Soumission

Journal de la Société Française de Statistique

Deux méthodes d'apprentissage non supervisé : synthèse sur la méthode des centres mobiles et présentation des courbes principales *

Title: Two methods in unsupervised learning : summary on k-means clustering and presentation of principal curves

Aurélie Fischer [1]

Résumé : Cet article propose une synthèse bibliographique sur le thème de l'apprentissage non supervisé. Après une introduction à la quantification et au problème connexe de classification par la méthode des centres mobiles, nous présentons la notion de courbe principale, qui peut être vue comme une généralisation de ces méthodes. Nous exposons différentes définitions de courbe principale et donnons un aperçu des applications de ces objets.

Abstract: This article proposes a review on unsupervised learning. After an introduction to quantization and to the related question of k-means clustering, the notion of principal curve, that may be seen as a generalization of these methods, is presented. We expound different definitions of principal curve and give an overview of its applications.

Mots-clés : apprentissage non supervisé, quantification, classification par la méthode des centres mobiles, courbes principales, synthèse bibliographique
Keywords: unsupervised learning, quantization, k-means clustering, principal curves, survey
Classification AMS 2000 : 62G05, 62G08, 62H30

1. Introduction

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 et l'apprentissage non supervisé. 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 $\{(X_1,Y_1),\dots,(X_n,Y_n)\}$. 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 construire un modèle permettant de représenter au mieux les observations $X_1,\dots,X_n$, de manière à la fois précise et compacte.

* L'auteur remercie l'ANR pour son soutien partiel via le projet TopData.
[1] Laboratoire de Probabilités et Modèles Aléatoires, Université Paris Diderot. E-mail : aurelie.fischer@univ-paris-diderot.fr

Soumis au Journal de la Société Française de Statistique
File: ArticleTheseSFdSAFischerRevision2.tex, compiled with jsfds, version : 2009/12/09
date: 17 mars 2014

L'objectif de cet article est de présenter deux méthodes d'apprentissage non supervisé, la quantification et l'estimation de courbes principales. Nous mettrons en relief leurs points communs.

Nous avons choisi de présenter la quantification dans $\mathbb{R}^d$, ce cadre étant celui qui permet l'analogie avec les courbes principales. Nous nous plaçons donc dans $(\mathbb{R}^d, \|\cdot\|)$ euclidien. Dans tout le document, $\mathbf{X}$ désigne un vecteur aléatoire de $\mathbb{R}^d$ et $\mathbf{f} = (f_1, \ldots, f_d)$ une courbe paramétrée.

Le plan général de l'article est le suivant. La section 2 est consacrée à la quantification et au problème connexe de classification par la méthode des centres mobiles, et la section 3 aux courbes principales. Puis nous conclurons en rappelant le lien entre les deux méthodes dans la section 4.

2. Quantification et méthode des centres mobiles

2.1. Présentation générale

La quantification et la classification par la méthode des centres mobiles, aussi connue sous le nom de méthode des *k-means*, sont deux aspects d'une même question, dont les applications concernent des domaines aussi variés que la biologie, l'informatique ou les sciences sociales. La première des deux notions correspond à la formulation probabiliste du problème, et la seconde, 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, un vecteur aléatoire $\mathbf{X}$ est représenté par $q(\mathbf{X})$, où l'application q envoie $\mathbb{R}^d$ dans un sous-ensemble fini d'éléments de $\mathbb{R}^d$. Une fonction appelée distorsion permet de contrôler l'erreur due à ce remplacement. Cette théorie est exposée de manière détaillée par Gersho et Gray (1992), Graf et Luschgy (2000) et Linder (2002). A la quantification correspond dans le contexte statistique l'une des formes les plus répandues de classification non supervisée, la méthode des *k-means* (voir, par exemple, Lloyd (1982)). La classification non supervisée, ou *clustering*, consiste, à partir d'un amas de données $\mathbf{X}_1, \ldots, \mathbf{X}_n$, supposées indépendantes et distribuées suivant la loi de $\mathbf{X}$, 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 *et al.* (2000)). Ces groupes jouent le rôle des éléments choisis comme représentants en quantification. Contrairement à la classification supervisée, il n'existe pas de classes déterminées à l'avance ou d'étiquettes. En effectuant une étape de *clustering*, on espère tirer des données certaines informations, en dégager des caractéristiques, y repérer des phénomènes sous-jacents.

2.2. Le principe de la quantification

On se donne un vecteur aléatoire $\mathbf{X}$ de $\mathbb{R}^d$ de loi μ tel que $\mathbb{E}\|\mathbf{X}\|^2 < \infty$. Soit $k \geq 1$ un entier. Un *k-quantificateur* est une application borélienne $q : \mathbb{R}^d \to \mathbf{c}$, où $\mathbf{c} = \{c_1, \ldots, c_\ell\}$, avec $\ell \leq k$,

Soumis au Journal de la Société Française de Statistique
File: ArticleTheseSFdSAFischerRevision2.tex, compiled with jsfds, version : 2009/12/09
date: 17 mars 2014

est un sous-ensemble de $\mathbb{R}^d$ appelé table de codage ou ensemble des centres associés au quantificateur q. Tout élément $\mathbf{x} \in \mathbb{R}^d$ est représenté par un unique $\hat{\mathbf{x}} = q(\mathbf{x}) \in \mathbf{c}$.

On souhaite représenter $\mathbf{X}$ aussi précisément que possible, donc limiter l'erreur commise en remplaçant $\mathbf{X}$ par $q(\mathbf{X})$. Cette erreur peut être évaluée au moyen de la distorsion

$$W(q) = \mathbb{E}[\|\mathbf{X} - q(\mathbf{X})\|^2] = \int_{\mathbb{R}^d} \|\mathbf{x} - q(\mathbf{x})\|^2 \mathrm{d}\mu(\mathbf{x}). \tag{1}$$

Chercher le meilleur k-quantificateur revient à minimiser cette distorsion. Soit

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

où Q_k désigne l'ensemble de tous les k-quantificateurs. Un k-quantificateur $q^\star$ vérifiant

$$W(q^\star) = W^\star$$

est dit optimal.

Chaque k-quantificateur est déterminé par sa table de codage $\{c_1, \ldots, c_\ell\}$ et une partition de $\mathbb{R}^d$ en cellules $S_j = \{\mathbf{x} \in \mathbb{R}^d : q(\mathbf{x}) = c_j\},\ j = 1, \ldots, \ell$. On a l'équivalence

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

Pour définir un quantificateur, il suffit donc de donner sa table de codage et sa partition.

Envisagé sous sa version *k-means*, le problème du *clustering* 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 $\mathbf{X}_1, \ldots, \mathbf{X}_n$ supposées être des variables aléatoires indépendantes toutes de loi μ. Soit μ_n la mesure empirique associée à $\mathbf{X}_1, \ldots, \mathbf{X}_n$, définie par

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

pour tout borélien A de $\mathbb{R}^d$. On introduit la distorsion empirique

$$W_n(q) = \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{X}_i - q(\mathbf{X}_i)\|^2, \tag{2}$$

qui est la distorsion (1) pour la loi μ_n. Comme pour la distorsion théorique, on note

$$W_n^\star = \inf_{q \in \mathcal{Q}_k} W_n(q).$$

Classer les données en groupes revient à chercher un quantificateur $q_n^\star$ optimal pour la distorsion empirique (2). En d'autres termes, $q_n^\star$ doit vérifier $W_n(q_n^\star) = W_n^\star$.

Soumis au Journal de la Société Française de Statistique
File: ArticleTheseSFdSAFischerRevision2.tex, compiled with jsfds, version : 2009/12/09
date: 17 mars 2014

2.3. Quantificateur des plus proches voisins

Puisqu'un quantificateur est caractérisé par sa partition et sa table de codage, pour obtenir le meilleur quantificateur possible (au sens de la distorsion $W(q)$), nous devons en principe déterminer la meilleure table de codage $\mathbf{c} = \{c_1, \ldots, c_\ell\}$ et la meilleure partition $\{S_1, \ldots, S_\ell\}$, où $\ell \leq k$.

Un type de partition important est la *partition de Voronoi*, définie par

$$S_1 = \{\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x} - c_1\| \leq \|\mathbf{x} - c_p\|, p = 1 \ldots, \ell\},$$

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

$$S_j = \{\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x} - c_j\| \leq \|\mathbf{x} - c_p\|, p = 1, \ldots, \ell\} \setminus \bigcup_{m=1}^{j-1} S_m.$$

Autrement dit, $\mathbf{x}$ est affecté à la cellule S_j si, et seulement si, cet élément est plus proche de c_j que de tous les autres centres, et en cas d'égalité, la cellule de plus petit indice est choisie. Un quantificateur associé à la partition de Voronoi est appelé *quantificateur des plus proches voisins*.

Un fait intéressant est que l'on connaît, à même table de codage, la meilleure partition (voir par exemple Linder (2002)).

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

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

S'il existe un k-quantificateur optimal, c'est donc nécessairement un quantificateur des plus proches voisins. Pour construire un bon quantificateur, il suffit par conséquent de considérer cette classe particulière de quantificateurs. Ainsi, trouver un quantificateur optimal, c'est trouver la table de codage optimale minimisant la distorsion récrite en fonction de $\mathbf{c}$,

$$W(\mathbf{c}) = \mathbb{E}\left[\min_{j=1,\ldots,k} \|\mathbf{X} - c_j\|^2\right].$$

De même, la distorsion empirique devient

$$W_n(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\ldots,k} \|\mathbf{X}_i - c_j\|^2.$$

D'autre part, on sait décrire la meilleure table de codage pour un quantificateur de partition donnée.

Lemme 2 (Meilleure table de codage)**.** *Soit q un quantificateur de partition associée $\{S_j\}_{j=1}^{\ell}$, où $\ell \leq k$, avec $\mu(S_j) > 0$ 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*

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

alors

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

Soumis au Journal de la Société Française de Statistique
File: ArticleTheseSFdSAFischerRevision2.tex, compiled with jsfds, version : 2009/12/09
date: 17 mars 2014

L'existence de quantificateurs optimaux a été démontrée par Pollard (1982b).

Théorème 1 (Existence). *Il existe une table de codage optimale* $\mathbf{c}^\star$ *telle que* $W(\mathbf{c}^\star) = W^\star$, *et donc, un quantificateur des plus proches voisins* $q^\star$ *tel que* $W(q^\star) = W^\star$. *En particulier, il existe* $\mathbf{c}_n^\star$ *tel que* $W_n(\mathbf{c}_n^\star) = W_n^\star$.

Supposons que nous disposons d'un quantificateur empirique optimal $q_n^\star$ construit à l'aide d'observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Se pose alors la question de savoir si, asymptotiquement, la performance de $q_n^\star$ approche celle d'un « vrai » quantificateur optimal $q^\star$. Ce problème de consistance de la distorsion a été étudié par Pollard (1982b, 1981, 1982a), Abaya et Wise (1984) et Graf et Luschgy (1994).

Théorème 2 (Consistance). *Soit* $\mathbf{c}_n^\star$ *une table de codage empirique optimale. On a*

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

On peut ensuite s'intéresser à la vitesse de convergence, pour avoir une idée de la taille de l'échantillon à partir de laquelle $W(\mathbf{c}_n^\star)$ devient effectivement très proche de la distorsion optimale $W^\star$. Nous donnons ici une borne non-asymptotique établie par Bartlett *et al.* (1998) (voir aussi Pollard (1982a); Linder *et al.* (1994); Linder (2000)).

Théorème 3 (Vitesse). *Supposons que* μ *vérifie* $\mathbb{P}(\|\mathbf{X}\| \leq 1) = 1$. *Alors, il existe une constante* C *telle que*

$$\mathbb{E}[W(\mathbf{c}_n^\star)] - W^\star \leq C \min\left(\sqrt{\frac{kd}{n}}, \sqrt{\frac{k^{1-2/d} d \ln n}{n}}\right).$$

En outre, il existe une loi μ *avec* $\mathbb{P}(\|\mathbf{X}\| \leq 1) = 1$, *telle que, pour une constante* c,

$$\mathbb{E}[W(\mathbf{c}_n^\star)] - W^\star \geq c\sqrt{\frac{k^{1-4/d}}{n}}.$$

Notons que sous certaines conditions de régularité sur la loi de $\mathbf{X}$, énoncées par Pollard (1982a), la vitesse de convergence peut être plus rapide, de l'ordre de $1/n$. Après un résultat en probabilité de Chou (1994) et une borne supérieure en $\mathcal{O}(\ln n/n)$ établie par Antos *et al.* (2005)), une borne en $\mathcal{O}(1/n)$ a récemment été obtenue par Levrard (2013) grâce au principe de localisation proposé par Blanchard *et al.* (2008) et Koltchinskii (2006).

2.4. L'aspect algorithmique

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 et 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 de l'algorithme des *k-means* (Steinhaus (1956); MacQueen (1967); Lloyd (1982); Linde *et al.* (1980)).

Explicitons l'algorithme lorsque μ est inconnue, c'est-à-dire dans le cas du *clustering*. A partir d'une table de codage initiale $\{c_{0,1}, \ldots, c_{0,\ell}\}$, $\mathbb{R}^d$ est partitionné en cellules de Voronoi $S_{0,1}, \ldots, S_{0,\ell}$ en affectant chaque donnée $\mathbf{X}_i$ au centre $c_{0,j}$ le plus proche. Ensuite,

les nouveaux centres $c_{1,1},\ldots,c_{1,\ell}$ sont calculés en effectuant la moyenne des $\mathbf{X}_i$ tombés dans la cellule S_j, et ces deux étapes sont itérées jusqu'à ce que la table de codage demeure inchangée, ce qui signifie qu'un minimum local a été atteint.

Les étapes de l'algorithme, illustrées dans la Figure 1, se résument donc ainsi :

1. Initialisation
   Choix des centres et détermination de la partition de Voronoi associée.
2. Etape espérance conditionnelle
   Calcul des nouveaux centres.
3. Etape de projection
   Détermination de la nouvelle partition de Voronoi.
4. Critère d'arrêt
   Convergence (vers un minimum local).

Nous verrons dans la suite qu'un algorithme de courbe principale repose sur le même type d'alternance entre étape de projection et étape de calcul d'espérance.

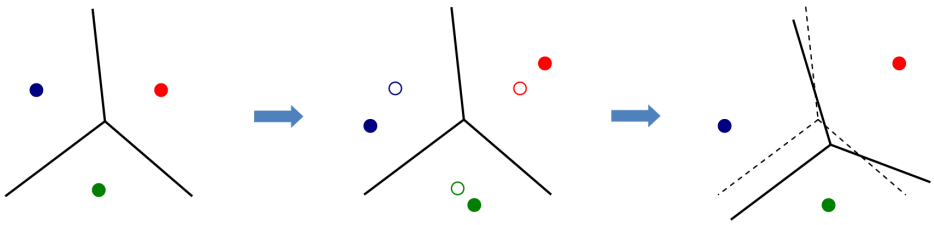

FIGURE 1. *Etapes de l'algorithme des* k-means.

2.5. Quelques idées pour le choix de k

Dans ce qui précède, nous avons supposé le nombre k de groupes fixé. En pratique, la question du choix de k est essentielle. Dans certaines situations, ce choix est dicté par les applications, mais bien souvent, il constitue un problème délicat.

Tout d'abord, remarquons qu'il semble naturel d'utiliser la distorsion empirique pour choisir la valeur de k appropriée, puisque toute l'information disponible est contenue dans cette quantité dépendant des observations (notée ici $W_n(k)$ pour mettre en évidence la dépendance en k). Cependant, la distorsion empirique est une fonction décroissante du nombre de groupes, de sorte que minimiser directement ce critère en k conduit à choisir la plus grande valeur possible, ce qui n'est pas raisonnable. Par exemple, si $k = n$, chaque observation forme un groupe à elle seule. Hastie *et al.* (2001) notent d'ailleurs qu'évaluer la distorsion sur un ensemble test de données indépendant ne suffit pas pour trouver k. 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. Il n'est donc pas possible de procéder par validation croisée comme en apprentissage supervisé. Néanmoins, une stratégie de choix de k peut se baser sur le fait que $W_n(k)$ 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.

Dans cette section, nous présentons quelques procédures permettant de déterminer k automatiquement. Nous verrons que la distorsion empirique intervient effectivement dans la plupart des moyens proposés dans la littérature pour choisir k.

On trouve une présentation de différentes méthodes dans Milligan et Cooper (1985) ainsi que Hardy (1996), et Gordon (1999) compare les performances des cinq meilleures règles exposées dans Milligan et Cooper (1985). Il faut distinguer les procédures globales, consistant à effectuer un *clustering* pour plusieurs valeurs de k afin de déterminer la valeur optimale 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 (1974) proposent de choisir la valeur de k maximisant un indice basé sur le quotient

$$\frac{B_n(k)/(k-1)}{W_n(k)/(n-k)},$$

où $B_n(k) = \sum_{j=1}^{k} \|\mathbf{c}_j - \bar{\mathbf{c}}\|$, avec $\bar{\mathbf{c}}$ la moyenne des observations, est l'inertie inter-classes. La méthode de Krzanowski et Lai (1985) consiste à maximiser $W_n(k)k^{2/d}$, ou plus précisément la quantité équivalente

$$\left| \frac{\mathrm{DIFF}(k)}{\mathrm{DIFF}(k+1)} \right|,$$

où $\mathrm{DIFF}(k) = W_n(k-1)(k-1)^{2/d} - W_n(k)(k)^{2/d}$, tandis que dans la règle d'Hartigan (1975), un nouveau groupe est ajouté tant que la quantité

$$H(k) = \left(\frac{W_n(k)}{W_n(k+1)} - 1 \right)(n-k-1)$$

dépasse un certain seuil. La statistique *Silhouette* de Kaufman et Rousseeuw (1990) 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 (1990) suggèrent donc de choisir la valeur de k maximisant la moyenne des $s(i)$ pour $i = 1, \ldots, n$. Dans la méthode de la *Gap Statistic* de Tibshirani *et al.* (2001), le choix de k est basé sur la comparaison de la variation du logarithme de la distorsion empirique pour le problème de *clustering* considéré et de celle obtenue pour des données uniformément distribuées. Kim *et al.* (2001) développent un indice donnant k en combinant deux fonctions de monotonies opposées qui présentent un saut autour de la valeur de k optimale, alors que Sugar et James (2003) proposent, pour visualiser un saut dans le tracé de $k \mapsto W_n(k)$, d'appliquer à la distorsion empirique une transformation de la forme $w \mapsto w^{-p}$ avec $p > 0$. Il existe également des méthodes basées

sur la stabilité des partitions, dans lesquelles k est sélectionné d'après les résultats obtenus en classant plusieurs sous-échantillons de l'ensemble des observations (voir par exemple Levine et Domany (2002) et Ben-Hur *et al.* (2002)). La relation entre le nombre k et la stabilité des groupes est analysée d'un point de vue théorique dans Shamir et Tishby (2008a,b), Ben-David *et al.* (2006), Ben-David *et al.* (2007) et Ben-David et von Luxburg (2008). Une position encore différente est adoptée dans Fischer (2011), où la question du choix de k est vue comme un problème de sélection de modèle par pénalisation (Birgé et Massart (1997)).

Remarque. *Nous avons présenté la quantification et le* clustering k-means *dans l'espace* $\mathbb{R}^d$ *muni de la norme euclidienne standard. En effet, il s'agit du cadre dans lequel l'analogie avec les courbes principales a un sens. Notons cependant qu'il est possible de considérer d'autres mesures de distorsion, par exemple dans le but de classer des observations de grande dimension ou de nature complexe. Dans le cadre du* clustering*, Biau* et al. *(2008) et Cadre et Paris (2012) considèrent le cas d'une norme hilbertienne au carré, et Laloë (2010) celui d'une norme* L^1*. Luschgy et Pagès (2002, 2006) étudient la quantification de processus gaussiens et de diffusions avec une norme* L^p*, tandis que Dereich et Vormoor (2011) traitent le problème de la quantification avec une norme d'Orlicz. Banerjee* et al. *(2005) proposent l'utilisation des divergences de Bregman (Bregman (1967)) comme notion de dissimilarité dans l'algorithme des* k-means *et Fischer (2010) analyse les propriétés théoriques de la quantification et du* clustering *basé sur cette classe de mesures de distorsion.*

A titre d'exemple, Biau et al. *(2008) obtiennent comme analogue du Théorème 3 dans le cadre hilbertien la borne supérieure suivante, indépendante de la dimension :*

$$\mathbb{E}[W(\mathbf{c}_n^\star)] - W^\star \leq C' \frac{k}{\sqrt{n}}.$$

3. Courbes principales

3.1. Contexte et notations

Parmi les méthodes statistiques utilisées pour résumer de l'information et représenter les données par certaines grandeurs « simplifiées », l'analyse en composantes principales est certainement l'une des plus célèbres. Cette technique, initiée au début du siècle dernier par les travaux de Pearson (1901) et Spearman (1904), puis développée par Hotelling (1933), 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 *et al.* (1979)). Que ce soit dans le cadre de la réduction de dimension ou de l'extraction de caractéristiques, elle peut fournir un 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 2). Il existe plusieurs moyens de donner un sens mathématique à cette idée. La définition dépend par exemple de la propriété des composantes

Soumis au Journal de la Société Française de Statistique
File: ArticleTheseSFdSAFischerRevision2.tex, compiled with jsfds, version : 2009/12/09
date: 17 mars 2014

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 $\mathbf{X}$.

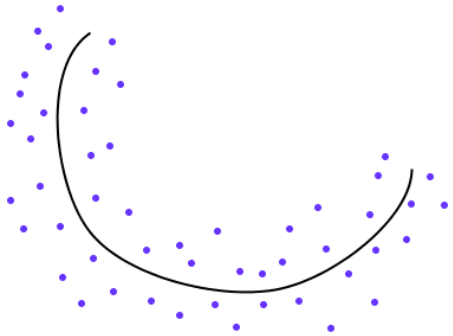

FIGURE 2. *Une courbe passant « au milieu » d'un nuage de points.*

La définition originelle est due à Hastie et Stuetzle (1989). Elle est basée sur la propriété d'auto-consistance des composantes principales. Plusieurs points de vue, plus ou moins proches de celui-ci, ont été introduits ensuite. Nous proposons dans cette section une présentation des différentes définitions, complétée par quelques propriétés des courbes principales d'ordre supérieur ainsi qu'un tour d'horizon des applications.

Soit $\mathbf{f} = (f_1, \ldots, f_d)$ une courbe paramétrée continue définie sur un intervalle fermé $I = [a, b]$. Si rien n'est précisé, $\mathbf{f}$ est supposée paramétrée par l'arc. La dérivée de la courbe paramétrée $\mathbf{f}$ est notée $\mathbf{f}'$ et la dérivée seconde $\mathbf{f}''$. Les définitions et résultats utiles concernant les courbes paramétrées sont rassemblés dans l'Appendice.

3.2. Des courbes paramétrées auto-consistantes

Dans cette section, nous présentons la définition originelle des courbes principales, donnée par Hastie et Stuetzle dans les années 1980 (Hastie (1984); Hastie et Stuetzle (1989)).

3.2.1. Définition de Hastie et Stuetzle

Tout d'abord, il nous faut introduire la notion d'indice de projection.

Définition 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*

$$t_{\mathbf{f}}(\mathbf{x}) = \sup\{t \in I, \|\mathbf{x} - \mathbf{f}(t)\| = \inf_{t'} \|\mathbf{x} - \mathbf{f}(t')\|\}. \tag{3}$$

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 $\|\mathbf{x} - \mathbf{f}(t)\|$ (Hastie et Stuetzle, 1989, 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 3.

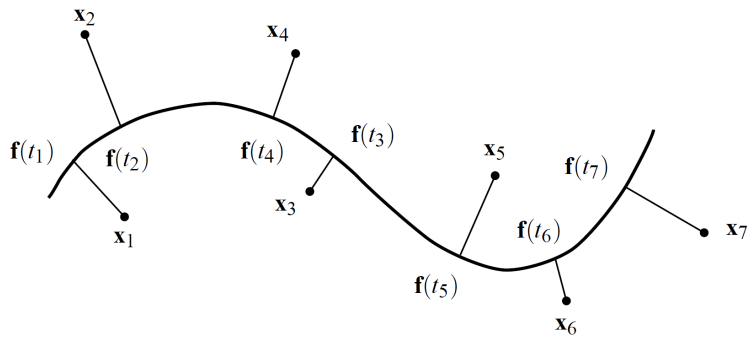

FIGURE 3. *Indice de projection. Pour tout i, t_i désigne $t_{\mathbf{f}}(\mathbf{x}_i)$.*

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 (1996).

Définition 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,*

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

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 sur ce point.

Finalement, une courbe principale est définie par Hastie et Stuetzle de la manière suivante :

Définition 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. *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.*

3.2.2. Lien avec l'analyse en composantes principales

Les courbes principales de Hastie et Stuetzle (1989) 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, elle correspond à 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 $\mathscr{G}$ est une classe de courbes paramétrées sur l'intervalle I et $\mathbf{f}_t = \mathbf{f} + t\mathbf{g}$, où $\mathbf{g} \in \mathscr{G}$, on dit que la courbe $\mathbf{f}$ est un point critique relativement à la classe $\mathscr{G}$ si, pour toute courbe $\mathbf{g} \in \mathscr{G}$,

$$\left.\frac{\mathrm{d}\mathbb{E}\|\mathbf{X} - \mathbf{f}_t(t_{\mathbf{f}_t}(\mathbf{X}))\|^2}{\mathrm{d}t}\right|_{t=0} = 0.$$

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 Hastie et Stuetzle (1989).

Proposition 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, pour tout* $t \in I$, $\|\mathbf{g}(t)\| \leq 1$ *et* $\|\mathbf{g}'(t)\| \leq 1$.

L'existence de courbes principales répondant à la définition de Hastie et Stuetzle (1989) est un problème ouvert en général. Duchamp et Stuetzle (1996) 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 (1989) généralisent cette notion de courbe principale aux dimensions supérieures, en recherchant des surfaces vérifiant la propriété d'auto-consistance.

3.2.3. 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 (1989), 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 la quantité $\mathbf{f}^{(j+1)}(t) = \mathbb{E}[\mathbf{X} | t_{\mathbf{f}^{(j)}}(\mathbf{X}) = t]$ 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 dont les sommets sont $\mathbf{f}^{(j)}(t_1^{(j-1)}), \ldots, \mathbf{f}^{(j)}(t_n^{(j-1)})$, 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 4). Avant de passer à l'étape suivante, les $t_i^{(j)}$ sont à nouveau rangés dans l'ordre croissant.

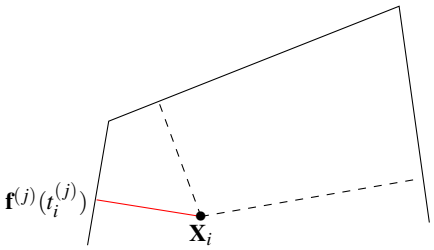

FIGURE 4. *Etape de projection. Pour déterminer* $t_i^{(j)}$*, il faut rechercher le segment minimisant la projection de* $\mathbf{X}_i$ *sur la ligne polygonale.*

Pour l'étape 3, remarquons qu'en général, pour un point $(t_i^{(j)}, \mathbf{f}^{(j)}(t_i^{(j)}))$ 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 lissage. Hastie et Stuetzle (1989) proposent ainsi d'appliquer à chaque fonction coordonnée un lissage LOWESS (*Locally Weighted Scatterplot Smoothing*) (Cleveland (1979)) ou encore d'utiliser des splines cubiques de lissage (Silverman (1985)).

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, chaque t_m étant affecté d'un poids qui dépend de sa distance à t_i. Le choix de ce 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. La taille du voisinage dépend de l'objectif pratique ; dans certains cas, il existe un choix naturel dicté par l'application. Un moyen envisagé par Hastie et Stuetzle (1989) pour sélectionner automatiquement ce paramètre est de procéder par validation croisée.

La seconde méthode consiste à chercher $\mathbf{f}$, une spline cubique paramétrée sur $[0,1]$, et $t_1, \ldots, t_n \in [0,1]$, minimisant

$$\sum_{i=1}^{n} \|\mathbf{X}_i - \mathbf{f}(t_i)\|^2 + \lambda \int_0^1 \|\mathbf{f}''(t)\|^2 \mathrm{d}t, \tag{4}$$

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 (4) 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. Dans ce cas, l'étape de projection de l'algorithme 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

$$\sum_{i=1}^{n}(X_{im}-f_m^{(j+1)}(t_i^{(j)}))^2+\lambda\int(f_m^{(j+1)''}(t))^2\mathrm{d}t,\quad m=1,\ldots,d.$$

3.2.4. Biais d'estimation et biais de modèle

La procédure d'Hastie et Stuetzle (1989) présente deux types de biais, ayant des effets opposés.

D'une part, comme nous l'avons mentionné plus haut, le résultat de l'étape de lissage LOWESS dépend fortement de la taille du voisinage. De même, dans la méthode utilisant des splines, la courbe principale obtenue dépend du coefficient de pénalité. Le biais d'estimation trouve son origine dans la procédure par moyennes locales, qui a tendance à aplatir la courbe. Plus les paramètres sont choisis grands, plus ce biais est important. Notons que Banfield et Raftery (1992), 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 (1998) remarquent qu'un meilleur résultat peut être obtenu en combinant l'algorithme de Banfield et Raftery (1992) avec celui de Hastie et Stuetzle (1989).

D'autre part, supposons que $\mathbf{X}=(X_1,\ldots,X_d)$ s'écrive sous la forme

$$X_j=f_j(S)+\varepsilon_j,\quad j=1,\ldots,d, \tag{5}$$

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 (1989). Un biais se produit lorsque la courbure est importante, comme l'illustre la Figure 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 (1989)).

Signalons 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.

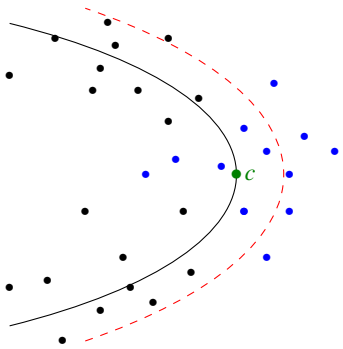

FIGURE 5. *Un biais lié à la courbure. Les observations se projetant au voisinage du point c étant plus nombreuses à l'extérieur de la courbure, la courbe principale ne peut être la courbe générative (en trait plein), elle est décalée (en pointillés).*

3.2.5. Auto-consistance et modèle de mélange

L'existence du biais de modèle conduit Tibshirani (1992), quelques années plus tard, à apporter une modification à cette définition de courbes principales, faisant appel aux modèles de mélange. Il s'agit de définir directement les courbes principales à partir du modèle (5).

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 (1992) s'écrit alors comme suit.

Définition 4. *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–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 de $\mathbb{R}$ telle que*

$$\mathbf{f}(s) = \mathbb{E}[\mathbf{X}|S = s].$$

Cette définition ne coïncide pas en général avec celle de Hastie et Stuetzle Hastie et Stuetzle (1989). 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 $g_{\mathbf{X}}$. 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 $g_{\mathbf{X}|S}$. Les $X_{i1},\ldots,X_{id}$ sont supposées conditionnellement indépendantes sachant S_i et $\mathbf{f}(s)=\mathbb{E}[\mathbf{X}|S=s]$. Dans le contexte des courbes principales, ce modèle de mélange peut s'interpréter en notant que les variables latentes permettent de définir n points « idéaux »échantillonnés sur une courbe, auxquels on ajoute une erreur pour construire les données. Par exemple, on peut imaginer que les observations ont été obtenues en bruitant des points distribués uniformément sur la courbe.

Tibshirani (1992) 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. Soit $\theta(s)=(\mathbf{f}(s),\Sigma(s))$. L'estimation peut être effectuée par maximum de vraisemblance. La log-vraisemblance observée s'écrit

$$\mathscr{L}(\theta;\mathbf{x})=\sum_{i=1}^{n}\ln\int g_{\mathbf{X}|S}(\mathbf{x}_i|\theta(s))g_S(s)\mathrm{d}s.$$

En pratique, la maximisation de la log-vraisemblance est effectuée à l'aide d'un algorithme de type EM (*Expectation-Maximization*, Dempster *et al.* (1977), Xu et Jordan (1996)). Cet algorithme utilise la log-vraisemblance des données complétées

$$\mathscr{L}(\theta;\mathbf{x},s)=\sum_{i=1}^{n}\ln g_{\mathbf{X}|S}(\mathbf{x}_i|\theta(s_i))+\sum_{i=1}^{n}\ln g_S(s_i).$$

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 (1989) utilisant des splines, 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 pénalisée est alors donnée par

$$\sum_{i=1}^{n}\ln g_{\mathbf{X}|S}(\mathbf{x}_i|\theta(s_i))+\sum_{i=1}^{n}\ln g_S(s_i)-(b'-a')\sum_{j=1}^{d}\lambda_j\int_{a'}^{b'}f_j''(s)^2\mathrm{d}s.$$

Soumis au Journal de la Société Française de Statistique
File: ArticleTheseSFdSAFischerRevision2.tex, compiled with jsfds, version : 2009/12/09
date: 17 mars 2014

3.3. Un problème de minimisation de moindres carrés

Le fait que l'existence de courbes principales au sens de Hastie et Stuetzle (1989) 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

$$\Delta(\mathbf{f}) = \mathbb{E}\left[\inf_{t\in I}\|\mathbf{X}-\mathbf{f}(t)\|^2\right] = \mathbb{E}\left[\|\mathbf{X}-\mathbf{f}(t_{\mathbf{f}}(\mathbf{X}))\|^2\right], \tag{6}$$

où $t_{\mathbf{f}}$ désigne l'indice de projection (voir Equation (3)), est étroitement lié à la propriété d'auto-consistance caractérisant la définition de Hastie et Stuetzle (1989). Nous verrons que ces deux définitions de courbe principale sont en rapport avec la notion de quantificateur optimal.

3.3.1. Courbes principales de longueur bornée

La définition de courbes principales sous forme de problème de moindres carrés a été proposée par Kégl *et al.* (2000). Ces auteurs considèrent des courbes principales de longueur bornée.

Définition 5. *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 (1989), 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.* (2000), en particulier en ce qui concerne le côté algorithmique. La définition de la longueur d'une courbe non supposée différentiable est donnée dans l'Appendice (Définition 12).

Avec cette approche, le problème de l'existence d'une courbe principale est résolu, puisque, comme le montre la proposition suivante due à Kégl *et al.*, la réponse est positive dans un cadre très général.

Proposition 2. *Dès que* $\mathbb{E}\|\mathbf{X}\|^2<+\infty$*, l'existence d'une courbe principale pour* $\mathbf{X}$ *est assurée.*

Il s'agit, comme en quantification, de minimiser un critère de type moindres carrés. En fait, cette analogie avec la quantification apparaît naturellement dès lors que l'on s'intéresse à la relation entre la définition de Hastie et Stuetzle (1989) et celle de Kégl *et al.* (2000). 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. En effet, 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 $\mathbf{y} \mapsto \mathbb{E}[\|\mathbf{X}-\mathbf{y}\|^2 | s(\mathbf{X}) = t]$ sur $\mathbb{R}^d$. Remarquons encore que le quantificateur q correspond dans le cadre des courbes principales à la fonction $\mathbf{f} \circ t_{\mathbf{f}}$.

Dans cette analogie, la définition de courbe principale de Hastie et Stuetzle (1989) 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 statistique, 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

$$\Delta_n(\mathbf{f}) = \frac{1}{n} \sum_{i=1}^{n} \inf_{t \in I} \|\mathbf{X}_i - \mathbf{f}(t)\|^2 = \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{X}_i - \mathbf{f}(t_{\mathbf{f}}(\mathbf{X}_i))\|^2. \tag{7}$$

Dans la section 2, nous avons évalué la performance d'un quantificateur empirique optimal en comparant sa distorsion avec la distorsion théorique optimale. 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.* (2000) 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 4. *Supposons que $\mathbb{P}(\mathbf{X} \in \mathscr{C}) = 1$, où $\mathscr{C}$ est un convexe fermé borné de $\mathbb{R}^d$. Soit $\mathscr{F}_L$ l'ensemble des courbes paramétrées de longueur au plus L, dont l'image est incluse dans $\mathscr{C}$. 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*

$$\Delta(\hat{\mathbf{f}}_{k,n}) - \min_{\mathbf{f} \in \mathscr{F}_L} \Delta(\mathbf{f}) = \mathscr{O}(n^{-1/3}).$$

D'un point de vue pratique, Kégl *et al.* (2000) proposent un algorithme itératif baptisé *Polygonal Line Algorithm* qui fournit une ligne brisée, approximation de courbe principale. En numérotant les sommets et segments d'une ligne polygonale comme indiqué dans la Figure 6, cet algorithme peut être décrit de la manière suivante :

1. Initialisation
   Soit $\mathbf{f}^{(1)}$ le plus petit segment correspondant à la première composante principale de $\mathbf{X}_1, \ldots, \mathbf{X}_n$ qui contienne toutes leurs projections.
2. Boucle externe
   Un sommet, et donc un segment, est ajouté à la ligne polygonale courante $\mathbf{f}^{(k)}$ constituée des sommets $v_1, \ldots, v_{k+1}$ et segments $s_1, \ldots, s_k$.
3. Boucle interne
   Les positions des sommets sont recalculées de manière cyclique, en déplaçant un sommet à la fois.

Soumis au Journal de la Société Française de Statistique
File: ArticleTheseSFdSAFischerRevision2.tex, compiled with jsfds, version : 2009/12/09
date: 17 mars 2014

- Etape de projection
  Les données $\mathbf{X}_1,\ldots,\mathbf{X}_n$ sont réparties dans au plus $2k+1$ ensembles disjoints, formant une partition de $\mathbb{R}^d$, en fonction du segment ou du sommet sur lequel elles se projettent.
- Etape d'optimisation
  La position du sommet v_j est ajustée en minimisant une version locale du critère empirique Δ_n.
- Sortie de la boucle
  Un critère d'arrêt reposant sur la variation du critère Δ_n est utilisé.

4. Condition d'arrêt
   L'algorithme se termine lorsque k dépasse un certain seuil, qui dépend du nombre d'observations n et du critère Δ_n.

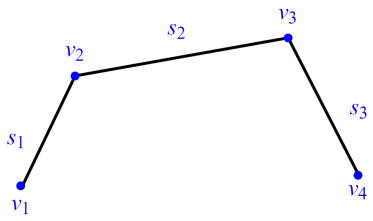

FIGURE 6. *Numérotation des segments et sommets pour $k=3$.*

Donnons quelques détails et commentaires sur les étapes 2 et 3 de l'algorithme. La première remarque est relative à la manière d'ajouter un sommet dans la boucle externe : on choisit pour nouveau sommet le milieu du segment sur lequel se projettent le plus grand nombre de données, ou, en cas d'égalité, le milieu du segment le plus long (voir Figure 7).

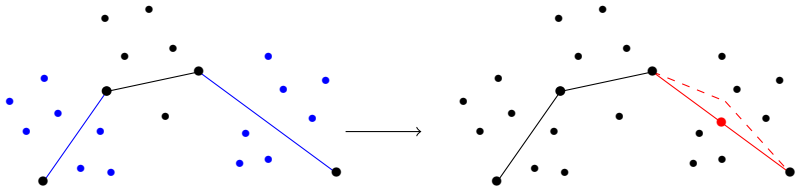

FIGURE 7. *Ajout d'un sommet. On cherche les segments sur lesquels se projettent le plus grand nombre de données : le nouveau sommet, qui sera ensuite ajusté, est le milieu du plus long d'entre eux.*

Concernant la boucle interne, notons, pour $\mathbf{x}\in\mathbb{R}^d$,

$$\Delta(\mathbf{x},\mathbf{f}) = \inf_{t\in I}\|\mathbf{x}-\mathbf{f}(t)\|^2$$
$$\Delta(\mathbf{x},s_j) = \inf_{\mathbf{y}\in s_j}\|\mathbf{x}-\mathbf{y}\|^2, \quad j=1,\ldots,k,$$
$$\Delta(\mathbf{x},v_j) = \|\mathbf{x}-v_j\|^2, \quad j=1,\ldots,k+1.$$

Soumis au Journal de la Société Française de Statistique
File: ArticleTheseSFdSAFischerRevision2.tex, compiled with jsfds, version : 2009/12/09
date: 17 mars 2014

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 $j = 1, \ldots, k$. Au cours de l'étape de projection, chaque observation est affectée à l'un des ensembles $V_1, \ldots, V_{k+1}, S_1, \ldots, S_k$. On constate une grande similitude entre cette étape de projection et celle qui a lieu en quantification, puisque la partition définie ici rappelle celle de Voronoi, comme l'illustre la Figure 8. Sa spécificité réside dans le fait qu'elle est construite à la fois par rapport aux sommets et aux segments.

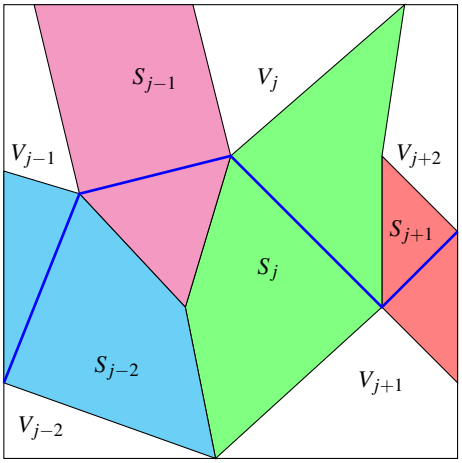

FIGURE 8. *Les ensembles $V_1, \ldots, V_{k+1}$ et $S_1, \ldots, S_k$ formant une partition de $\mathbb{R}^2$.*

Dans l'étape d'optimisation, le critère à minimiser pour déterminer la nouvelle localisation du sommet v_j est la quantité

$$\frac{1}{n}\left[\sum_{\mathbf{X}_i \in S_{j-1}} \Delta(\mathbf{X}_i, s_{j-1}) + \sum_{\mathbf{X}_i \in V_j} \Delta(\mathbf{X}_i, v_j) + \sum_{\mathbf{X}_i \in S_j} \Delta(\mathbf{X}_i, s_j)\right], \tag{8}$$

à laquelle s'ajoute une pénalité sur les angles. Remarquons que l'expression (8) ne fait intervenir que 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.

3.3.2. Courbes principales de courbure intégrale bornée

Dans sa thèse, Kégl (1999) 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 (2002), qui observent en outre que le point de vue de Kégl *et al.* (2000) 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 (6) et (7) que Kégl *et al.* (2000), 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 9, cette quantité est alors décrite simplement, de manière géométrique.

Définition 6 (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* $\overrightarrow{v_jv_{j+1}}$, *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*

$$\mathscr{K}(\mathbf{f}) = \sum_{j=2}^{k} \phi_j.$$

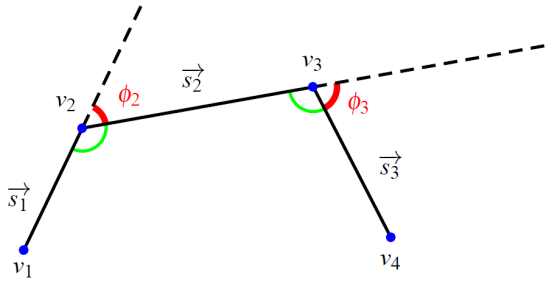

FIGURE 9. *En notant* $\overrightarrow{s_j}$ *le vecteur* $\overrightarrow{v_jv_{j+1}}$ *pour tout* $j=1,\dots,k$, *la courbure intégrale de la ligne polygonale de sommets les* v_j *est la somme des angles* $\phi_{j+1}=(\overrightarrow{s_j},\overrightarrow{s_{j+1}})$.

La définition pour une courbe générale s'obtient en approximant cette courbe par des fonctions linéaires par morceaux (voir également l'Appendice).

Définition 7 (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* $\mathbf{f}(t_0),\dots,\mathbf{f}(t_p)$, *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 au livre d'Alexandrov et Reshetnyak (1989).

Sandilya et Kulkarni (2002) 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 $1/2$. 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. 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 8. *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* $\mathscr{C}_{K,\tau}$ *définie par*

$$\mathscr{C}_{K,\tau} = \{\mathbf{f} : \mathscr{K}(\mathbf{f}) \leq K, \mathscr{K}(\mathbf{f}) - \mathscr{K}(\mathbf{f}|_{B_R}) \leq \tau(R)\},$$

où τ *est une fonction continue qui décroît vers 0.*

La classe $\mathscr{C}_{K,\tau}$ 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 (2002) sont en mesure de démontrer l'existence des courbes principales de courbure intégrale bornée.

Proposition 3. *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 $\mathscr{C}_{k,\tau}$ 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 de Sandilya et Kulkarni (2002), 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.* (2000) énoncé dans le Théorème 4. L'hypothèse $\mathbb{P}(\mathbf{X} \in \mathscr{C}) = 1$ 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 5. *Soit* $\alpha > 0$*. 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* $\hat{\mathbf{f}}_{n,k} \in \mathscr{C}_{k,\tau}$ *est une ligne brisée à* k *segments d'image incluse dans* B_{R_k}*, minimisant le critère empirique* $\Delta_n(\mathbf{f})$*, alors*

$$\Delta(\hat{\mathbf{f}}_{k,n}) - \min_{\mathbf{f} \in \mathscr{C}_{K,\tau}} \Delta(\mathbf{f}) = \mathscr{O}(n^{-\frac{\alpha}{6+3\alpha}}).$$

Remarque. *La sélection des paramètres de longueur, courbure et nombre de segments dans les définitions de courbe principale de Kégl et al. (2000) et Sandilya et Kulkarni (2002) est étudiée dans Biau et Fischer (2012) et Fischer (2013) sous l'angle de la sélection de modèle non-asymptotique par critère pénalisé introduite par Birgé et Massart (1997) dans les années 1990.*

Par ailleurs, dans le but de s'affranchir des contraintes sur la complexité des courbes, Gerber et Whitaker (2013) ont proposé récemment de renverser le problème de minimisation, en fixant la courbe $\mathbf{f} = \mathbb{E}[\mathbf{X}|t(\mathbf{X})]$ *et en développant un critère à minimiser en le paramètre* t*. Ce critère alternatif, qui fait intervenir directement la notion d'orthogonalité, s'écrit*

$$\mathbb{E}\left[\left\langle \mathbf{f}(t(\mathbf{X})) - \mathbf{X}, \frac{\mathrm{d}}{\mathrm{d}s}\mathbf{f}(s)\Big|_{s=t(\mathbf{X})} \right\rangle^2\right].$$

3.4. Quelques définitions reposant sur une analyse locale

Les différentes 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 (1989) ainsi que celle de Tibshirani (1992), et nous avons vu que réaliser le minimum du critère de type moindres carrés de Kégl *et al.* (2000) et Sandilya et Kulkarni (2002) 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.

3.4.1. Courbes principales de points orientés principaux

La définition de *courbe principale de points orientés*, proposée par Delicado (2001) et étendue dans Delicado et Huerta (2003), 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 $m^\star(\mathbf{x}) = \{m(\mathbf{x},\mathbf{y}), \mathbf{y} \in \mathbf{y}^\star(\mathbf{x})\}$ 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 4. *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 (2001) introduit le concept de *points orientés principaux*.

Définition 9 (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 10. *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 (2001) 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 (1989). 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 (2001) 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.

3.4.2. Composantes principales locales

Dans l'approche de Einbeck *et al.* (2005b), une courbe principale est calculée à partir d'un ensemble de premières composantes principales locales. Ce point de vue local, qui repose sur l'introduction d'un noyau, 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 $\hat{\Sigma}^{\mathbf{x}} = (\hat{\sigma}_{jk}^{\mathbf{x}})_{1 \leq j,k \leq d}$ 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 Verbeek *et al.* (2001) ont également proposé un algorithme de courbes principales basé sur une analyse en composantes principales locale. Appelé « k-segment algorithm », il consiste à construire une courbe principale en assemblant plusieurs segments.

3.4.3. Les courbes principales comme lignes de crête d'une densité

Dans la propriété d'auto-consistance, chaque point de la courbe principale peut être vu comme la moyenne des données qui se projettent au voisinage de ce point, c'est-à-dire qui se trouvent dans le sous-espace orthogonal à la courbe à cet endroit.

Considérant des variables aléatoires à densité, Ozertem et Erdogmus (2011) reprennent en partie cette idée, en proposant la modification suivante : les points de la courbe principale ne correspondent plus à une moyenne mais à un maximum. Chaque point de la courbe est un maximum local de la densité de probabilité dans le sous-espace orthogonal local. La courbe obtenue est alors la ligne de crête de la densité.

La définition s'énonce en termes de conditions sur le gradient et la hessienne de la densité. En effet, sur la ligne de crête, l'un des vecteurs propres de la hessienne de la densité est colinéaire au gradient et les autres sont associés à une valeur propre négative, de sorte qu'on a bien une ligne de crête et non une « vallée ». Notons que le cas limite est celui d'un mode de la densité, où le gradient de la densité s'annule et au voisinage duquel la hessienne est négative.

Cette démarche aboutit à la définition plus formelle suivante.

Définition 11. *Supposons que le vecteur aléatoire* $\mathbf{X}$ *admet une densité* $g_{\mathbf{X}}$ *de classe* C^2 *qui ne s'annule pas. Notons*

$$(\lambda_1(\mathbf{x}), v_1(\mathbf{x})), (\lambda_2(\mathbf{x}), v_2(\mathbf{x})), \ldots, (\lambda_n(\mathbf{x}), v_n(\mathbf{x}))$$

les valeurs propres, supposées distinctes et non nulles, et vecteurs propres, de la matrice hessienne de $g_{\mathbf{X}}$ au point $\mathbf{x}$. *Soit $\mathscr{C}$ l'ensemble des points tels que le gradient de la densité soit orthogonal à $n-1$ vecteurs propres v_i, $i \in I_\perp$ de la hessienne. L'ensemble des points $\mathbf{x} \in \mathscr{C}$ tels que $\lambda_i(\mathbf{x}) < 0$ pour tout $i \in I_\perp$, c'est-à-dire les maxima locaux dans l'espace vectoriel engendré par les v_i, $i \in I_\perp$, forme alors une courbe principale pour le vecteur aléatoire* $\mathbf{X}$.

Cette définition se généralise à des objets de dimension supérieure en notant $\mathscr{C}_d$ l'ensemble des points tels que le gradient de la densité soit orthogonal à $n-d$ vecteurs propres de la hessienne. Par exemple, $\mathscr{C}_0$ est constitué des points critiques de la densité, et une surface principale de dimension 2 peut être obtenue à l'aide de l'ensemble $\mathscr{C}_2$.

Concernant l'estimation de telles courbes, elle peut se faire au moyen d'un estimateur à noyau de la densité. La consistance de cette approche est étudiée par Genovese *et al.* (2012b).

Remarquons que la problématique de l'estimation non paramétrique de filaments, analysée dans un précédent article Genovese *et al.* (2012a), peut être rapprochée de celle des courbes principales. Nous terminons cette section en la présentant brièvement.

3.4.4. Estimation de filaments

Considérons le modèle

$$\mathbf{X}_i = \mathbf{f}(U_i) + \varepsilon_i,\ i = 1, \ldots, n,$$

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.* (2012a) 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, 1998, Chapitre 5)) pour ce problème, Genovese *et al.* (2012a) montrent que l'une de leurs méthodes permet d'atteindre la vitesse correspondante.

3.5. Plusieurs courbes principales

Une courbe principale, dans les définitions que nous avons présentées, correspond dans le cas linéaire à la première composante principale. Cependant, 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 10 constitue un exemple typique de configuration pour laquelle il semble opportun de rechercher plus d'une courbe principale.

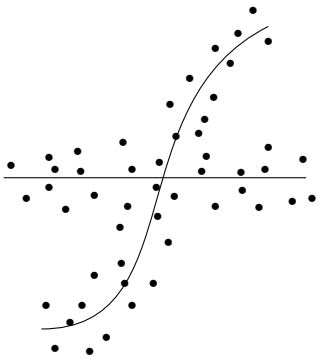

FIGURE 10. *Deux courbes principales pour un ensemble d'observations.*

On peut trouver dans la littérature quelques mentions de cette notion de courbes principales successives.

Duchamp et Stuetzle (1996), qui étudient les courbes principales dans le plan (définition de Hastie et Stuetzle (1989)), constatent que les courbes principales multiples vérifient certaines propriétés analogues à l'orthogonalité des composantes principales.

Proposition 5. *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 (2002), pour traiter le cas des courbes principales multiples, l'algorithme de lignes polygonales de Kégl *et al.* (2000) 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 11, on peut distinguer, 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.

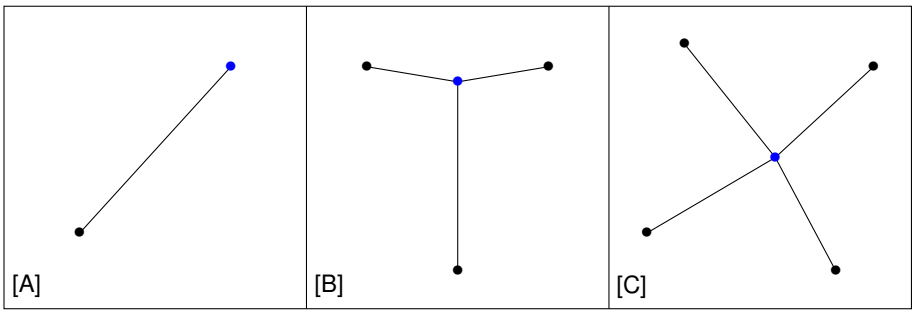

FIGURE 11. *Exemples de types de sommets. [A] Extrémité. [B] Nœud en T. [C] Etoile à 4 branches.*

Remarquons que Delicado (2001) propose une extension de sa définition à plusieurs courbes principales en utilisant une quantité généralisant la variance totale. Einbeck *et al.* (2005b) indiquent que leur algorithme de composantes principales locales permet de trouver plusieurs courbes principales grâce à différents choix d'initialisation. Dans Einbeck *et al.* (2005a), ces auteurs suggèrent de considérer des composantes principales locales correspondant au deuxième axe principal.

3.6. Quelques domaines d'application

Les applications de la notion de courbe principale sont nombreuses et variées. Nous donnons ici un aperçu de cette grande diversité de champs d'application.

La première application, décrite dans l'article originel de Hastie et Stuetzle (1989), 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 (1989)).

Les courbes principales sont utilisées par Kégl et Krzyżak (2002) en squelettisation, tâche importante qui constitue généralement une étape préalable à la reconnaissance optique de caractères. Dans Reinhard et Niranjan (1999), elles interviennent dans une technique de reconnaissance vocale, tandis que Ozertem et Erdogmus (2008) 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 *et al.* (2000)).

D'autres applications ont trait à la géographie ou à la géologie. Ainsi, avec la multiplication des systèmes GPS (*Global Positioning System*), on peut disposer facilement d'observations donnant la trace d'individus qui se déplacent à pied ou dans un véhicule. De telles 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 (2007) est alors de moyenner ces observations au moyen de 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 (2000) pour détecter des failles sismiques ou des champs de mines sur des images de reconnaissance aérienne, et par Banfield et Raftery (1992), comme nous l'avons mentionné plus haut, pour identifier les contours de morceaux de banquise sur des images satellite. Einbeck *et al.* (2005b), quant à eux, analysent à l'aide des courbes principales une zone de plaines inondables, retrouvant les rivières et les vallées correspondantes. Dans Einbeck *et al.* (2005a), 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 *et al.* (2005a) 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 (1999) 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 *et al.* (2004), 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 (2008); Wong *et al.* (2009)). Dans la perspective d'étudier l'impact de produits microbicides sur la transmission du VIH, Caffo *et al.* (2008) 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 (1989) présentent une méthode de comparaison d'estimations de titrage en or ou autres métaux de deux laboratoires différents. Par ailleurs, Zayed et Einbeck (2010) proposent une méthode basée sur les courbes principales destinée à construire à partir de plusieurs sous-indices un nouvel indice économique 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 (2010)). 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 (1995) et Wilson *et al.* (1999)). Enfin, Zhang *et al.* (2006) 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.

4. Conclusion

Après avoir présenté le contexte de la quantification et du *clustering k-means*, puis explicité plusieurs définitions de courbe principale, en donnant un aperçu de leurs applications, résumons, pour conclure, la relation entre ces deux méthodes d'apprentissage non supervisé.

Une table de codage optimale $\mathbf{c}^\star$ et une courbe principale $\mathbf{f}^\star$ peuvent être définies par la minimisation de critères de type moindres carrés très semblables

$$\frac{1}{n}\sum_{i=1}^{n}\min_{j=1,\dots,k}\|\mathbf{X}_i - c_j\|^2, \qquad \frac{1}{n}\sum_{i=1}^{n}\inf_{t\in I}\|\mathbf{X}_i - \mathbf{f}(t)\|^2. \tag{9}$$

Dans le premier cas, on recherche un ensemble de $\ell \leq k$ vecteurs de $\mathbb{R}^d$, dans le second une courbe paramétrée en dimension d de complexité donnée. La différence réside dans le fait que la minimisation apparaissant dans le critère se fait pour le *clustering* sur un ensemble fini $\{1,\dots,k\}$, correspondant aux centres, alors qu'elle a lieu sur un intervalle dans le cas d'une courbe principale, puisqu'elle porte sur le paramètre t réel de la courbe paramétrée.

Dans les deux contextes, les éléments optimaux sont en fait déterminés par un couple : pour la quantification, il s'agit de l'ensemble des vecteurs $\{c_1,\dots,c_\ell\}$ et d'une partition $\{S_1,\dots,S_\ell\}$ de $\mathbb{R}^d$, et pour les courbes principales, de la courbe $\mathbf{f}$ et d'une fonction $t_\mathbf{f}$ de $\mathbb{R}^d$ dans $\mathbb{R}$. Les deux notions peuvent être caractérisées à l'aide d'une version implicite de la définition (9), faisant intervenir une espérance conditionnelle : l'ensemble des centres et la courbe principale doivent vérifier respectivement

$$\forall j,\, c_j = \mathbb{E}[\mathbf{X}|\mathbf{X}\in S_j(\mathbf{c})], \qquad \forall t,\, \mathbf{f}(t) = \mathbb{E}[\mathbf{X}|t_\mathbf{f}(\mathbf{X}) = t]. \tag{10}$$

Si l'on fixe un élément du couple, le second est déterminé de manière similaire dans les deux cadres. D'une part, à table de codage fixée, on connaît la meilleure partition $\{S_1,\dots,S_\ell\}$, qui se trouve être la partition de Voronoi. Celle-ci est construite en affectant un point $\mathbf{x}$ à la cellule S_j si, et seulement si, ce point est plus proche du centre c_j que de tout autre centre. Pour les courbes principales, l'indice de projection est en quelque sorte une généralisation continue de cette partition, puisqu'on associe à un point $\mathbf{x}$ l'indice t lorsque sa projection sur $\mathbf{f}$ correspond à $\mathbf{f}(t)$, ce qui signifie que $\mathbf{x}$ est plus proche de $\mathbf{f}(t)$ que de tout autre point de la courbe paramétrée. D'autre part, la partition de Voronoi ou l'indice de projection étant fixé, les espérances conditionnelles (10) fournissent la table de codage ou la courbe paramétrée correspondante. Cette alternance intervient d'ailleurs dans toute procédure algorithmique itérative destinée à approcher un quantificateur optimal ou une courbe principale.

Annexe A: 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 (1989).

Soumis au Journal de la Société Française de Statistique
File: ArticleTheseSFdSAFischerRevision2.tex, compiled with jsfds, version : 2009/12/09
date: 17 mars 2014

A.1. Définition

Soient $I \subset \mathbb{R}$ un intervalle et $\mathbf{f}$ l'arc paramétré défini par

$$\begin{aligned} \mathbf{f} : I &\to \mathbb{R}^d \\ t &\mapsto (f_1(t), \ldots, f_d(t)), \end{aligned}$$

où $f_1, \ldots, f_d$ sont des fonctions de I dans $\mathbb{R}$. On dit que $\mathbf{f}$ est de classe C^k lorsque les fonctions coordonnées $f_1, \ldots, f_d$ sont de classe C^k.

A.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 12 (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 13 (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 de $\mathbf{f}'$ et $\mathbf{f}''$, dérivées d'ordre 1 et 2.

Définition 14 (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 15 (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*

$$s(t) = \int_{t_0}^{t} \|\mathbf{f}'(u)\| \mathrm{d}u.$$

Soumis au Journal de la Société Française de Statistique
File: ArticleTheseSFdSAFischerRevision2.tex, compiled with jsfds, version : 2009/12/09
date: 17 mars 2014

Remarquons que si $\|\mathbf{f}'(t)\| = 1$ pour tout $t \in I$, alors 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 16 (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*

$$\mathscr{K}(\mathbf{f}, \alpha, \beta) = \int_\alpha^\beta \|\mathbf{f}''(t)\| \mathrm{d}t.$$

Remerciements

Je tiens à remercier l'éditeur, l'éditeur de la rubrique « Etats de l'art », l'éditeur associé, ainsi que les rapporteurs anonymes. Cet article a considérablement gagné en clarté grâce à leurs commentaires avisés et leurs remarques judicieuses.

Références

ABAYA, E. A. et WISE, G. L. (1984). Convergence of vector quantizers with applications to optimal quantization. *SIAM Journal on Applied Mathematics*, 44:183–189.

ALEXANDROV, A. D. et RESHETNYAK, Y. G. (1989). *General Theory of Irregular Curves*. Mathematics and its Applications. Kluwer Academic Publishers, Dordrecht.

ANTOS, A., GYÖRFI, L. et GYÖRGY, A. (2005). Individual convergence rates in empirical vector quantizer design. *IEEE Transactions on Information Theory*, 51:4013–4022.

BANERJEE, A., MERUGU, S., DHILLON, I. S. et GHOSH, J. (2005). Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749.

BANFIELD, J. D. et RAFTERY, A. E. (1992). Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *Journal of the American Statistical Association*, 87:7–16.

BARTLETT, P. L., LINDER, T. et LUGOSI, G. (1998). The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, 44:1802–1813.

BEN-DAVID, S., PÁL, D. et SIMON, H. U. (2007). Stability of k-means clustering. *In* BSHOUTY, N. et GENTILE, C., éditeurs : *Proceedings of the 20th Annual Conference on Learning Theory (COLT 2007)*, pages 20–34. Springer.

BEN-DAVID, S. et von LUXBURG, U. (2008). Relating clustering stability to properties of cluster boundaries. *In* SERVEDIO, R. A. et ZHANG, T., éditeurs : *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pages 379–390, Madison. Omnipress.

BEN-DAVID, S., von LUXBURG, U. et PÁL, D. (2006). A sober look on clustering stability. *In* LUGOSI, G. et SIMON, H. U., éditeurs : *Proceedings of the 19th Annual Conference on Learning Theory (COLT 2006)*, pages 5–19, Berlin. Springer.

BEN-HUR, A., ELISSEEFF, A. et GUYON, I. (2002). A stability based method for discovering structure in clustered data. *In Proceedings of the 7th Pacific Symposium on Biocomputing*, volume 7, pages 6–17.

Soumis au Journal de la Société Française de Statistique
File: ArticleTheseSFdSAFischerRevision2.tex, compiled with jsfds, version : 2009/12/09
date: 17 mars 2014

BIAU, G., DEVROYE, L. et LUGOSI, G. (2008). On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54:781–790.

BIAU, G. et FISCHER, A. (2012). Parameter selection for principal curves. *IEEE Transactions on Information Theory*, 58:1924–1939.

BIRGÉ, L. et MASSART, P. (1997). From model selection to adaptive estimation. *In* POLLARD, D., TORGERSEN, E. et YANG, G., éditeurs : *Festschrift for Lucien Le Cam : Research Papers in Probability and Statistics*, pages 55–87. Springer, New York.

BLANCHARD, G., BOUSQUET, O. et MASSART, P. (2008). Statistical performance of support vector machines. *The Annals of Statistics*, 36:489–531.

BREGMAN, L. M. (1967). 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.

BRUNSDON, C. (2007). Path estimation from GPS tracks. *In Proceedings of the 9th International Conference on GeoComputation, National Centre for Geocomputation, National University of Ireland, Maynooth, Eire*.

CADRE, B. et PARIS, Q. (2012). On hölder fields clustering. *TEST*, 21:301–316.

CAFFO, B. S., CRAINICEANU, C. M., DENG, L. et HENDRIX, C. W. (2008). 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.

CALINSKI, R. B. et HARABASZ, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3:1–27.

CHANG, K. et GHOSH, J. (1998). Principal curves for nonlinear feature extraction and classification. *SPIE Applications of Artificial Neural Networks in Image Processing III*, 3307:120–129.

CHOU, P. A. (1994). The distortion of vector quantizers trained on n vectors decreases to the optimum as $o_p(1/n)$. *In Proceedings of the IEEE International Symposium on Information Theory*, Trondheim, Norway.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

CORKERON, P. J., ANTHONY, P. et MARTIN, R. (2004). 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.

DE'ATH, G. (1999). Principal curves : a new technique for indirect and direct gradient analysis. *Ecology*, 80:2237–2253.

DELICADO, P. (2001). Another look at principal curves and surfaces. *Journal of Multivariate Analysis*, 77:84–116.

DELICADO, P. et HUERTA, M. (2003). Principal curves of oriented points : theoretical and computational improvements. *Computational Statistics*, 18:293–315.

DEMPSTER, A. P., LAIRD, N. M. et RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, pages 1–38.

DEREICH, S. et VORMOOR, C. (2011). The high resolution vector quantization problem with orlicz norm distortion. *Journal of Theoretical Probability*, 24:517–544.

DONG, D. et MCAVOY, T. J. (1995). Nonlinear principal component analysis-based principal curves and neural networks. *Computers and Chemical Engineering*, 20:65–78.

DUCHAMP, T. et STUETZLE, W. (1996). Geometric properties of principal curves in the plane. *In* RIEDER, H., 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.

DUDA, R. O., HART, P. E. et STORK, D. G. (2000). *Pattern Classification*. Wiley-Interscience, New York.

EINBECK, J. et DWYER, J. (2010). Using principal curves to analyse traffic patterns on freeways. *Transportmetrica*.

EINBECK, J., TUTZ, G. et EVERS, L. (2005a). Exploring multivariate data structures with local principal curves. *In* WEIHS, C. et GAUL, W., 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.

EINBECK, J., TUTZ, G. et EVERS, L. (2005b). Local principal curves. *Statistics and Computing*, 15:301–313.

FISCHER, A. (2010). Quantization and clustering with Bregman divergences. *Journal of Multivariate Analysis*, 101:2207–2221.

FISCHER, A. (2011). On the number of groups in clustering. *Statistics & Probability Letters*, 81:1771–1781.

FISCHER, A. (2013). Selecting the length of a principal curve within a Gaussian model. *Electronic Journal of*

Statistics, 7:342–363.

FRIEDSAM, H. et OREN, W. A. (1989). The application of the principal curve analysis technique to smooth beamlines. *In Proceedings of the 1st International Workshop on Accelerator Alignment*.

GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. et WASSERMAN, L. (2012a). The geometry of nonparametric filament estimation. *Journal of the American Statistical Association*, 107:788–799.

GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. et WASSERMAN, L. (2012b). Nonparametric ridge estimation. Available at http://arxiv.org/pdf/1212.5156v1.pdf.

GERBER, S. et WHITAKER, R. (2013). Regularization-free principal curve estimation. *Journal of Machine Learning Research*, 14:1285–1302.

GERSHO, A. et GRAY, R. M. (1992). *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Norwell.

GORDON, A. D. (1999). *Classification*, volume 82 de *Monographs on Statistics and Applied Probability*. Chapman Hall/CRC, Boca Raton.

GRAF, S. et LUSCHGY, H. (1994). Consistent estimation in the quantization problem for random vectors. *In Transactions of the Twelfth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 84–87.

GRAF, S. et LUSCHGY, H. (2000). *Foundations of Quantization for Probability Distributions*. Lecture Notes in Mathematics. Springer-Verlag, Berlin, Heidelberg.

HARDY, A. (1996). On the number of clusters. *Computational Statistics and Data Analysis*, 23:83–96.

HARTIGAN, J. A. (1975). *Clustering Algorithms*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York.

HASTIE, T. (1984). Principal curves and surfaces. Rapport technique, Stanford Linear Accelerator Center.

HASTIE, T. et STUETZLE, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84:502–516.

HASTIE, T., TIBSHIRANI, R. et FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

HERMANN, T., MEINICKE, P. et RITTER, H. (2000). Principle curve sonification. *In Proceedings of the 6th International Conference on Auditory Display Curve Sonification (ICAD2000), Atlanta, USA*, pages 81–86.

HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 498–520.

KAUFMAN, L. et ROUSSEEUW, P. (1990). *Finding Groups in Data : An Introduction to Cluster Analysis*. Wiley Series in Probability and Mathematical Statistics. Wiley-Interscience, Hoboken.

KÉGL, B. (1999). *Principal Curves : Learning, Design, and Applications*. Thèse de doctorat, Concordia University, Montréal, Québec, Canada.

KÉGL, B. et KRZYŻAK, A. (2002). Piecewise linear skeletonization using principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:59–74.

KÉGL, B., KRZYŻAK, A., LINDER, T. et ZEGER, K. (2000). Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:281–297.

KIM, D. J., PARK, Y. W. et PARK, D. J. (2001). A novel validity index for determination of the optimal number of clusters. *IEICE Transactions on Information and System*, E84D:281–285.

KOLTCHINSKII, V. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34:2593–2656.

KRZANOWSKI, W. J. et LAI, Y. T. (1985). A criterion for determining the number of clusters in a data set. *Biometrics*, 44:23–34.

LALOË, T. (2010). L1-quantization and clustering in Banach spaces. *Mathematical Methods of Statistics*, 19:136–150.

LEHMANN, E. L. et CASELLA, G. (1998). *Theory of Point Estimation*. Springer-Verlag, New York.

LEVINE, E. et DOMANY, E. (2002). Resampling method for unsupervised estimation of cluster validity. *Journal of Neural Computation*, 13:2573–2593.

LEVRARD, C. (2013). Fast rates for empirical vector quantization. *Electronic Journal of Statistics*, 7:1716–1746.

LINDE, Y., BUZO, A. et GRAY, R. M. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communication*, 28:801–804.

LINDER, T. (2000). On the training distortion of vector quantizers. *IEEE Transactions on Information Theory*, 46:1617–1623.

LINDER, T. (2002). Learning-theoretic methods in vector quantization. *In* GYÖRFI, L., éditeur : *Principles of*

Nonparametric Learning. Springer-Verlag, Wien.

LINDER, T., LUGOSI, G. et ZEGER, K. (1994). Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Transactions on Information Theory*, 40:1728–1740.

LLOYD, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137.

LUSCHGY, H. et PAGÈS, P. (2002). Functional quantization of gaussian processes. *Journal of Functional Analysis*, 196:486–531.

LUSCHGY, H. et PAGÈS, P. (2006). Functional quantization of a class of brownian diffusions : a constructive approach. *Stochastic Processes and their Applications*, 116:310–336.

MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

MARDIA, K. V., KENT, J. T. et BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, London.

MILLIGAN, G. W. et COOPER, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–79.

OZERTEM, U. et ERDOGMUS, D. (2008). 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.

OZERTEM, U. et ERDOGMUS, D. (2011). Locally defined principal curves and surfaces. *Journal of Machine Learning Research*, 12:1249–1286.

PEARSON, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572.

POLLARD, D. (1981). Strong consistency of *k*-means clustering. *Annals of Statistics*, 9:135–140.

POLLARD, D. (1982a). A central limit theorem for *k*-means clustering. *Annals of Probability*, 10:919–926.

POLLARD, D. (1982b). Quantization and the method of *k*-means. *IEEE Transactions on Information Theory*, 28.

REINHARD, K. et NIRANJAN, M. (1999). Parametric subspace modeling of speech transitions. *Speech Communication*, 27:19–42.

SANDILYA, S. et KULKARNI, S. R. (2002). Principal curves with bounded turn. *IEEE Transactions on Information Theory*, 48:2789–2793.

SHAMIR, O. et TISHBY, N. (2008a). Cluster stability for finite samples. *In* PLATT, J. C., KOLLER, D., SINGER, Y. et ROWSEIS, S., éditeurs : *Advances in Neural Information Processing Systems 20*, pages 1297–1304, Cambridge. MIT Press.

SHAMIR, O. et TISHBY, N. (2008b). Model selection and stability in *k*-means clustering. *In* SERVEDIO, R. A. et ZHANG, T., éditeurs : *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pages 367–378, Madison. Omnipress.

SILVERMAN, B. W. (1985). Some aspects of spline smoothing approaches to non-parametric regression curve fitting. *Journal of the Royal Statistical Society*, 47:1–52.

SPEARMAN, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–293.

STANFORD, D. C. et RAFTERY, A. E. (2000). Finding curvilinear features in spatial point patterns : principal curve clustering with noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:2237–2253.

STEINHAUS, H. (1956). Sur la division des corps matériels en parties. *Bulletin de l'Académie Polonaise des Sciences, Classe III*, IV:801–804.

SUGAR, C. A. et JAMES, G. M. (2003). Finding the number of clusters in a data set : an information theoretic approach. *Journal of the American Statistical Association*, 98:750–763.

TARPEY, T. et FLURY, B. (1996). Self-consistency : a fundamental concept in statistics. *Statistical Science*, 11:229–243.

TIBSHIRANI, R. (1992). Principal curves revisited. *Statistics and Computing*, 2:183–190.

TIBSHIRANI, R., WALTHER, G. et HASTIE, T. (2001). Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society*, 63:411–423.

VERBEEK, J. J., VLASSIS, N. et KRÖSE, B. (2001). A soft *k*-segments algorithm for principal curves. *In Proceedings of International Conference on Artificial Neural Networks 2001*, pages 450–456.

WILSON, D. J. H., IRWIN, G. W. et LIGHTBODY, G. (1999). RBF principal manifolds for process monitoring. *IEEE Transactions of Neural Networks*, 10:1424–1434.

WONG, W. C. K. et CHUNG, A. C. S. (2008). 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.

WONG, W. C. K., SO, R. W. K. et CHUNG, A. C. S. (2009). Principal curves : a technique for preliminary carotid lumen segmentation and stenosis grading. *MIDAS Journal*.

XU, L. et JORDAN, M. I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8:129–151.

ZAYED, M. et EINBECK, J. (2010). Constructing economic summary indexes via principal curves. *In* LECHEVALLIER, Y. et SAPORTA, G., éditeurs : *Proceedings of the 19th International Conference on Computational Statistics, COMPSTAT 2010*, pages 1709–1716. Springer.

ZHANG, F., WU, B., ZHANG, L., HUANG, H. et TIAN, Y. (2006). 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.