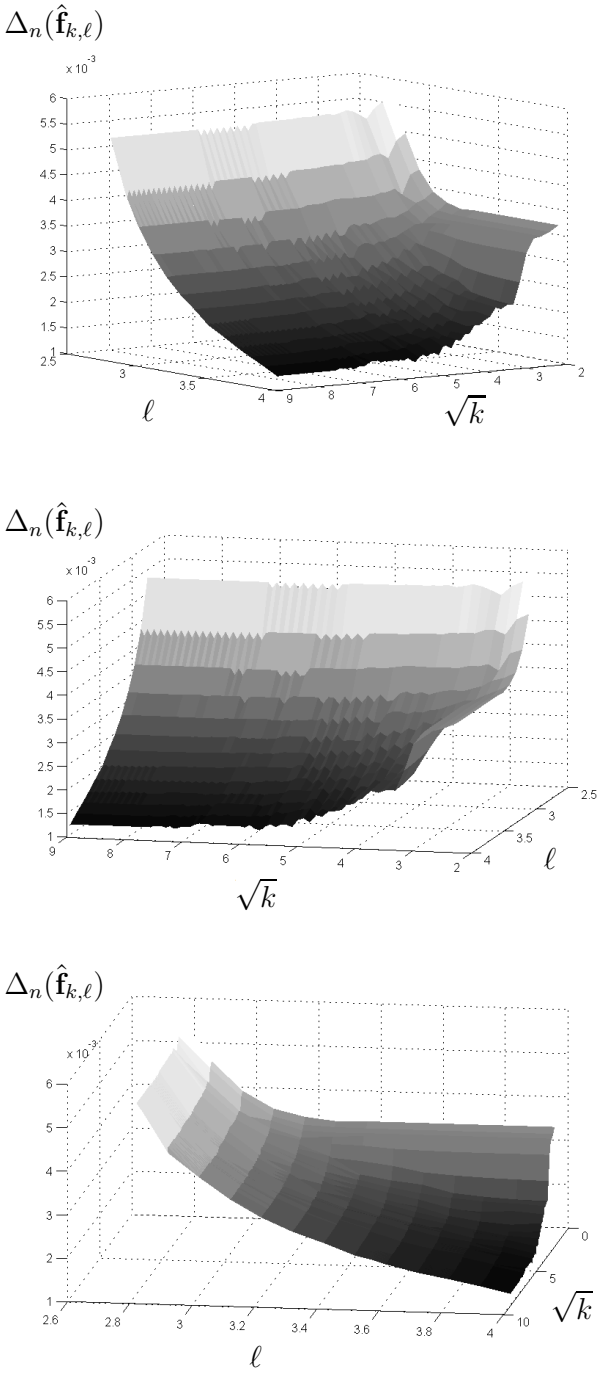


FIGURE F.7.: Criterion $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ as a function of $\sqrt{k}$ and ℓ for the half-circle data (n=100).

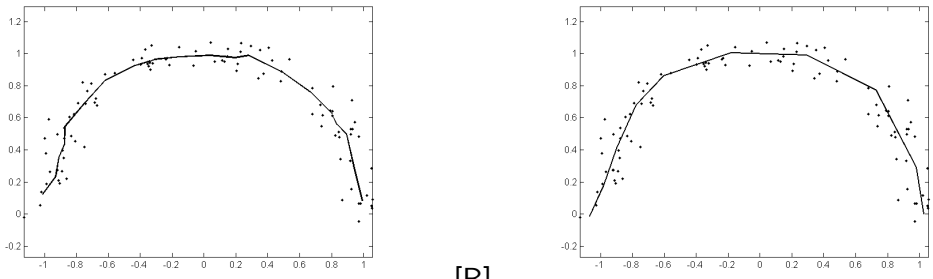


FIGURE F.8.: Selected principal curves for the half-circle data (n=100). [A] Method **MS1**: $\hat{k} = 20$, $\hat{\ell} = 3$. [B] Method **MS2**: $\hat{k} = 9$.

It can be noted that the outputs of both algorithms have approximately the same quality. Indeed, the **MS1** principal curve shows a few irregularities not visible on the **MS2** result, which on the other hand seems rougher, due to the relatively low value of $\hat{k}$.

The methods **MS1** and **MS2** were also tested on a larger sample, shown in Figure F.11. The results for the half-circle data set with $n = 250$ are given in Figure F.12. We observe that both principal curves obtained with this sample size are very accurate.

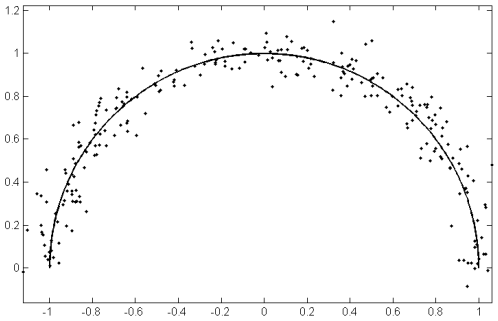


FIGURE F.11.: 250 observations distributed around a half-circle with radius 1.

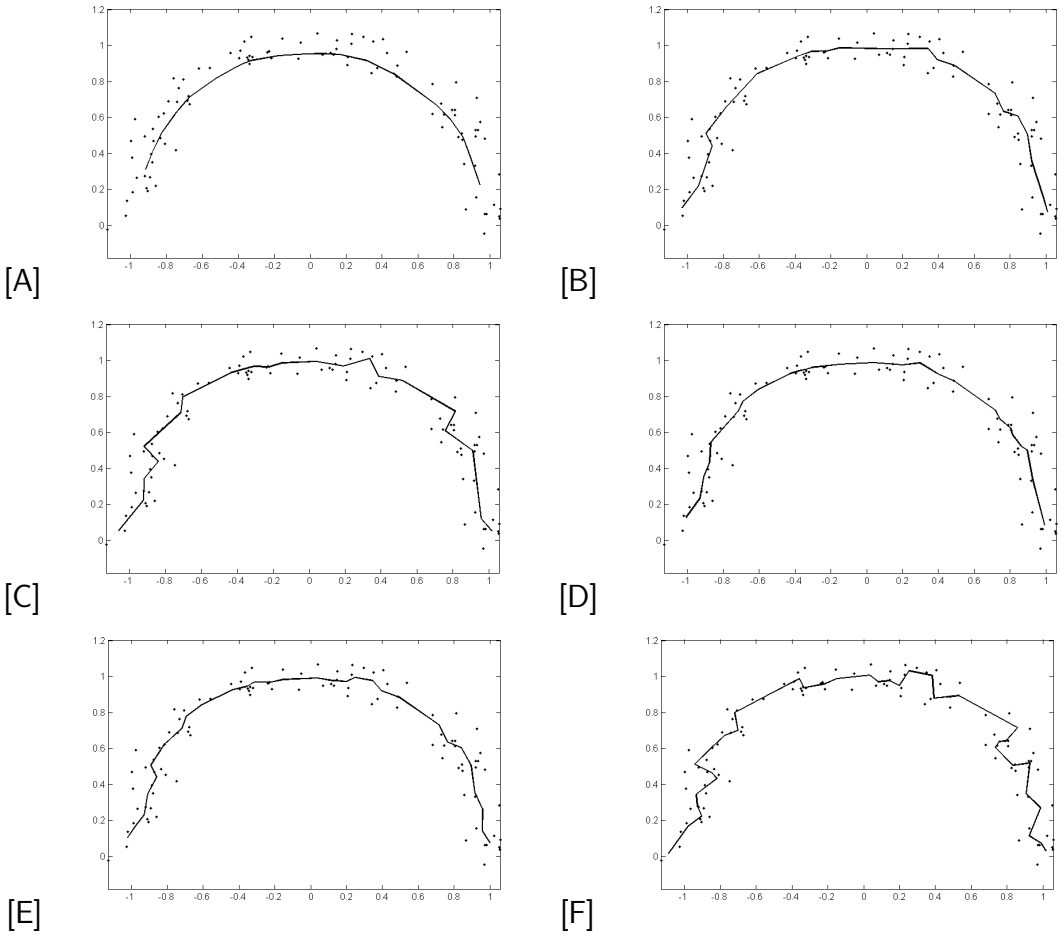


FIGURE F.9.: Method **MS1**: Examples of principal curves for some prespecified values of k and ℓ (n=100). [A] $k = 20$, $\ell = 2.5$. [B] $k = 20$, $\ell = 3.1$. [C] $k = 20$, $\ell = 3.4$. [D] $k = 25$, $\ell = 3$. [E] $k = 30$, $\ell = 3.1$. [F] $k = 35$, $\ell = 4$.

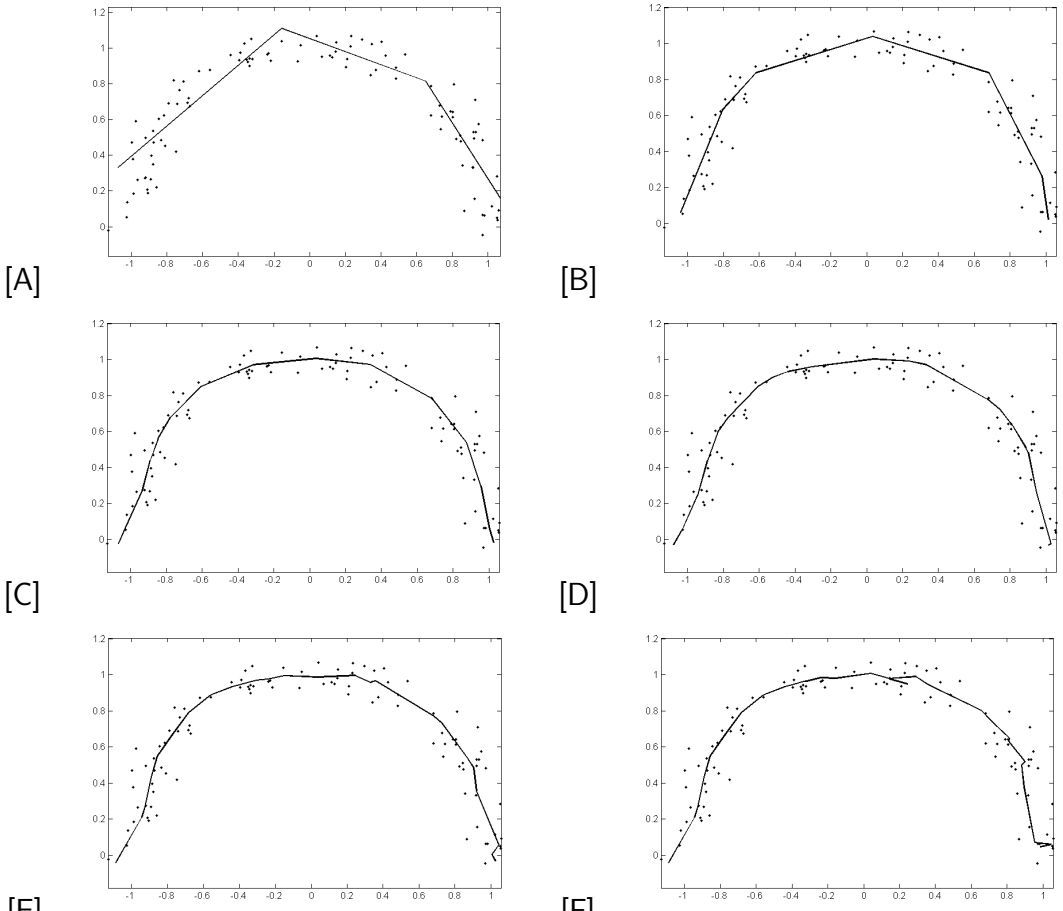
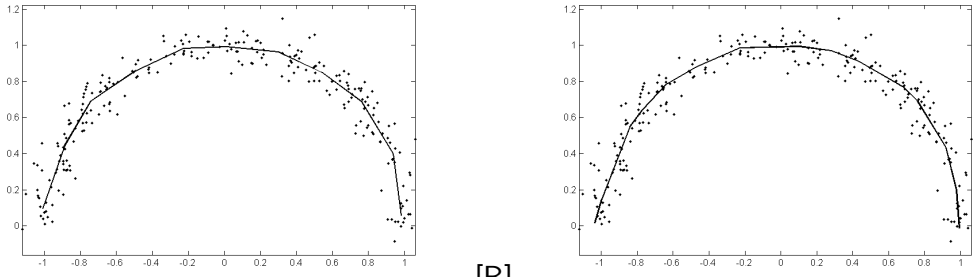


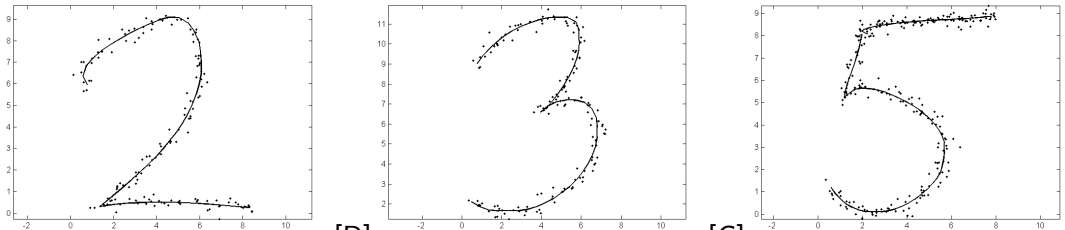
FIGURE F.10.: Method **MS2**: Examples of principal curves for some prespecified values of k (n=100). [A] $k = 3$. [B] $k = 6$. [C] $k = 14$. [D] $k = 20$. [E] $k = 26$. [F] $k = 35$.



[A] [B]

FIGURE F.12.: Selected principal curves for the half-circle data (n=250). [A] Method **MS1**: $\hat{k} = 12$, $\hat{\ell} = 3$. [B] Method **MS2**: $\hat{k} = 14$.

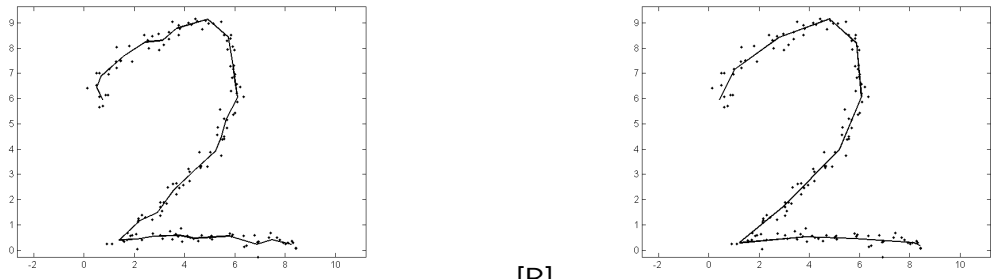
In a second set of numerical examples, we took handwritten-type digits as generative curves, with noise variance 0.04. As depicted in Figure F.13, 150 observations were sampled around the digit 2 and the digit 3 and 250 observations around the digit 5.



[A] [B] [C]

FIGURE F.13.: [A] 150 observations sampled around the digit 2. [B] 150 observations sampled around the digit 3. [C] 250 observations sampled around the digit 5.

Figure F.14 presents the results obtained for the digit 2 with the algorithms **MS1** and **MS2**, whereas Figure F.15 and Figure F.16 show curves corresponding to other choices of the parameters.

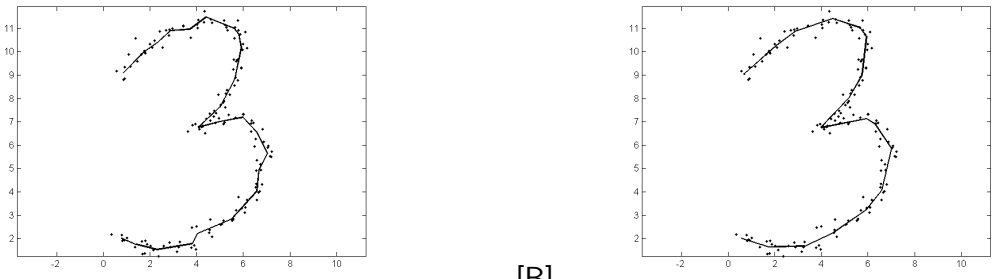


[A] [B]

FIGURE F.14.: Selected principal curves for the digit 2 data (n=150). [A] Method **MS1**: $\hat{k} = 27$, $\hat{\ell} = 24$. [B] Method **MS2**: $\hat{k} = 12$.

With respect to the digit 2 data, the **MS1** principal curve follows the observations more closely than what is expected. On the other hand, the **MS2** output looks better, even if a faintly larger $\hat{k}$ could make the principal curve a bit smoother.

The principal curves fitted for the digits 3 and 5 are shown in Figure F.17 and Figure F.18. For the digit 3, we note again that the algorithm **MS1** slightly overfits the data, whereas the **MS2** output looks a little rough. However, the resulting principal curves for the digit 5 are visually fully satisfactory. On this last example, both algorithms performed similarly and, interestingly, selected almost the same value $\hat{k}$.



[A] [B]

FIGURE F.17.: Selected principal curves for the digit 3 (n=150). [A] Method **MS1**: $\hat{k} = 23$, $\hat{\ell} = 23$. [B] Method **MS2**: $\hat{k} = 18$.

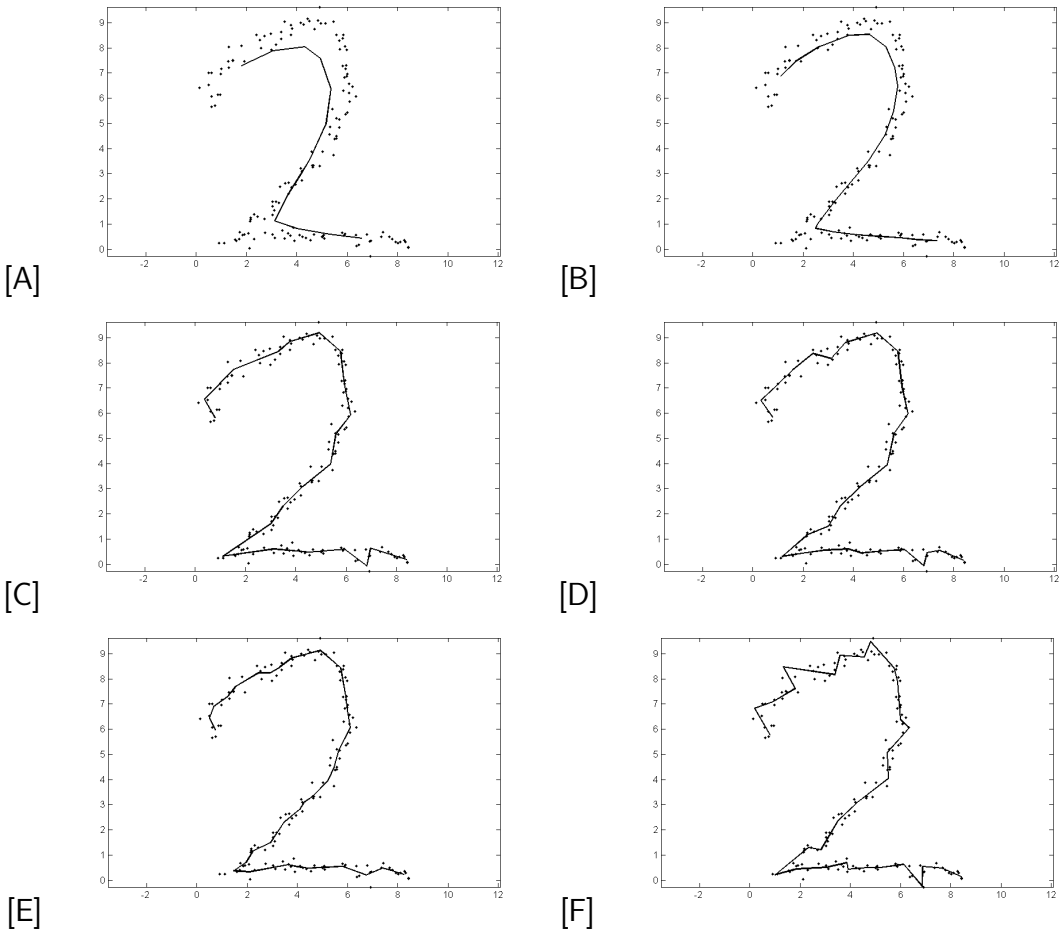


FIGURE F.15.: Method **MS1**: Examples of principal curves for some prespecified values of k and ℓ (n=150). [A] $k = 12$, $\ell = 14$. [B] $k = 20$, $\ell = 18$. [C] $k = 20$, $\ell = 26$. [D] $k = 27$, $\ell = 26$. [E] $k = 35$, $\ell = 24$. [F] $k = 30$, $\ell = 30$.

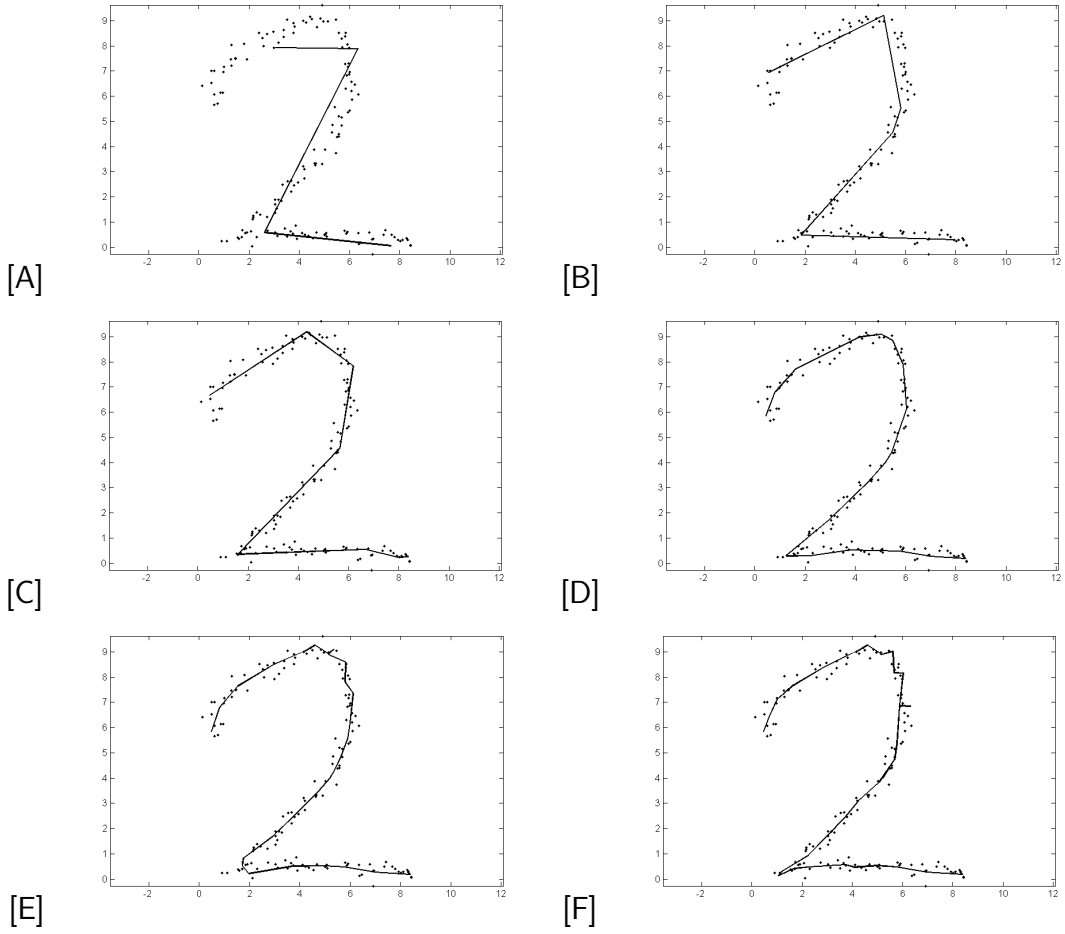


FIGURE F.16.: Method **MS2**: Examples of principal curves for some prespecified values of k (n=150). [A] $k = 3$. [B] $k = 5$. [C] $k = 7$. [D] $k = 20$. [E] $k = 30$. [F] $k = 40$.

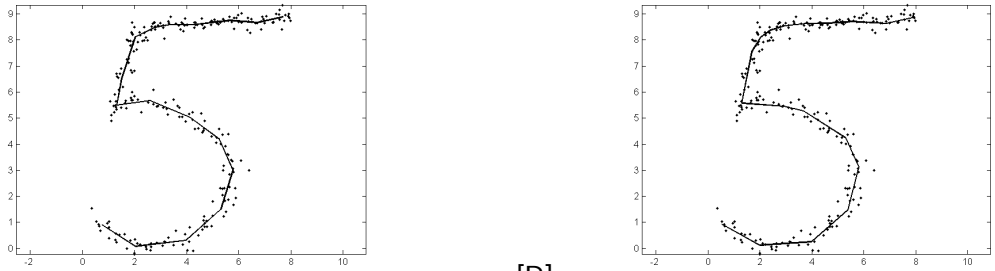


FIGURE F.18.: Selected principal curves for the digit 5 (n=250). [A] Method **MS1**: $\hat{k} = 17$, $\hat{\ell} = 21$. [B] Method **MS2**: $\hat{k} = 18$.

This small simulation study reveals, as expected, that a good automatic choice of the parameters $\hat{k}$ and $\hat{\ell}$ is crucial to obtain a suitable principal curve. On the whole, the visual quality of both algorithms is satisfactory, even if they sometimes lead to somewhat different results. In fact, the principal curves fitted by algorithm **MS1** often follow the data quite closely, in particular when the sample size is not very large, whereas the **MS2** outputs tend to be a bit angular due to the selection of a relatively small $\hat{k}$. Besides, from a computational point of view, **MS1** is, by construction, more CPU–time consuming than **MS2**, since the former algorithm involves the computation of the criterion $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ for a range of values of the length $\ell \leq L$.

Remark F.4.1. A general observation regarding the slope heuristics in **MS2** is that the graph of the criterion Δ_n as a function of k is not always smooth enough to obtain relevant results in the software CAPUSHE. Indeed, by construction of the **KKLZ** Polygonal Line Algorithm, Δ_n does not necessarily decrease from one step to another, which may disturb the slope estimation. Two strategies have been tried to overcome this problem.

The first one consists in deleting all points in the graph of $-\Delta_n$ corresponding to the values of k such that $-\Delta_n(\mathbf{f}_k) < -\Delta_n(\mathbf{f}_{k-1})$. The main drawback of this approach is that it removes a substantial number of values of k. In particular, a whole range of values of k around the right number of segments might be removed, so that selecting a suitable $\hat{k}$ is practically impossible.

Another attempt is to perform several times the vertex optimization step, choosing at random the order in which the vertices are successively updated, and keep the output matching the lowest criterion. This approach effectively allows to smooth the criterion Δ_n, but it induces some annoying variability: it was found that different principal curves were obtained with the same number of segments, leading to more or less good results.

None of the two strategies described above does completely solve the problem. This discussion is illustrated in Figure F.19 with an example of output of the package CAPUSHE.

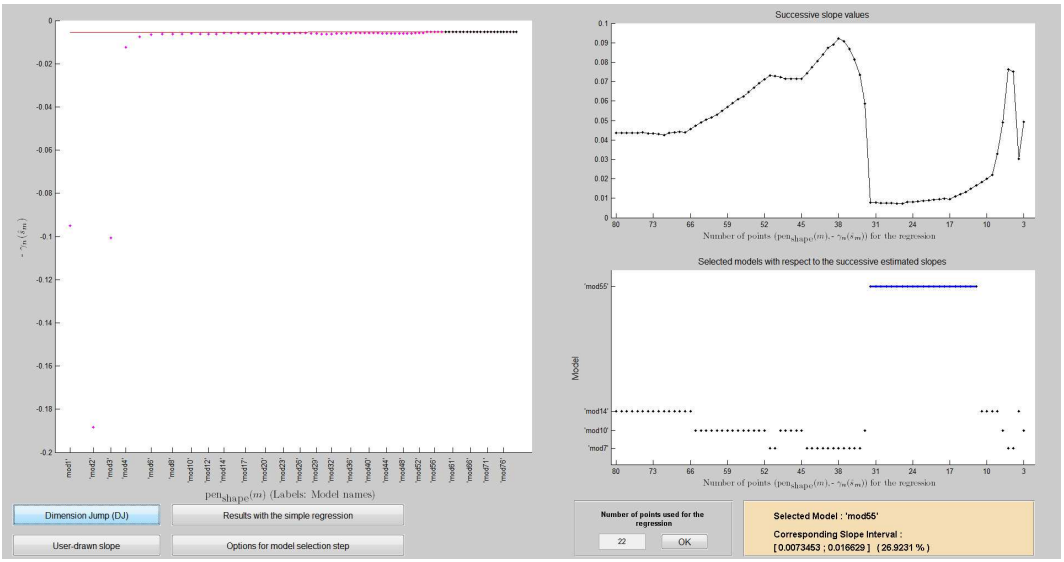


FIGURE F.19.: Example of CAPUSHE output for the half-circle data ($n = 250$): The slope heuristics selects $\hat{k} = 55$. **Left**: Graph of the criterion $-\Delta_n(\hat{\mathbf{f}}_k)$ as a function of $\sqrt{k/n}$. **Upper right**: Successive estimated slope values versus the number of points used for the slope estimation. **Bottom right**: Selected values of $\hat{k}$ versus the number of points used for the slope estimation.

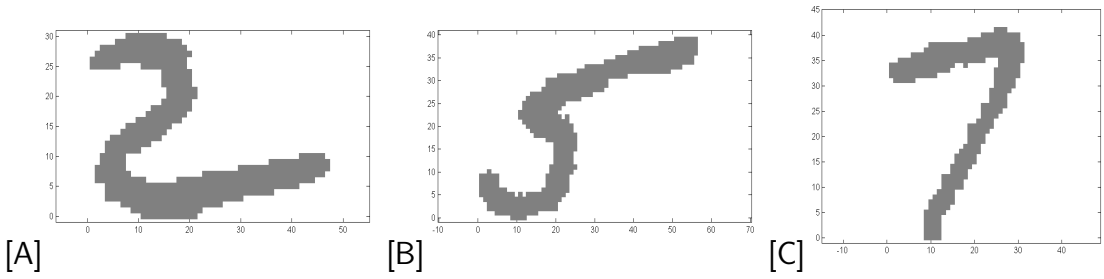
F.4.2. Real data sets

F.4.2.1. NIST database digits

The first real-life data set used in this second series of experiments originated from NIST Special Database 19 (`http://www.nist.gov/srd/nistsd19.cfm`), containing handwritten characters from 3600 writers. The data consists in binary images scanned at 11.8 dots per millimeter (300 dpi), which uniformly fill the area corresponding to the thickness of the pen stroke. Determining the medial axis of such handwritten characters often constitutes a preliminary step to perform character recognition (see, e.g., Deutsch [16] and Alcorn and Hoggar [1]).

Algorithms **MS1** and **MS2** were applied to the three NIST database digits visible in Figure F.20. Figure F.21 shows the surface corresponding to the criterion $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ plotted as a function of $\sqrt{k}$ and ℓ. Principal curves for this digit obtained with the

two methods are depicted in Figure F.22, whereas Figure F.23 and Figure F.24 show some results for other prespecified values of the parameters.



[A] [B] [C]

FIGURE F.20.: Three NIST database handwritten digits.

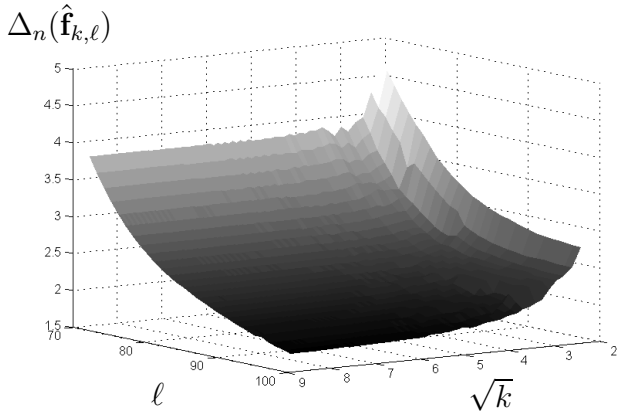
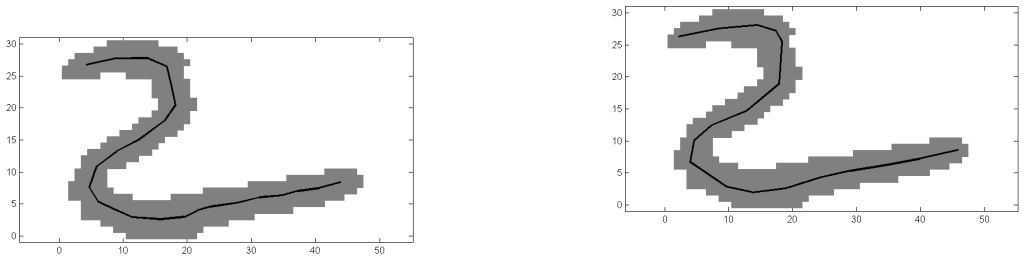
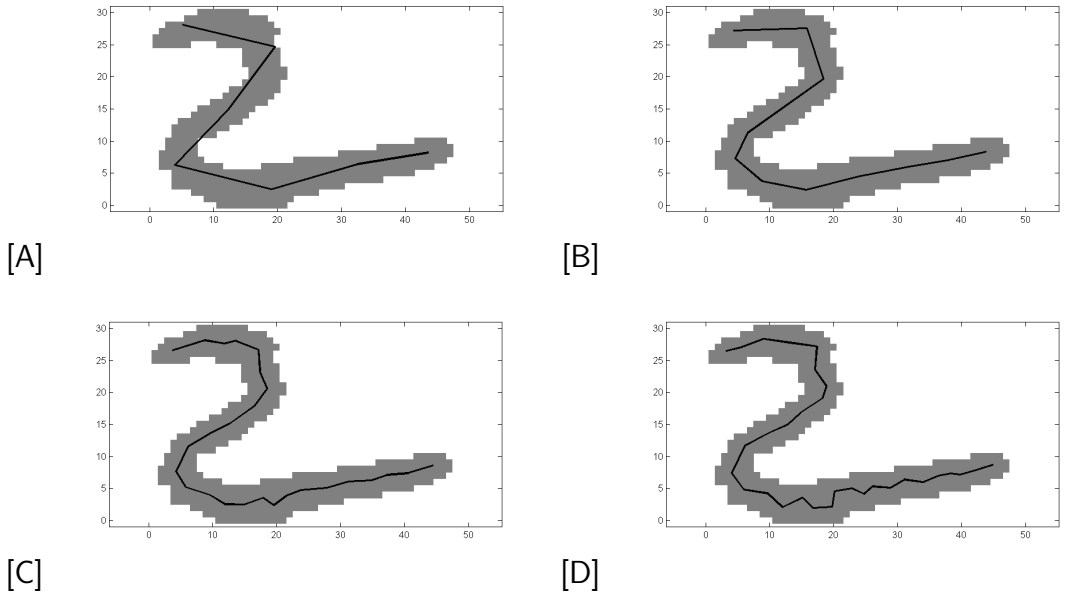


FIGURE F.21.: Criterion $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ as a function of $\sqrt{k}$ and ℓ for the NIST database digit 2 (n=458).



[A] [B]

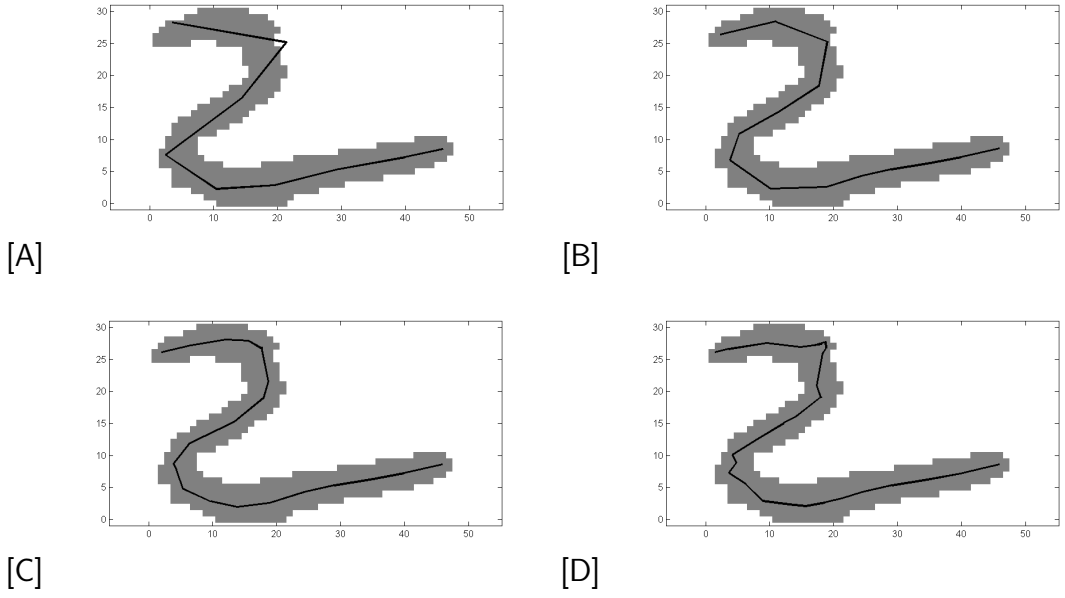
FIGURE F.22.: Selected principal curves for the digit 2 (n=458). [A] Method **MS1**: $\hat{k} = 23$, $\hat{\ell} = 80$. [B] Method **MS2**: $\hat{k} = 17$.



[A] [B]

[C] [D]

FIGURE F.23.: Method **MS1**: Examples of principal curves for some prespecified values of k and ℓ. [A] $k = 6$, $\ell = 80$. [B] $k = 10$, $\ell = 80$. [C] $k = 25$, $\ell = 84$. [D] $k = 30$, $\ell = 90$.



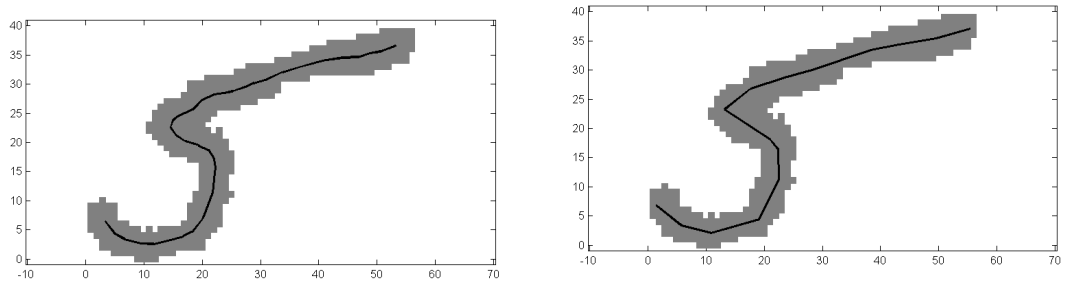
[A] [B]

[C] [D]

FIGURE F.24.: Method **MS2**: Examples of principal curves for some prespecified values of k. [A] $k = 8$. [B] $k = 13$. [C] $k = 20$. [D] $k = 30$.

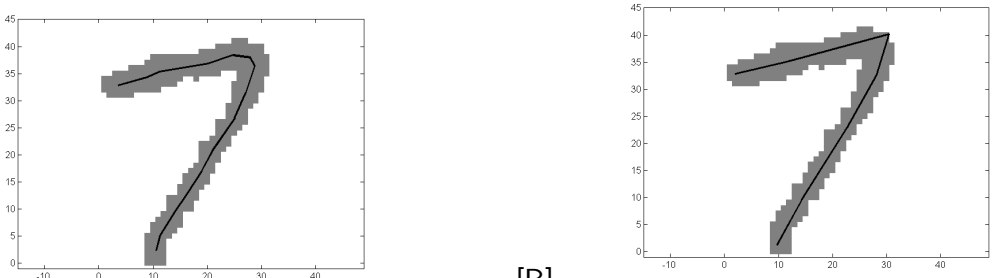
We observe that both results for the digit 2 are quite similar, with a slight advantage to **MS1** however. Indeed, this algorithm yields a smoother curve which better recovers the loop of the digit.

Figure F.25 and F.26 show the outputs of the algorithms **MS1** and **MS2** for the NIST digits 5 and 7.



[A] [B]

FIGURE F.25.: Selected principal curves for the digit 5 (n=513). [A] Method **MS1**: $\hat{k} = 38$, $\hat{\ell} = 82$. [B] Method **MS2**: $\hat{k} = 14$.



[A] [B]

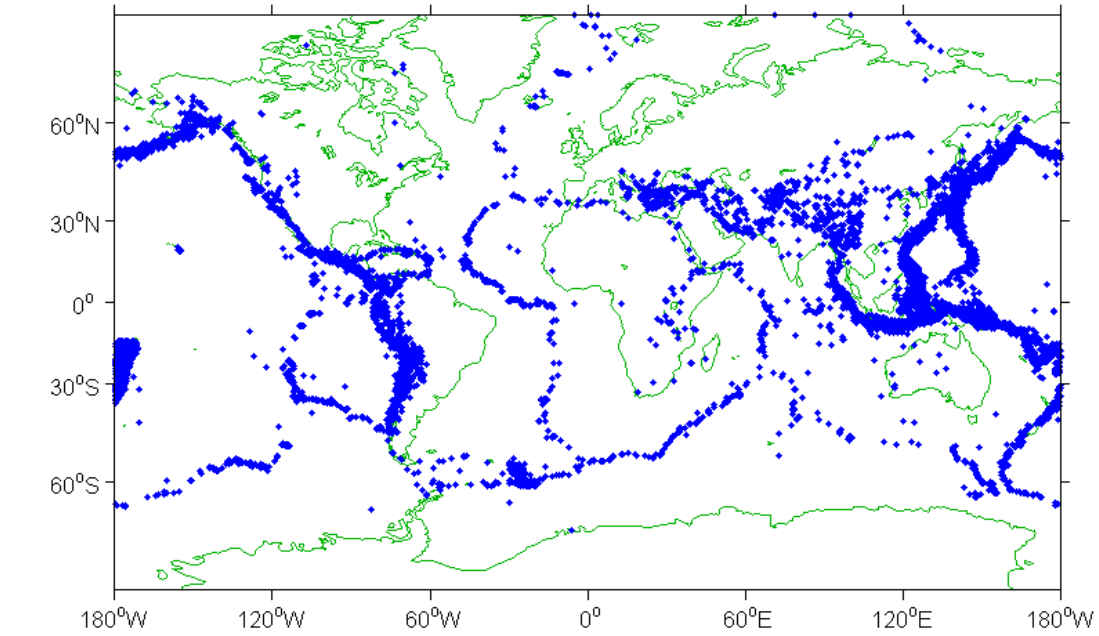
FIGURE F.26.: Selected principal curves for the digit 7 (n=334). [A] Method **MS1**: $\hat{k} = 15$, $\hat{\ell} = 66$. [B] Method **MS2**: $\hat{k} = 6$.

As a general conclusion on these NIST digit data sets, we found that both methods perform quite well. Here, **MS1** does not seem to overfit, probably because the sample size is large enough. As in the simulated data examples, it nevertheless appears that the curves estimated by the algorithm **MS2** could be smoother with a slightly larger number $\hat{k}$ of segments.

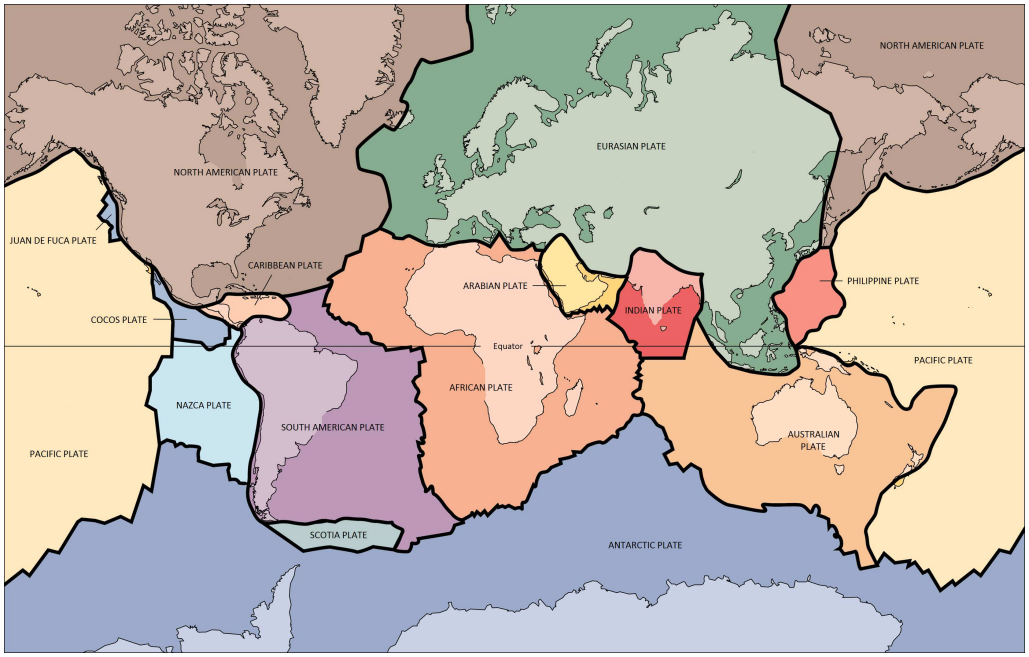
F.4.2.2. Seismic data

Together with satellite images, the localization of earthquakes is an essential source of information in geology for the study of seismic faults, whether in accretion or subduction regions. As an illustration, Figure F.27 depicts seismic impacts in the world—the map is drawn using Miller's projection—, as well as a world map from the USGS (United States Geological Survey) showing the various lithospheric plates. The data set, which can be downloaded on the USGS website (`http://earthquake.usgs.gov/research/data/centennial.php`), is part of the "Centennial Catalog", listing the major earthquakes registered since 1900 (Engdahl et Villaseñor [21]). In this subsection, we employ algorithms **MS1** and **MS2** as a means to recover the borders of lithospheric plates using the earthquake localization data of Figure F.27.

We decided to focus on two particularly representative seismic active zones. The first one (**Z1** hereafter) is located in the Atlantic Ocean, to the west of the African continent (about 60°S 50°W to 40°N 0°), and the second one (**Z2** hereafter) extends from the south of Africa to the south of Australia (about 65°S 0° to 25°S 160°E). The localization of these two regions on the world map is visible in Figure F.28. The results for **Z1** are shown in Figure F.29 and for **Z2** in Figure F.30.



[A]



[B]

FIGURE F.27.: [A] Earthquake impacts and [B] lithospheric plate borders.

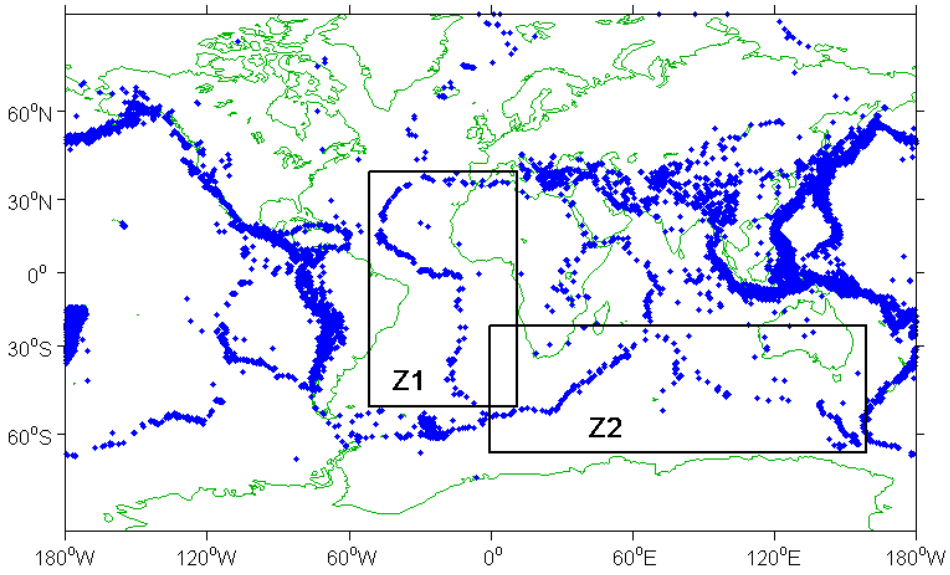
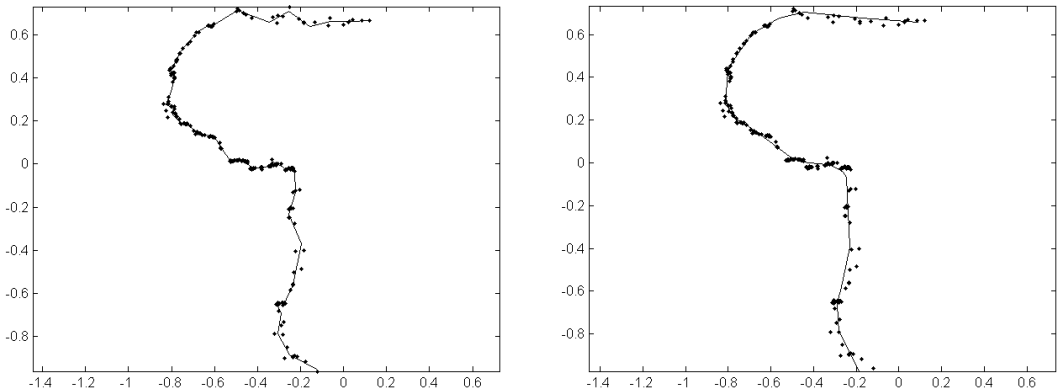


FIGURE F.28.: Localization of the two considered seismic zones **Z1** (about 60°S 50°W to 40°N 0°) and **Z2** (about 65°S 0° to 25°S 160°E).



[A] [B]

FIGURE F.29.: Selected principal curves for the seismic zone **Z1** (n=252). [A] Method **MS1**: $\hat{k} = 55$, $\hat{\ell} = 31$. [B] Method **MS2**: $\hat{k} = 30$.

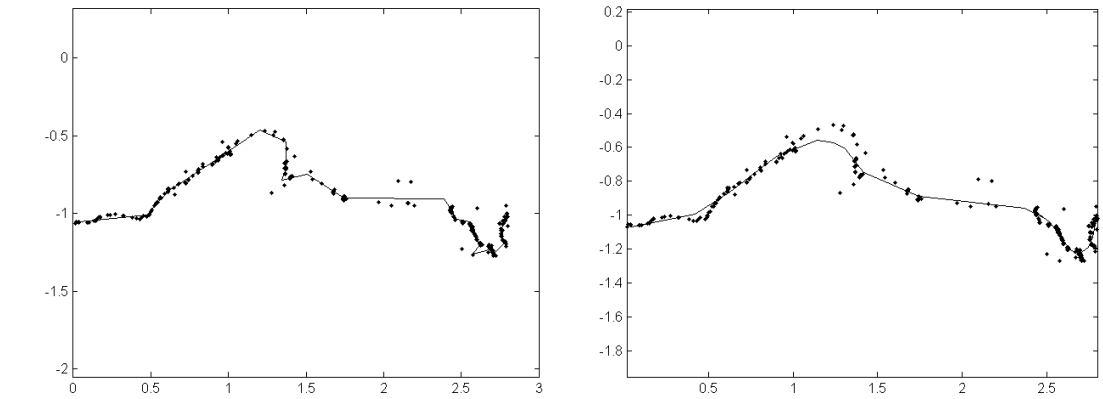


FIGURE F.30.: Selected principal curves for the seismic zone **Z2** (n=322). [A] Method **MS1**: $\hat{k} = 22$, $\hat{\ell} = 38$. [B] Method **MS2**: $\hat{k} = 20$.

In Figure F.29, we see, for the seismic zone **Z1**, that the method **MS1** again yields a principal curve following the data points quite closely. On the contrary, the algorithm **MS2** provides a smoother curve, which at first sight seems a better result. However, the border of the lithospheric plate is probably more likely to look like the more irregular **MS1** principal curve, as suggested by Figure F.27 [B]. The same observation holds for **Z2** (Figure F.30). Moreover, in this case, the **MS2** output does not recover the shape of the plate border, which certainly passes through the most northern points and not several degrees south. Apparently, the local penalty on the angles leads here to overpenalization. Thus, on this seismic data set, **MS1** results seem to be more relevant.

It is noteworthy that using this type of earthquake data to draw faults could be especially useful to locate some faults which cannot be easily spotted and necessitate monitoring for seismic risk prevention. With this respect, Harding and Berghoff [24], employing a method based on airborne laser mapping, study for instance seismic hazards in a zone densely covered by vegetation, located in the Puget Lowland of Washington State, USA. Using a principal curve approach to solve this kind of problems is undoubtedly an interesting project for future research.

F.5. Proofs

F.5.1. Proof of Theorem F.2.1

Theorem F.2.1 is an adaptation of Theorem 8.1 in Massart [33]. We first recall the following lemma, which is a consequence of McDiarmid's inequality [34] (see

Massart [33, Theorem 5.3]).

Lemma F.5.1. *If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent random variables and $\mathcal{G}$ is a finite or countable class of real-valued functions such that $a \leq g \leq b$ for every function $g \in \mathcal{G}$, then, setting $Z = \sup_{g\in\mathcal{G}} \sum_{i=1}^n (g(\mathbf{X}_i) - \mathbb{E}[g(\mathbf{X}_i)])$, we have, for every $\varepsilon \geq 0$,*

$$\mathbb{P}\left\{Z - \mathbb{E}[Z] \geq \varepsilon\right\} \leq \exp\left(-\frac{2\varepsilon^2}{n(b-a)^2}\right).$$

Proof of the theorem. Let $\bar{\Delta}_n(\mathbf{f}) = \Delta_n(\mathbf{f}) - \mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})]$ denote the centered empirical process. For all $k \geq 1$ and $\ell \in \mathcal{L}$, for any $\mathbf{f}_{k,\ell} \in \mathcal{F}_{k,\ell}$, we have, by definition of $\tilde{\mathbf{f}}$,

$$\Delta_n(\tilde{\mathbf{f}}) + \mathrm{pen}(\hat{k}, \hat{\ell}) \leq \Delta_n(\mathbf{f}_{k,\ell}) + \mathrm{pen}(k, \ell).$$

Equivalently,

$$\Delta_n(\tilde{\mathbf{f}}) - \Delta_n(\mathbf{f}_{k,\ell}) \leq \mathrm{pen}(k, \ell) - \mathrm{pen}(\hat{k}, \hat{\ell}).$$

Since $\Delta_n(\tilde{\mathbf{f}}) = \mathbb{E}[\Delta(\tilde{\mathbf{f}}, \mathbf{X})] + \bar{\Delta}_n(\tilde{\mathbf{f}})$ and $\Delta_n(\mathbf{f}_{k,\ell}) = \mathbb{E}[\Delta(\mathbf{f}_{k,\ell}, \mathbf{X})] + \bar{\Delta}_n(\mathbf{f}_{k,\ell})$, this inequality becomes

$$\mathbb{E}[\Delta(\tilde{\mathbf{f}}, \mathbf{X})] - \mathbb{E}[\Delta(\mathbf{f}_{k,\ell}, \mathbf{X})] \leq \bar{\Delta}_n(\mathbf{f}_{k,\ell}) - \bar{\Delta}_n(\tilde{\mathbf{f}}) + \mathrm{pen}(k, \ell) - \mathrm{pen}(\hat{k}, \hat{\ell}). \tag{F.4}$$

Moreover, for every $\mathbf{f} \in \mathcal{F}$,

$$\mathcal{D}(\mathbf{f}^\star, \mathbf{f}) = \mathbb{E}[\Delta(\mathbf{f}, \mathbf{X}) - \Delta(\mathbf{f}^\star, \mathbf{X})],$$

so that

$$\mathbb{E}[\Delta(\tilde{\mathbf{f}}, \mathbf{X})] - \mathbb{E}[\Delta(\mathbf{f}_{k,\ell}, \mathbf{X})] = \mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) - \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}). \tag{F.5}$$

Therefore, combining (F.4) and (F.5),

$$\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) \leq \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) - \bar{\Delta}_n(\tilde{\mathbf{f}}) + \mathrm{pen}(k, \ell) - \mathrm{pen}(\hat{k}, \hat{\ell}). \tag{F.6}$$

Consider now a family of nonnegative weights $\{x_{k,\ell}\}_{k\geq 1, \ell\in\mathcal{L}}$ such that

$$\sum_{k\geq 1, \ell\in\mathcal{L}} e^{-x_{k,\ell}} = \Sigma < \infty,$$

and let $z > 0$. Applying Lemma F.5.1, we get, for all $k' \geq 1$, $\ell' \in \mathcal{L}$ and $\varepsilon \geq 0$,

$$\mathbb{P}\left\{\sup_{\mathbf{f}\in\mathcal{F}_{k',\ell'}} (-\bar{\Delta}_n(\mathbf{f})) \geq \mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k',\ell'}} (-\bar{\Delta}_n(\mathbf{f}))\right] + \varepsilon\right\} \leq \exp\left(-\frac{2n\varepsilon^2}{\delta^4}\right).$$

This may be rewritten, for $\varepsilon = \delta^2\sqrt{\frac{x_{k',\ell'}+z}{2n}}$,

$$\mathbb{P}\left\{\sup_{\mathbf{f}\in\mathcal{F}_{k',\ell'}} (-\bar{\Delta}_n(\mathbf{f})) \geq \mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k',\ell'}} (-\bar{\Delta}_n(\mathbf{f}))\right] + \delta^2\sqrt{\frac{x_{k',\ell'}+z}{2n}}\right\} \leq e^{-x_{k',\ell'}-z}.$$

Setting $E_{k',\ell'} = \mathbb{E}\Big[\sup_{\mathbf{f}\in\mathcal{F}_{k',\ell'}}(-\bar{\Delta}_n(\mathbf{f}))\Big]$, we thus have, for all $k' \geq 1$ and $\ell' \in \mathcal{L}$,

$$\sup_{\mathbf{f}\in\mathcal{F}_{k',\ell'}}(-\bar{\Delta}_n(\mathbf{f})) \leq E_{k',\ell'} + \delta^2\sqrt{\frac{x_{k',\ell'}+z}{2n}},$$

except on a set of probability not larger than Σe^{-z}. Then, inequality (F.6) implies

$$\begin{aligned}\mathcal{D}(\mathbf{f}^\star,\tilde{\mathbf{f}}) &\leq \mathcal{D}(\mathbf{f}^\star,\mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + E_{\hat{k},\hat{\ell}} + \delta^2\sqrt{\frac{x_{\hat{k},\hat{\ell}}+z}{2n}} - \mathrm{pen}(\hat{k},\hat{\ell}) + \mathrm{pen}(k,\ell)\\ &\leq \mathcal{D}(\mathbf{f}^\star,\mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + E_{\hat{k},\hat{\ell}} + \delta^2\sqrt{\frac{x_{\hat{k},\hat{\ell}}}{2n}} - \mathrm{pen}(\hat{k},\hat{\ell}) + \mathrm{pen}(k,\ell) + \delta^2\sqrt{\frac{z}{2n}},\end{aligned}$$

except on a set of probability not larger than Σe^{-z}. Consequently, if for all $k' \geq 1$ and $\ell' \in \mathcal{L}$,

$$\mathrm{pen}(k',\ell') \geq E_{k',\ell'} + \delta^2\sqrt{\frac{x_{k',\ell'}}{2n}},$$

then

$$\mathcal{D}(\mathbf{f}^\star,\tilde{\mathbf{f}}) \leq \mathcal{D}(\mathbf{f}^\star,\mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + \mathrm{pen}(k,\ell) + \delta^2\sqrt{\frac{z}{2n}},$$

except on a set of probability not larger than Σe^{-z}. Put differently,

$$\mathbb{P}\left\{\delta^{-2}\sqrt{2n}[\mathcal{D}(\mathbf{f}^\star,\tilde{\mathbf{f}}) - \mathcal{D}(\mathbf{f}^\star,\mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + \mathrm{pen}(k,\ell)] \geq \sqrt{z}\right\} \leq \Sigma e^{-z},$$

or, letting $z = u^2$,

$$\mathbb{P}\left\{[\delta^{-2}\sqrt{2n}[\mathcal{D}(\mathbf{f}^\star,\tilde{\mathbf{f}}) - \mathcal{D}(\mathbf{f}^\star,\mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + \mathrm{pen}(k,\ell)] \geq u\right\} \leq \Sigma e^{-u^2}.$$

Recalling that $\int_0^\infty e^{-u^2}\mathrm{d}u = \frac{\sqrt{\pi}}{2}$ and letting $g_+ = \max(g,0)$, we obtain

$$\mathbb{E}\Big[(\mathcal{D}(\mathbf{f}^\star,\tilde{\mathbf{f}}) - \mathcal{D}(\mathbf{f}^\star,\mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + \mathrm{pen}(k,\ell))_+\Big] \leq \frac{\delta^2\Sigma}{2^{3/2}}\sqrt{\frac{\pi}{n}}.$$

Hence, as $\mathbb{E}[\bar{\Delta}_n(\mathbf{f}_{k,\ell})] = 0$,

$$\mathbb{E}[\mathcal{D}(\mathbf{f}^\star,\tilde{\mathbf{f}})] \leq \mathcal{D}(\mathbf{f}^\star,\mathbf{f}_{k,\ell}) + \mathrm{pen}(k,\ell) + \frac{\delta^2\Sigma}{2^{3/2}}\sqrt{\frac{\pi}{n}}.$$

Since this is true for all k and ℓ, we finally get

$$\mathbb{E}[\mathcal{D}(\mathbf{f}^\star,\tilde{\mathbf{f}})] \leq \inf_{k\geq 1,\ell\in\mathcal{L}}\Big[\mathcal{D}(\mathbf{f}^\star,\mathcal{F}_{k,\ell}) + \mathrm{pen}(k,\ell)\Big] + \frac{\delta^2\Sigma}{2^{3/2}}\sqrt{\frac{\pi}{n}},$$

where $\mathcal{D}(\mathbf{f}^\star,\mathcal{F}_{k,\ell}) = \inf_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\mathcal{D}(\mathbf{f}^\star,\mathbf{f})$. This concludes the proof of Theorem F.2.1.

F.5.2. Proof of Proposition F.2.1

The first step consists in proving that the quantity

$$\mathbb{E}\bigg[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\left(\mathbb{E}[\Delta(\mathbf{f},\mathbf{X})]-\Delta_n(\mathbf{f})\right)\bigg]$$

may be upper bounded by means of the Rademacher average

$$\mathbb{E}\bigg[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta(\mathbf{f},\mathbf{X}_i)\bigg],$$

where $\varepsilon_1,\ldots,\varepsilon_n$ are independent Rademacher random variables, defined by

$$\mathbb{P}\left\{\varepsilon_i=1\right\}=\mathbb{P}\left\{\varepsilon_i=-1\right\}=1/2,$$

independent of $\mathbf{X}_1,\ldots,\mathbf{X}_n$. Let $\mathbf{X}'_1,\ldots,\mathbf{X}'_n$ be independent copies of $\mathbf{X}_1,\ldots,\mathbf{X}_n$, also independent of $\varepsilon_1,\ldots,\varepsilon_n$. A symmetrization argument yields

$$\begin{aligned}
&\mathbb{E}\bigg[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\left(\mathbb{E}[\Delta(\mathbf{f},\mathbf{X})]-\Delta_n(\mathbf{f})\right)\bigg]\\
&\quad=\mathbb{E}\bigg[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\Big(\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\Delta(\mathbf{f},\mathbf{X}'_i)\Big|\mathbf{X}_1,\ldots,\mathbf{X}_n\Big]-\frac{1}{n}\sum_{i=1}^{n}\Delta(\mathbf{f},\mathbf{X}_i)\Big)\bigg]\\
&\quad\leq\mathbb{E}\bigg[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\frac{1}{n}\sum_{i=1}^{n}\Big(\Delta(\mathbf{f},\mathbf{X}'_i)-\Delta(\mathbf{f},\mathbf{X}_i)\Big)\bigg]\\
&\quad=\mathbb{E}\bigg[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Big(\Delta(\mathbf{f},\mathbf{X}'_i)-\Delta(\mathbf{f},\mathbf{X}_i)\Big)\bigg]\\
&\quad\leq\mathbb{E}\bigg[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta(\mathbf{f},\mathbf{X}'_i)\bigg]+\mathbb{E}\bigg[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\frac{1}{n}\sum_{i=1}^{n}(-\varepsilon_i)\Delta(\mathbf{f},\mathbf{X}_i)\bigg]\\
&\quad=2\mathbb{E}\bigg[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta(\mathbf{f},\mathbf{X}_i)\bigg].
\end{aligned}$$

Next, the Rademacher average

$$\mathbb{E}\bigg[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta(\mathbf{f},\mathbf{X}_i)\bigg]$$

may be bounded by resorting to a Dudley integral. More precisely, let

$$S_{k,\ell}=\{\Delta(\mathbf{f},\cdot),\mathbf{f}\in\mathcal{F}_{k,\ell}\}$$

be a subset of the continuous functions from $\mathcal{C}$ to $\mathbb{R}^+$, endowed with the sup-norm $\|\cdot\|_\infty$, and denote by $\mathcal{N}(S_{k,\ell},\|\cdot\|_\infty,\varepsilon)$ the covering number of $S_{k,\ell}$, i.e., the minimal

number of closed balls of radius ε needed to cover $S_{k,\ell}$. According to Dudley [18], there exists an absolute constant $c > 0$ such that, for all $\mathbf{X}_1, \ldots, \mathbf{X}_n$,

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \Delta(\mathbf{f}, \mathbf{X}_i)\right] \leq \frac{c}{\sqrt{n}} \int_0^\infty \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon.$$

To evaluate the covering number of $S_{k,\ell}$, we may use Lemma 2 in Kégl [27], which ensures that

$$\mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon) \leq 2^{\ell\delta/\varepsilon + 3k+1} V_d^{k+1} \left[\frac{\delta^2\sqrt{d}}{\varepsilon} + \sqrt{d}\right]^d \left[\frac{\ell\delta\sqrt{d}}{k\varepsilon} + 3\sqrt{d}\right]^{kd},$$

where V_d denotes the volume of the d-dimensional unit ball. Observe that

$$\begin{aligned}
&\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon) \\
&\quad\leq \left[\frac{\ell\delta}{\varepsilon} + 3k + 1\right] \ln 2 + (k+1)\ln V_d + d\ln\left[\frac{\delta^2\sqrt{d}}{\varepsilon} + \sqrt{d}\right] + kd\ln\left[\frac{\ell\delta\sqrt{d}}{k\varepsilon} + 3\sqrt{d}\right] \\
&\quad= \frac{\ell\delta}{\varepsilon}\ln 2 + (3k+1)\ln 2 + (k+1)\ln V_d + d(k+1)\ln\sqrt{d} + d\ln\left[\frac{\delta^2}{\varepsilon} + 1\right] \\
&\qquad + kd\ln 3 + kd\ln\left[\frac{\ell\delta}{3k\varepsilon} + 1\right] \\
&\quad= \frac{\ell\delta}{\varepsilon}\ln 2 + d\ln\left(\frac{\delta^2}{\varepsilon} + 1\right) + kd\ln\left(\frac{\ell\delta}{3k\varepsilon} + 1\right) + kd\ln 3 + (3k+1)\ln 2 \\
&\qquad + (k+1)(\ln V_d + \frac{d}{2}\ln d).
\end{aligned}$$

Hence, recalling that the support of $\mathbf{f}$ is included in a set $\mathcal{C}$ with diameter δ, we obtain

$$\begin{aligned}
\int_0^\infty \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon &= \int_0^{\delta^2} \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon \\
&\leq \sqrt{\ell\delta\ln 2} I_1 + \sqrt{d} I_2 + \sqrt{kd} I_3 + \delta^2 A(k,d),
\end{aligned}$$

where $I_1 = \displaystyle\int_0^{\delta^2} \frac{1}{\sqrt{\varepsilon}}\mathrm{d}\varepsilon$, $I_2 = \displaystyle\int_0^{\delta^2} \sqrt{\ln\left(\frac{\delta^2}{\varepsilon} + 1\right)}\mathrm{d}\varepsilon$, $I_3 = \displaystyle\int_0^{\delta^2} \sqrt{\ln\left(\frac{\ell\delta}{3k\varepsilon} + 1\right)}\mathrm{d}\varepsilon$, and

$$A(k,d) = \left[kd\ln 3 + (3k+1)\ln 2 + (k+1)(\ln V_d + \frac{d}{2}\ln d)\right]^{1/2}.$$

Control of $\mathbf{I_1}$. Clearly,

$$I_1 = \int_0^{\delta^2} \frac{1}{\sqrt{\varepsilon}}\mathrm{d}\varepsilon = 2\delta.$$

Control of $\mathbf{I_2}$. We have

$$\begin{aligned} I_2 &\leq \int_0^{\delta^2} \sqrt{\ln\Big(\frac{2\delta^2}{\varepsilon}\Big)}\mathrm{d}\varepsilon \\ &= 2\delta^2 \int_0^{1/2} \sqrt{\ln\frac{1}{u}}\mathrm{d}u \\ &\leq \delta^2(\sqrt{\ln 2} + \sqrt{\pi}). \end{aligned}$$

Control of $\mathbf{I_3}$. Let $M = \max(3k, L/\delta)$. Clearly, for all $\ell \in \mathcal{L}$, $\delta \geq \frac{\ell}{M}$, and then $\delta^2 \geq \frac{\ell\delta}{M}$. Let us cut up the integral I_3 and write

$$\begin{aligned} I_3 &= \int_0^{\delta^2} \sqrt{\ln\Big(\frac{\ell\delta}{3k\varepsilon}+1\Big)}\mathrm{d}\varepsilon \\ &= \int_0^{\ell\delta/M} \sqrt{\ln\Big(\frac{\ell\delta}{3k\varepsilon}+1\Big)}\mathrm{d}\varepsilon + \int_{\ell\delta/M}^{\delta^2} \sqrt{\ln\Big(\frac{\ell\delta}{3k\varepsilon}+1\Big)}\mathrm{d}\varepsilon. \end{aligned} \tag{F.7}$$

Observe, since $\varepsilon \leq \frac{\ell\delta}{M}$, that $\frac{\ell\delta}{3k\varepsilon} \geq 1$. Consequently,

$$\begin{aligned} \int_0^{\ell\delta/M} \sqrt{\ln\Big(\frac{\ell\delta}{3k\varepsilon}+1\Big)}\mathrm{d}\varepsilon &\leq \int_0^{\ell\delta/M} \sqrt{\ln\Big(\frac{2\ell\delta}{3k\varepsilon}\Big)}\mathrm{d}\varepsilon \\ &= \frac{2\ell\delta}{3k}\int_0^{3k/2M} \sqrt{\ln\frac{1}{u}}\mathrm{d}u \\ &\leq \frac{\ell\delta}{M}\bigg(\sqrt{\ln\Big(\frac{2M}{3k}\Big)} + \sqrt{\pi}\bigg). \end{aligned}$$

The second integral in equality (F.7) may be bounded using the fact that the integrand is a decreasing function of ε:

$$\begin{aligned} \int_{\ell\delta/M}^{\delta^2} \sqrt{\ln\Big(\frac{\ell\delta}{3k\varepsilon}+1\Big)}\mathrm{d}\varepsilon &\leq \Big(\delta^2 - \frac{\ell\delta}{M}\Big)\sqrt{\ln\Big(\frac{M}{3k}+1\Big)} \\ &\leq \Big(\delta^2 - \frac{\ell\delta}{M}\Big)\sqrt{\ln\Big(\frac{2M}{3k}\Big)}. \end{aligned}$$

As a result,

$$I_3 \leq \delta^2\sqrt{\ln\Big(\frac{2M}{3k}\Big)} + \frac{\ell\delta}{M}\sqrt{\pi}.$$

Thus,

$$\begin{aligned}
&\int_0^\infty \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon \\
&\quad \leq 2\delta\sqrt{\delta\ell \ln 2} + \sqrt{d}\delta^2(\sqrt{\ln 2} + \sqrt{\pi}) + \frac{\ell\delta}{M}\sqrt{kd\pi} + \delta^2\sqrt{kd\ln\left(\frac{2M}{3k}\right)} + \delta^2 A(k,d) \\
&\quad = 2\delta\sqrt{\delta\ell \ln 2} + \frac{\ell\delta}{M}\sqrt{kd\pi} + a_0 \\
&\qquad + \sqrt{k}\delta^2\left[d\ln\left(\frac{2M}{3k}\right) + d\ln 3 + \frac{d}{2}\ln d + \ln V_d + 3\ln 2\right]^{1/2},
\end{aligned}$$

where a_0 is a nonnegative constant. Finally,

$$\int_0^\infty \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon \leq a_1\sqrt{k} + a_2\sqrt{\ell} + a_3\frac{\ell}{\sqrt{k}} + a_0,$$

where the nonnegative constants $a_0, \ldots, a_3$ only depend on the maximal length L, the dimension d and the diameter δ of the convex set $\mathcal{C}$.

F.5.3. Proof of Proposition F.3.1

Let

$$S_{k,\kappa} = \{\Delta(\mathbf{f}, \cdot), \mathbf{f} \in \mathcal{F}_{k,\kappa}\}$$

be a subset of the continuous functions from $\mathcal{C}$ to $\mathbb{R}^+$, endowed with the sup-norm $\|\cdot\|_\infty$. Starting as in the proof of Proposition F.2.1, we know that, for all $\mathbf{X}_1, \ldots, \mathbf{X}_n$,

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\kappa}}\Big(\mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})] - \Delta_n(\mathbf{f})\Big)\right] \leq \frac{c}{\sqrt{n}}\int_0^\infty \sqrt{\ln \mathcal{N}(S_{k,\kappa}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon,$$

for some absolute constant $c > 0$. Now, according to Lemma 5 in Sandilya and Kulkarni [37], we may write, for each $\varepsilon > 0$,

$$\begin{aligned}
&\ln \mathcal{N}(S_{k,\kappa}, \|\cdot\|_\infty, \varepsilon) \\
&\quad \le \left(\frac{\zeta(\kappa)\delta^2}{\varepsilon} + 2k + 1\right) \ln 2 + (k+1) \ln V_d + d \ln \left(\frac{\delta^2\sqrt{d}}{\varepsilon} + \sqrt{d}\right) \\
&\qquad + kd \ln \left(\frac{\zeta(\kappa)\delta^2\sqrt{d}}{k\varepsilon} + 3\sqrt{d}\right) \\
&\quad = \frac{\zeta(k)\delta^2}{\varepsilon} \ln 2 + (2k+1) \ln 2 + (k+1) \ln V_d + d(k+1) \ln \sqrt{d} + d \ln \left(\frac{\delta^2}{\varepsilon} + 1\right) \\
&\qquad + kd \ln 3 + kd \ln \left(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\right) \\
&\quad = \frac{\zeta(k)\delta^2}{\varepsilon} \ln 2 + d \ln \left(\frac{\delta^2}{\varepsilon} + 1\right) + kd \ln \left(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\right) + kd \ln 3 + (2k+1) \ln 2 \\
&\qquad + (k+1)(\ln V_d + \frac{d}{2} \ln d).
\end{aligned}$$

Consequently,

$$\begin{aligned}
\int_0^\infty \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)} \mathrm{d}\varepsilon &= \int_0^{\delta^2} \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)} \mathrm{d}\varepsilon \\
&\le \delta \sqrt{\zeta(\kappa) \ln 2} I_1 + \sqrt{d} I_2 + \sqrt{kd} I_3 + \delta^2 A(k,d),
\end{aligned}$$

where $I_1 = \displaystyle\int_0^{\delta^2} \frac{1}{\sqrt{\varepsilon}} \mathrm{d}\varepsilon$, $I_2 = \displaystyle\int_0^{\delta^2} \sqrt{\ln\left(\frac{\delta^2}{\varepsilon} + 1\right)} \mathrm{d}\varepsilon$, $I_3 = \displaystyle\int_0^{\delta^2} \sqrt{\ln\left(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\right)} \mathrm{d}\varepsilon$, and

$$A(k,d) = \delta^2 \Big[kd \ln 3 + (2k+1) \ln 2 + (k+1)(\ln V_d + \frac{d}{2} \ln d) \Big]^{1/2}.$$

Control of $\mathbf{I_1}$. We clearly have

$$I_1 = \int_0^{\delta^2} \frac{1}{\sqrt{\varepsilon}} \mathrm{d}\varepsilon = 2\delta.$$

Control of $\mathbf{I_2}$. We have

$$\begin{aligned}
I_2 &\le \int_0^{\delta^2} \sqrt{\ln\left(\frac{2\delta^2}{\varepsilon}\right)} \mathrm{d}\varepsilon \\
&= 2\delta^2 \int_0^{1/2} \sqrt{\ln \frac{1}{u}} \mathrm{d}u \\
&\le \delta^2 (\sqrt{\ln 2} + \sqrt{\pi}).
\end{aligned}$$

Control of $\mathbf{I_3}$. Assume first that $\dfrac{\zeta(\kappa)}{3k} \geq 1$. Then

$$\begin{aligned} I_3 &= \int_0^{\delta^2} \sqrt{\ln\Big(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon}+1\Big)}\mathrm{d}\varepsilon \\ &\leq \int_0^{\delta^2} \sqrt{\ln\Big(\frac{2\zeta(\kappa)\delta^2}{3k\varepsilon}\Big)}\mathrm{d}\varepsilon \\ &= \frac{2\zeta(\kappa)\delta^2}{3k}\int_0^{3k/2\zeta(\kappa)} \sqrt{\ln\frac{1}{u}}\mathrm{d}u \\ &\leq \delta^2\bigg(\sqrt{\ln\frac{2\zeta(\kappa)}{3k}}+\sqrt{\pi}\bigg). \end{aligned}$$

On the other hand, if $\dfrac{\zeta(\kappa)}{3k} < 1$, we cut up I_3 into two pieces and write

$$\begin{aligned} I_3 &= \int_0^{\delta^2} \sqrt{\ln\Big(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon}+1\Big)}\mathrm{d}\varepsilon \\ &= \int_0^{\zeta(\kappa)\delta^2/3k} \sqrt{\ln\Big(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon}+1\Big)}\mathrm{d}\varepsilon + \int_{\zeta(\kappa)\delta^2/3k}^{\delta^2} \sqrt{\ln\Big(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon}+1\Big)}\mathrm{d}\varepsilon. \end{aligned} \tag{F.8}$$

The first integral is bounded by using the inequality $\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} \geq 1$ for all $\varepsilon \in]0, \frac{\zeta(\kappa)\delta^2}{3k}]$. We obtain

$$\begin{aligned} \int_0^{\zeta(\kappa)\delta^2/3k} \sqrt{\ln\Big(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon}+1\Big)}\mathrm{d}\varepsilon &\leq \int_0^{\zeta(\kappa)\delta^2/3k} \sqrt{\ln\Big(\frac{2\zeta(\kappa)\delta^2}{3k\varepsilon}\Big)}\mathrm{d}\varepsilon \\ &= \frac{2\zeta(\kappa)\delta^2}{3k}\int_0^{1/2} \sqrt{\ln\frac{1}{u}}\mathrm{d}u \\ &\leq \frac{\zeta(\kappa)\delta^2}{3k}(\sqrt{\ln 2}+\sqrt{\pi}). \end{aligned}$$

With respect to the second integral in (F.8), we note that the function under the integral is decreasing in ε, so that

$$\int_{\zeta(\kappa)\delta^2/3k}^{\delta^2} \sqrt{\ln\Big(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon}+1\Big)}\mathrm{d}\varepsilon \leq \Big(\delta^2 - \frac{\zeta(\kappa)\delta^2}{3k}\Big)\sqrt{\ln 2}.$$

Thus, we have

$$I_3 \leq \begin{cases} \delta^2\bigg(\sqrt{\ln\dfrac{\zeta(\kappa)}{3k}}+\sqrt{\pi}+\sqrt{\ln 2}\bigg) & \text{if} \quad \dfrac{\zeta(\kappa)}{3k} \geq 1 \\ \delta^2\bigg(\dfrac{\zeta(\kappa)}{3k}\sqrt{\pi}+\sqrt{\ln 2}\bigg) & \text{if} \quad \dfrac{\zeta(\kappa)}{3k} < 1. \end{cases}$$

Hence, collecting the different results,

$$\begin{aligned}
&\int_0^\infty \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon \\
&\leq 2\delta^2\sqrt{\zeta(\kappa)\ln 2} + \sqrt{d}\delta^2(\sqrt{\ln 2}+\sqrt{\pi}) + \delta^2\sqrt{kd}\left(\sqrt{\ln\frac{\zeta(\kappa)}{3k}}+\sqrt{\pi}+\sqrt{\ln 2}\right)\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}\geq 1\}} \\
&\quad + \delta^2\sqrt{kd}\left(\frac{\zeta(\kappa)}{3k}\sqrt{\pi}+\sqrt{\ln 2}\right)\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}<1\}} + \delta^2 A(k,d) \\
&\leq \delta^2\Big(2\sqrt{\zeta(\kappa)\ln 2} + \frac{\zeta(\kappa)}{3\sqrt{k}}\sqrt{\pi d}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}<1\}} + \sqrt{kd\ln\frac{\zeta(\kappa)}{3k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}\geq 1\}} \\
&\quad + \sqrt{k}\Big[\sqrt{d}(\sqrt{\pi}+\sqrt{\ln 2}) + \Big(d\ln 3 + \frac{d}{2}\ln d + \ln V_d + 2\ln 2\Big)^{1/2} + a_0\Big]\Big),
\end{aligned}$$

where a_0 is a nonnegative constant. Finally,

$$\begin{aligned}
&\int_0^\infty \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon \\
&\leq \delta^2\Big(a_1\sqrt{k} + a_2\sqrt{\zeta(\kappa)} + a_3\frac{\zeta(\kappa)}{\sqrt{k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}<1\}} + a_4\sqrt{k\ln\frac{\zeta(\kappa)}{k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}\geq 1\}} + a_0\Big),
\end{aligned}$$

where the nonnegative constants $a_0, \ldots, a_4$ only depend on the dimension d.

References

[1] T. M. Alcorn and C. W. Hoggar. Preprocessing of data for character recognition. *Marconi Review*, pages 61–81, 1969.

[2] A. D. Alexandrov and Y. G. Reshetnyak. *General Theory of Irregular Curves*. Mathematics and its Applications. Kluwer Academic Publishers, Dordrecht, 1989.

[3] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279, 2009.

[4] J. D. Banfield and A. E. Raftery. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *Journal of the American Statistical Association*, 87:7–16, 1992.

[5] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

[6] P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2001.

[7] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 2011. In press. Available at `http://hal.archives-ouvertes.fr/docs/00/46/16/39/PDF/RR-7223.pdf`.

[8] L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 55–87. Springer, New York, 1997.

[9] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138:33–73, 2007.

[10] C. Brunsdon. Path estimation from GPS tracks. In *Proceedings of the 9th International Conference on GeoComputation, National Centre for Geocomputation, National University of Ireland, Maynooth, Eire*, 2007.

[11] B. S. Caffo, C. M. Crainiceanu, L. Deng, and C. W. Hendrix. A case study in pharmacologic colon imaging using principal curves in single photon emission computed tomography. *Journal of the American Statistical Association*, 103:1470–1480, 2008.

[12] K. Chang and J. Ghosh. Principal curves for nonlinear feature extraction and classification. *SPIE Applications of Artificial Neural Networks in Image Processing III*, 3307:120–129, 1998.

[13] P. J. Corkeron, P. Anthony, and R. Martin. Ranging and diving behaviour of two 'offshore' bottlenose dolphins, Tursiops sp., off eastern Australia. *Journal of the Marine Biological Association of the United Kingdom*, 84:465–468, 2004.

[14] G. De'ath. Principal curves: a new technique for indirect and direct gradient analysis. *Ecology*, 80:2237–2253, 1999.

[15] P. Delicado. Another look at principal curves and surfaces. *Journal of Multivariate Analysis*, 77:84–116, 2001.

[16] E. S. Deutsch. Preprocessing for character recognition. In *Proceedings of the IEE–NPL Conference on Pattern Recognition*, pages 179–190, 1968.

[17] R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1:290–330, 1967.

[18] R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 1999.

[19] J. Einbeck, G. Tutz, and L. Evers. Exploring multivariate data structures with local principal curves. In C. Weihs and W. Gaul, editors, *Classification – The Ubiquitous Challenge, Proceedings of the 28th Annual Conference of the Gesellschaft für Klassifikation, University of Dortmund*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 256–263. Springer, Berlin, Heidelberg, 2005.

[20] J. Einbeck, G. Tutz, and L. Evers. Local principal curves. *Statistics and Computing*, 15:301–313, 2005.

[21] E. R. Engdahl and A. Villaseñor. Global seismicity: 1900–1999. In W.H.K. Lee, H. Kanamori, P.C. Jennings, and C. Kisslinger, editors, *International Handbook of Earthquake and Engineering Seismology*, pages 665–690. Academic Press, London, 2002.

[22] H. Friedsam and W. A. Oren. The application of the principal curve analysis technique to smooth beamlines. In *Proceedings of the 1st International Workshop on Accelerator Alignment*, 1989.

[23] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. The geometry of nonparametric filament estimation. 2010. Available at `http://arxiv.org/PS_cache/arxiv/pdf/1003/1003.5536v2.pdf`.

[24] D. J. Harding and G. S. Berghoff. Fault scarp detection beneath dense vegetation cover: airborne lidar mapping of the Seattle fault zone, Bainbridge Island, Washington State. In *Proceedings of the American Society of Photogrammetry and Remote Sensing Annual Conference*, 2000.

[25] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

[26] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 498–520, 1933.

[27] B. Kégl. *Principal Curves: Learning, Design, and Applications*. PhD thesis, Concordia University, Montréal, Québec, Canada, 1999.

[28] B. Kégl and A. Krzyżak. Piecewise linear skeletonization using principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:59–74, 2002.

[29] B. Kégl, A. Krzyżak, T. Linder, and K. Zeger. Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:281–297, 2000.

[30] A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis*. Dover Publications, Mineola, 1975.

[31] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47:1902–1914, 2001.

[32] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979.

[33] P. Massart. *Concentration Inequalities and Model Selection*. Ecole d'Eté de Probabilités de Saint-Flour XXXIII – 2003, Lecture Notes in Mathematics. Springer, Berlin, Heidelberg, 2007.

[34] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press, Cambridge, 1989.

[35] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[36] K. Reinhard and M. Niranjan. Parametric subspace modeling of speech transitions. *Speech Communication*, 27:19–42, 1999.

[37] S. Sandilya and S. R. Kulkarni. Principal curves with bounded turn. *IEEE Transactions on Information Theory*, 48:2789–2793, 2002.

[38] C. Spearman. General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.

[39] D. C. Stanford and A. E. Raftery. Finding curvilinear features in spatial point patterns: principal curve clustering with noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:2237–2253, 2000.

[40] R. Tibshirani. Principal curves revisited. *Statistics and Computing*, 2:183–190, 1992.

[41] J. J. Verbeek, N. Vlassis, and B. Kröse. A soft k-segments algorithm for principal curves. In *Proceedings of International Conference on Artificial Neural Networks 2001*, pages 450–456, 2001.

[42] W. C. K. Wong and A. C. S. Chung. Principal curves to extract vessels in 3D angiograms. In *Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'08)*, pages 1–8, 2008.

Bibliographie

[1] C. Abraham, P. A. Cornillon, E. Matzner-Løber et N. Molinari : Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics*, 30:581–595, 2003.

[2] H. Akaike : Information theory and an extension of the maximum likelihood principle. *In Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281, 1973.

[3] B. Al-Shboul et S.-H. Myaeng : Initializing k-means using genetic algorithms. *World Academy of Science, Engineering and Technology*, 54:114–118, 2009.

[4] Y. Alber et D. Butnariu : Convergence of Bregman projection methods for solving consistent convex feasibility problems in reflexive Banach spaces. *Journal of Optimization Theory and Applications*, 92:33–61, 1997.

[5] R. J. Alcock et Y. Manolopoulos : Time-series similarity queries employing a feature-based approach. *In 7th Hellenic Conference on Informatics*, Ioannina, Greece, 1999.

[6] T. M. Alcorn et C. W. Hoggar : Preprocessing of data for character recognition. *Marconi Review*, pages 61–81, 1969.

[7] A. D. Alexandrov et Y. G. Reshetnyak : *General Theory of Irregular Curves*. Mathematics and its Applications. Kluwer Academic Publishers, Dordrecht, 1989.

[8] A. Antoniadis, X. Brossat, J. Cugliari et J.-M. Poggi : Clustering functional data with wavelets. *In* Y. Lechevallier et G. Saporta, éditeurs : *Proceedings of the 19th International Conference on Computational Statistics, COMPSTAT 2010*, pages 697–704. Springer, 2010.

[9] W. Arendt, J. K. Batty, M. Hieber et F. Neubrander : *Vector-valued Laplace Transforms and Cauchy Problems*. Monographs in Mathematics. Birkhäuser, Basel, 2001.

[10] S. Arlot et P. Massart : Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279, 2009.

[11] N. Aronszajn : Theory of reproducing kernels. *Transactions of American Mathematical Society*, 68:337–404, 1950.

[12] D. Art, R. Gnanadesikan et J. R. Kettenring : Data-based metrics for cluster analysis. *Utilitas Mathematica*, 21A:75–99, 1982.

[13] B. Auder, A. De Crecy, B. Iooss et M. Marquès : Screening and metamodeling of computer experiments with functional outputs. Application to thermal-hydraulic computations. 2010. Available at `http://hal.archives-ouvertes.fr/docs/00/52/54/91/PDF/ress_samo10_BA.pdf`.

[14] K. S. Azoury et M. K. Warmuth : Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43:211–246, 2001.

[15] A. Banerjee, X. Guo et H. Wang : Optimal Bregman prediction and Jensen's equality. *In Proceedings of the International Symposium on Information Theory (ISIT), Chicago*, volume 169, 2004.

[16] A. Banerjee, X. Guo et H. Wang : On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51, 2005.

[17] A. Banerjee, S. Merugu, I. S. Dhillon et J. Ghosh : Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[18] J. D. Banfield et A. E. Raftery : Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *Journal of the American Statistical Association*, 87:7–16, 1992.

[19] O. Barndorff-Nielsen : *Information and Exponential Families in Statistical Theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, Chichester, 1979.

[20] E. del Barrio, P. Deheuvels et S. van de Geer : *Lectures on Empirical Processes*. European Mathematical Society, Zürich, 2007.

[21] A. Barron, L. Birgé et P. Massart : Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

[22] P. L. Bartlett, S. Boucheron et G. Lugosi : Model selection and error estimation. *Machine Learning*, 48:85–113, 2001.

[23] P. L. Bartlett, T. Linder et G. Lugosi : The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, 44:1802–1813, 1998.

[24] M. Basseville : Divergence measures for statistical data processing. Rapport technique, IRISA, 2010. Available at `http://hal.inria.fr/docs/00/54/23/37/PDF/PI-1961.pdf`.

[25] J.-P. Baudry, C. Maugis et B. Michel : Slope heuristics: overview and implementation. *Statistics and Computing*, 22:455–470, 2012.

[26] H. H. Bauschke, J. M. Borwein et P. L. Combettes : Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces. *Communications in Contemporary Mathematics*, 3:615–647, 2001.

[27] S. Ben-David, D. Pál et H. U. Simon : Stability of k-means clustering. *In* N. Bshouty et C. Gentile, éditeurs : *Proceedings of the 20th Annual Conference on Learning Theory (COLT 2007)*, pages 20–34. Springer, 2007.

[28] S. Ben-David et U. von Luxburg : Relating clustering stability to properties of cluster boundaries. *In* R. A. Servedio et T. Zhang, éditeurs : *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pages 379–390, Madison, 2008. Omnipress.

[29] S. Ben-David, U. von Luxburg et D. Pál : A sober look on clustering stability. *In* G. Lugosi et H. U. Simon, éditeurs : *Proceedings of the 19th Annual Conference on Learning Theory (COLT 2006)*, pages 5–19, Berlin, 2006. Springer.

[30] A. Ben-Hur, A. Elisseeff et I. Guyon : A stability based method for discovering structure in clustered data. *In Proceedings of the 7th Pacific Symposium on Biocomputing*, volume 7, pages 6–17, 2002.

[31] A. T. Bharucha-Reid : *Random Integral Equations*, volume 96 de *Mathematics in Science and Engineering*. Academic Press, New York, 1972.

[32] G. Biau, L. Devroye et G. Lugosi : On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54:781–790, 2008.

[33] C. Biernacki, G. Celeux et G. Govaert : Assessing a mixture model for clustering with the integrated classification likelihood. Rapport technique 3521, INRIA, 1998.

[34] P. Billingsley : *Convergence of Probability Measures*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, 1999.

[35] L. Birgé et P. Massart : From model selection to adaptive estimation. *In* D. Pollard, E. Torgersen et G. Yang, éditeurs : *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 55–87. Springer, New York, 1997.

[36] L. Birgé et P. Massart : Gaussian model selection. *Journal of the European Mathematical Society*, 3:203–268, 2001.

[37] L. Birgé et P. Massart : Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138:33–73, 2007.

[38] J. Borwein et R. Goebel : Notions of relative interior in banach spaces. *Journal of Mathematical Sciences*, 115:2542–2553, 2003.

[39] S. Boucheron, O. Bousquet et G. Lugosi : Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[40] L. M. Bregman : The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[41] H. Brezis : *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer, New York, 2010.

[42] J. Bruna et S. Mallat : Classification with scattering operators. 2010. Available at `http://arxiv.org/PS_cache/arxiv/pdf/1011/1011.3023v3.pdf`.

[43] C. Brunsdon : Path estimation from GPS tracks. *In Proceedings of the 9th International Conference on GeoComputation, National Centre for Geocomputation, National University of Ireland, Maynooth, Eire*, 2007.

[44] B. Cadre et Q. Paris : On hölder fields clustering. *Test*, 21:301–316, 2012.

[45] B. S. Caffo, C. M. Crainiceanu, L. Deng et C. W. Hendrix : A case study in pharmacologic colon imaging using principal curves in single photon emission computed tomography. *Journal of the American Statistical Association*, 103:1470–1480, 2008.

[46] C. Caillerie et B. Michel : Model selection for simplicial approximation. *Foundations of Computational Mathematics*, 11:707–731, 2011.

[47] R. B. Calinski et J. Harabasz : A dendrite method for cluster analysis. *Communications in Statistics*, 3:1–27, 1974.

[48] H. Cartan : *Cours de Calcul Différentiel*. Méthodes. Hermann, Paris, 1977.

[49] N. Cesa-Bianchi et G. Lugosi : *Prediction, Learning, and Games*. Cambridge University Press, New York, 2006.

[50] K. Chang et J. Ghosh : Principal curves for nonlinear feature extraction and classification. *SPIE Applications of Artificial Neural Networks in Image Processing III*, 3307:120–129, 1998.

[51] J. M. Chiou et P. L. Li : Functional clustering of longitudinal data. *In* S. Dabo-Niang et F. Ferraty, éditeurs : *Functional and Operatorial Statistics*, Physica-Verlag, pages 103–107. Springer, 2008.

[52] W. S. Cleveland : Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

[53] A. Cohen, I. Daubechies et P. Vial : Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis*, 1:54–81, 1993.

[54] R. R. Coifman et M. V. Wickerhauser : Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38:713–718, 1992.

[55] P. J. Corkeron, P. Anthony et R. Martin : Ranging and diving behaviour of two 'offshore' bottlenose dolphins, Tursiops sp., off eastern Australia. *Journal of the Marine Biological Association of the United Kingdom*, 84:465–468, 2004.

[56] I. Csiszár : Generalized projections for non-negative functions. *Acta Mathematica Hungarica*, 68:161–185, 1995.

[57] I. Csiszár, F. Gamboa et E. Gassiat : MEM pixel correlated solutions for generalized moment and interpolation problems. *IEEE Transactions on Information Theory*, 45, 1999.

[58] I. Csiszár et F. Matús : Generalized maximum likelihood estimates for infinite dimensional exponential families. *In Prague Stochastics 2006*, pages 288–297, 2006.

[59] I. Csiszár et F. Matús : On minimization of entropy functionals under moment constraints. *In Proceedings of the International Symposium on Information Theory (ISIT 2008), Toronto, Canada*, pages 2101–2105, 2008.

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

[61] I. Daubechies : *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[62] G. De'ath : Principal curves: a new technique for indirect and direct gradient analysis. *Ecology*, 80:2237–2253, 1999.

[63] P. Deheuvels : Strong bounds for multidimensional spacings. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 64:411–424, 1983.

[64] P. Delicado : Another look at principal curves and surfaces. *Journal of Multivariate Analysis*, 77:84–116, 2001.

[65] P. Delicado et M. Huerta : Principal curves of oriented points: theoretical and computational improvements. *Computational Statistics*, 18:293–315, 2003.

[66] A. P. Dempster, N. M. Laird et D. B. Rubin : Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, pages 1–38, 1977.

[67] M. B. Denckla et R. G. Rudel : Rapid "automatized" naming (R.A.N.): dyslexia differentiated from other learning disabilities. *Neuropsychologia*, 14:471–479, 1976.

[68] S. Dereich et C. Vormoor : The high resolution vector quantization problem with Orlicz norm distortion. *Journal of Theoretical Probability*, 24:517–544, 2010.

[69] E. S. Deutsch : Preprocessing for character recognition. *In Proceedings of the IEE–NPL Conference on Pattern Recognition*, pages 179–190, 1968.

[70] R. A. DeVore et G. G. Lorentz : *Constructive Approximation*. Springer-Verlag, Berlin, Heidelberg, 1993.

[71] L. Devroye, L. Györfi et G. Lugosi : *A Probabilistic Theory of Pattern Recognition*. Applications of Mathematics. Springer, New York, 1996.

[72] D. Dong et T. J. McAvoy : Nonlinear principal component analysis-based principal curves and neural networks. *Computers and Chemical Engineering*, 20:65–78, 1995.

[73] D. L. Donoho et I. M. Johnstone : Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[74] D. L. Donoho et I. M. Johnstone : Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26:879–921, 1998.

[75] T. Duchamp et W. Stuetzle : Geometric properties of principal curves in the plane. *In* H. Rieder, éditeur : *Robust Statistics, Data Analysis, and Computer Intensive Methods: in Honor of Peter Huber's 60th Birthday*, volume 109 de *Lecture Notes in Statistics*, pages 135–152. Springer-Verlag, New York, 1996.

[76] R. O. Duda, P. E. Hart et D. G. Stork : *Pattern Classification*. Wiley-Interscience, New York, 2000.

[77] R. M. Dudley : The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1:290–330, 1967.

[78] R. M. Dudley : *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 1999.

[79] R. M. Dudley : *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2002.

[80] J. Einbeck et J. Dwyer : Using principal curves to analyse traffic patterns on freeways. *Transportmetrica*, 2010.

[81] J. Einbeck, G. Tutz et L. Evers : Exploring multivariate data structures with local principal curves. *In* C. Weihs et W. Gaul, éditeurs : *Classification – The Ubiquitous Challenge, Proceedings of the 28th Annual Conference*

of the Gesellschaft für Klassifikation, University of Dortmund, Studies in Classification, Data Analysis, and Knowledge Organization, pages 256–263. Springer, Berlin, Heidelberg, 2005.

[82] J. Einbeck, G. Tutz et L. Evers : Local principal curves. *Statistics and Computing*, 15:301–313, 2005.

[83] E. R. Engdahl et A. Villaseñor : Global seismicity: 1900–1999. *In* W.H.K. Lee, H. Kanamori, P.C. Jennings et C. Kisslinger, éditeurs : *International Handbook of Earthquake and Engineering Seismology*, pages 665–690. Academic Press, London, 2002.

[84] F. Ferraty et P. Vieu : *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. Springer-Verlag, New York, 2006.

[85] A. Frank et A. Asuncion : UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2010. `http://archive.ics.uci.edu/ml`.

[86] J. H. Friedman, M. Jacobson et W. Stuetzle : Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–846, 1981.

[87] H. Friedsam et W. A. Oren : The application of the principal curve analysis technique to smooth beamlines. *In Proceedings of the 1st International Workshop on Accelerator Alignment*, 1989.

[88] B. A. Frigyik, S. Srivastava et M. R. Gupta : An introduction to functional derivatives. Rapport technique UWEETR-2008-0001, Department of Electrical Engineering, University of Washington, Seattle, 2008.

[89] B. A. Frigyik, S. Srivastava et M. R. Gupta : Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 54:5130–5139, 2008.

[90] S. Gaffney : *Probabilistic Curve-Aligned Clustering and Prediction with Mixture Models*. Thèse de doctorat, Department of Computer Science, University of California, Irvine, 2004.

[91] F. Gamboa et E. Gassiat : Bayesian methods and maximum entropy for ill-posed inverse problems. *The Annals of Statistics*, 25:328–350, 1997.

[92] F. Gamboa, J.-M. Loubes et P. Rochet : Maximum entropy estimation for survey sampling. *Journal of Statistical Planning and Inference*, 141:305–317, 2011.

[93] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli et L. Wasserman : The geometry of nonparametric filament estimation. *Journal of the American Statistical Association*, 107:788–799, 2012.

[94] A. Gersho et R. M. Gray : *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Norwell, 1992.

[95] R. G. Ghanem et P. D. Spanos : *Stochastic Finite Elements: A Spectral Approach*. Springer-Verlag, New York, 1991.

[96] E. Giné et J. Zinn : Some limit theorems for empirical processes. *Annals of Probability*, 12:929–989, 1984.

[97] A. D. Gordon : *Classification*, volume 82 de *Monographs on Statistics and Applied Probability*. Chapman Hall/CRC, Boca Raton, 1999.

[98] S. Graf et H. Luschgy : *Foundations of Quantization for Probability Distributions*. Lecture Notes in Mathematics. Springer-Verlag, Berlin, Heidelberg, 2000.

[99] R. M. Gray, A. Buzo, A. H. Gray et Y. Matsuyama : Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28:367–376, 1980.

[100] D. J. Harding et G. S. Berghoff : Fault scarp detection beneath dense vegetation cover: airborne lidar mapping of the Seattle fault zone, Bainbridge Island, Washington State. *In Proceedings of the American Society of Photogrammetry and Remote Sensing Annual Conference*, 2000.

[101] A. Hardy : On the number of clusters. *Computational Statistics and Data Analysis*, 23:83–96, 1996.

[102] J. A. Hartigan : *Clustering Algorithms*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, 1975.

[103] T. Hastie : Principal curves and surfaces. Rapport technique, Stanford Linear Accelerator Center, 1984.

[104] T. Hastie et W. Stuetzle : Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

[105] T. Hastie, R. Tibshirani et J. Friedman : *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, 2001.

[106] N. E. Heckman et R. H. Zamar : Comparing the shapes of regression functions. *Biometrika*, 87:135–144, 2000.

[107] T. Hermann, P. Meinicke et H. Ritter : Principle curve sonification. *In Proceedings of the 6th International Conference on Auditory Display Curve Sonification (ICAD2000), Atlanta, USA*, pages 81–86, 2000.

[108] F. Hirsch et G. Lacombe : *Eléments d'Analyse Fonctionelle*. Dunod, Paris, 2003.

[109] H. Hotelling : Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 498–520, 1933.

[110] G. M. James et C. A. Sugar : Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98:397–408, 2003.

[111] L. Jones et C. Byrne : General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis. *IEEE Transactions on Information Theory*, 36, 1990.

[112] M. I. Jordan, X. Nguyen et M. J. Wainwright : Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56:5847–5861, 2010.

[113] L. Kaufman et P. Rousseeuw : *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Mathematical Statistics. Wiley-Interscience, Hoboken, 1990.

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

[115] B. Kégl : *Principal Curves: Learning, Design, and Applications*. Thèse de doctorat, Concordia University, Montréal, Québec, Canada, 1999.

[116] B. Kégl et A. Krzyżak : Piecewise linear skeletonization using principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:59–74, 2002.

[117] B. Kégl, A. Krzyżak, T. Linder et K. Zeger : Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:281–297, 2000.

[118] S. S. Khan et A. Ahmad : Cluster center initialization algorithm for k-means clustering. *Pattern Recognition Letters*, 25:1293–1302, 2004.

[119] D. J. Kim, Y. W. Park et D. J. Park : A novel validity index for determination of the optimal number of clusters. *IEICE Transactions on Information and System*, E84D:281–285, 2001.

[120] J. Kogan, M. Teboulle, P. Berkhin, I. Dhillon et Y. Guan : Clustering with entropy-like k-means algorithms. *In Grouping Multidimensional Data: Recent Advances in Clustering*, chapitre 5, pages 127–160. Springer, Berlin, Heidelberg, 2006.

[121] A. N. Kolmogorov et S. V. Fomin : *Introductory Real Analysis*. Dover Publications, Mineola, 1975.

[122] V. Koltchinskii : Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47:1902–1914, 2001.

[123] J. Kruskal : On the shortest spanning subtree of a graph and the traveling salesman problem. *In Proceedings of the American Mathematical Society*, volume 7, pages 48–50, 1956.

[124] W. J. Krzanowski et Y. T. Lai : A criterion for determining the number of clusters in a data set. *Biometrics*, 44:23–34, 1985.

[125] T. Laloë : L1-quantization and clustering in Banach spaces. *Mathematical Methods of Statistics*, 19:136–150, 2010.

[126] M. Ledoux et M. Talagrand : *Probability in Banach Spaces*. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer, Berlin, Heidelberg, 1991.

[127] E. L. Lehmann et G. Casella : *Theory of Point Estimation*. Springer-Verlag, New York, 1998.

[128] E. Levine et E. Domany : Resampling method for unsupervised estimation of cluster validity. *Journal of Neural Computation*, 13:2573–2593, 2002.

[129] Y Linde, A Buzo et R M Gray : An algorithm for vector quantizer design. *IEEE Transactions on Communication*, 28:801–804, 1980.

[130] T. Linder : On the training distortion of vector quantizers. *IEEE Transactions on Information Theory*, 46:1617–1623, 2000.

[131] T. Linder : Learning-theoretic methods in vector quantization. *In* L. Györfi, éditeur : *Principles of Nonparametric Learning*. Springer-Verlag, Wien, 2002.

[132] S. P. Lloyd : Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[133] W. L. Loh : On latin hypercube sampling. *The Annals of Statistics*, 24:2058–2080, 1996.

[134] S. Lojasiewicz : *An Introduction to the Theory of Real Functions*. John Wiley and Sons, New York, 1988.

[135] H. Luschgy et G. Pagès : Functional quantization rate and mean regularity of processes with an application to Levy processes. *The Annals of Applied Probability*, 18:427–469, 2008.

[136] H. Luschgy et P. Pagès : Functional quantization of gaussian processes. *Journal of Functional Analysis*, 196:486–531, 2002.

[137] H. Luschgy et P. Pagès : Functional quantization of a class of brownian diffusions: a constructive approach. *Stochastic Processes and their Applications*, 116:310–336, 2006.

[138] S. Mallat : *A Wavelet Tour of Signal Processing, The Sparse Way*. Academic Press, San Diego, 2008.

[139] C. L. Mallows : Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[140] K. V. Mardia, J. T. Kent et J. M. Bibby : *Multivariate Analysis*. Academic Press, London, 1979.

[141] P. Massart : *Concentration Inequalities and Model Selection*. Ecole d'Eté de Probabilités de Saint-Flour XXXIII – 2003, Lecture Notes in Mathematics. Springer, Berlin, Heidelberg, 2007.

[142] P. Massart et E. Nédélec : Risk bounds for statistical learning. *Annals of Statistics*, 34:2326–2366, 2006.

[143] C. McDiarmid : On the method of bounded differences. *In Surveys in Combinatorics*, pages 148–188. Cambridge University Press, Cambridge, 1989.

[144] M. D. McKay, W. J. Conover et R. J. Beckman : A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–245, 1979.

[145] N. Merhav et J. Ziv : On the amount of statistical side information required for lossy data compression. *IEEE Transactions on Information Theory*, 43: 1112–1121, 1997.

[146] Y. Meyer : *Wavelet and Operators*. Cambridge University Press, Cambridge, 1992.

[147] G. W. Milligan et M. C. Cooper : An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–79, 1985.

[148] W. J. Nash, T. L. Sellers, S. R. Talbot, A. J. Cawthorn et W. B Ford : The population biology of Abalone (Haliotis species) in Tasmania. 1, Blacklip Abalone (H. rubra) from the north coast and islands of Bass Strait. Rapport technique 48, Sea Fisheries Division, 1994.

[149] F. Nielsen, J.D. Boissonnat et R Nock : Bregman Voronoi diagrams: properties, algorithms and applications. Rapport technique 6154, INRIA, 2007.

[150] U. Ozertem et D. Erdogmus : Signal denoising using principal curves: application to timewarping. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)*, pages 3709–3712, 2008.

[151] K. Pearson : On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[152] J. M. Pena, J. A. Lozano et P. Larranaga : An empirical comparison of four initialization methods for the K-means algorithm. *Pattern Recognition Letters*, 20:1027–1040, 1999.

[153] G. T. Perim, E. D. Wandekokem et F. M. Varejão : K-means initialization methods for improving clustering by simulated annealing. *In Advances in artificial intelligence – Iberamia 2008*, volume 5290, pages 133–142. Springer-Verlag, Berlin, Heidelberg, 2008.

[154] R. R. Phelps : *Convex Functions, Monotone Operators, and Differentiability*. Springer-Verlag, Berlin, Heidelberg, 1993.

[155] J. N. Pierce : Asymptotic quantizing error for unbounded random variables. *IEEE Transactions on Information Theory*, 16:81–83, 1970.

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

[157] D. Pollard : Quantization and the method of k-means. *IEEE Transactions on Information Theory*, 28, 1982.

[158] R. C. Prim : Shortest connection networks and some generalizations. *Bell System Technology Journal*, 36:1389–1401, 1957.

[159] J. O. Ramsay et B. W. Silverman : *Functional Data Analysis*. Springer, New York, 2006.

[160] K. Reinhard et M. Niranjan : Parametric subspace modeling of speech transitions. *Speech Communication*, 27:19–42, 1999.

[161] R. T. Rockafellar : *Convex Analysis*. Princeton University Press, Princeton, New Jersey, 1970.

[162] F. Rossi, B. Conan-Guez et A. El Golli : Clustering functional data with the SOM algorithm. *In Proceedings of the European Symposium on Artificial Neural Networks (ESANN'2004)*, pages 305–312, Bruges, 2004.

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

[164] F. Rouvière : *Petit Guide de Calcul Différentiel*. Cassini, Paris, 2003.

[165] M. J. Sabin et R. M. Gray : Global convergence and empirical consistency of the Generalized Lloyd Algorithm. *IEEE Transactions on Information Theory*, 32:148–155, 1986.

[166] S. Sandilya et S. R. Kulkarni : Principal curves with bounded turn. *IEEE Transactions on Information Theory*, 48:2789–2793, 2002.

[167] G. Schwarz : Estimating the dimension of a model. *Annals of Statistics*, 6:33–73, 1978.

[168] O. Shamir et N. Tishby : Cluster stability for finite samples. *In* J. C. Platt, D. Koller, Y. Singer et S. Rowseis, éditeurs : *Advances in Neural Information Processing Systems 20*, pages 1297–1304, Cambridge, 2008. MIT Press.

[169] O. Shamir et N. Tishby : Model selection and stability in k-means clustering. *In* R. A. Servedio et T. Zhang, éditeurs : *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pages 367–378, Madison, 2008. Omnipress.

[170] G. R. Shorack et J. A. Wellner : *Empirical Processes with Applications to Statistics*. Society for Industrial and Applied Mathematics, Philadelphia, 2009.

[171] B. W. Silverman : Some aspects of spline smoothing approaches to non-parametric regression curve fitting. *Journal of the Royal Statistical Society*, 47:1–52, 1985.

[172] C. Spearman : General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.

[173] D. C. Stanford et A. E. Raftery : Finding curvilinear features in spatial point patterns: principal curve clustering with noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:2237–2253, 2000.

[174] H. Steinhaus : Sur la division des corps matériels en parties. *Bulletin de l'Académie Polonaise des Sciences, Classe III*, IV:801–804, 1956.

[175] A. Strehl et J. Ghosh : Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[176] T. Su et J. Dy : A deterministic method for initializing k-means clustering. *In Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, volume 141, pages 784–786, 2004.

[177] C. A. Sugar et G. M. James : Finding the number of clusters in a data set: an information theoretic approach. *Journal of the American Statistical Association*, 98:750–763, 2003.

[178] T. Tarpey et B. Flury : Self-consistency: a fundamental concept in statistics. *Statistical Science*, 11:229–243, 1996.

[179] A Tarsitano : Mahalanobis metrics for k-means algorithms. *In Atti del Convegno intermedio SIS, Napoli*, 2003.

[180] R. Tibshirani : Principal curves revisited. *Statistics and Computing*, 2:183–190, 1992.

[181] R. Tibshirani, G. Walther et T. Hastie : Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society*, 63:411–423, 2001.

[182] A. B. Tsybakov : On the best rate of adaptive estimation in some inverse problems. *Comptes-rendus de l'Académie des Sciences, Paris*, 2000.

[183] A. B. Tsybakov : *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009.

[184] A. van der Vaart et J. Wellner : *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer, New York, 1996.

[185] J. J. Verbeek, N. Vlassis et B. Kröse : A soft k-segments algorithm for principal curves. *In Proceedings of International Conference on Artificial Neural Networks 2001*, pages 450–456, 2001.

[186] U. von Luxburg, O. Bousquet et B. Schölkopf : A compression approach to support vector model selection. *Journal of Machine Learning Research*, 5:293–323, 2004.

[187] M. V. Wickerhauser : *Adapted Wavelet Analysis from Theory to Software*. A. K. Peters, Wellesley, 1994.

[188] R. C. Williamson, A. J. Smola et B. Schölkopf : Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47:2516–2532, 2001.

[189] D. J. H. Wilson, G. W. Irwin et G. Lightbody : RBF principal manifolds for process monitoring. *IEEE Transactions of Neural Networks*, 10:1424–1434, 1999.

[190] W. C. K. Wong et A. C. S. Chung : Principal curves to extract vessels in 3D angiograms. *In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'08)*, pages 1–8, 2008.

[191] W. C. K. Wong, R. W. K. So et A. C. S. Chung : Principal curves: a technique for preliminary carotid lumen segmentation and stenosis grading. *MIDAS Journal*, 2009.

[192] J. Wu, H. Xiong, J. Chen et W. Zhou : A generalization of proximity functions for K-means. *In Proceedings of the 2007 7th IEEE International Conference on Data Mining*, pages 361–370, 2007.

[193] L. Xu et M. I. Jordan : On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8:129–151, 1996.

[194] M. Zayed et J. Einbeck : Constructing economic summary indexes via principal curves. *In* Y. Lechevallier et G. Saporta, éditeurs : *Proceedings of the 19th International Conference on Computational Statistics, COMPSTAT 2010*, pages 1709–1716. Springer, 2010.

[195] F. Zhang, B. Wu, L. Zhang, H. Huang et Y. Tian : Illicit vessel identification in inland waters using sar image. *In Proceedings of the IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS 2006)*, pages 3144–3147, 2006.