Journal de la Société Française de Statistique
Vol. 156 No. 1 (2015)

On two extensions of the vector quantization scheme *

Titre: Sur deux extensions du principe de quantification vectorielle.

Aurélie Fischer[1]

Abstract: In this paper, we present results pertaining to two different extensions of vector quantization and the related question of $k-$means clustering. The first part of the paper is about the theoretical performance of quantization and clustering with Bregman divergences. The second one is dedicated to model selection issues for principal curves. Some numerical illustrations are provided in each case.

Résumé : Dans cet article, nous présentons des résultats relatifs à deux extensions différentes de la quantification vectorielle et de la question liée de classification par la méthode des centres mobiles. La première partie de l'article concerne la performance théorique de la quantification et du clustering avec des divergences de Bregman ; la seconde est dédiée à des problèmes de sélection de modèle pour les courbes principales. Chaque partie est complétée par quelques illustrations numériques.

Keywords: quantization, clustering, Bregman divergences, principal curves, model selection
Mots-clés : quantification, classification non supervisée, divergences de Bregman, courbes principales, sélection de modèle
AMS 2000 subject classifications: 62G05, 62G08, 62H30

1. Introduction

Let X be a random variable with distribution μ taking its values in a set $\mathscr{X}$, and $X_1,\ldots,X_n$ a sample of independent copies of X. In this paper, we are interested in procedures based on the minimization of a criterion of the form

$$\Delta(\psi) = \mathbb{E}d(X,\psi(X)). \tag{1}$$

Here, $\mathbb{E}$ denotes expectation with respect to μ, ψ is a measurable mapping from $\mathscr{X}$ to a subset $\mathscr{U}$ of $\mathscr{X}$, and d is some dissimilarity measure. For instance, when $\mathscr{X} = \mathbb{R}^d$, equipped with the Euclidean norm, we can set $d(x,y) = \|x-y\|^r$, $r \geq 1$, the most common case $r = 2$ leading to the squared Euclidean distance. When $\mathscr{U}$ is discrete, the quantity (1) is the distortion used in quantization, whereas the situation where $\mathscr{U}$ is a one-dimensional structure corresponds to principal curves estimation. In practice, μ is unknown, which motivates the introduction of the empirical counterpart of $\Delta(\psi)$

$$\Delta_n(\psi) = \frac{1}{n}\sum_{i=1}^{n} d(X_i,\psi(X_i)).$$

* The author was supported by the ANR project TopData ANR-13-BS01-0008.
[1] Laboratoire de Probabilités et Modèles Aléatoires, Université Paris Diderot.
E-mail: aurelie.fischer@univ-paris-diderot.fr

 ISSN: 2102-6238

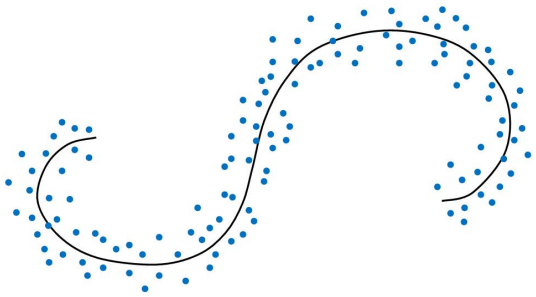


FIGURE 1. *An example of principal curve.*

This is $\Delta(\psi)$ for the empirical measure μ_n associated with $X_1,\dots,X_n$, given by $\mu_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i\in A\}}$ for any Borel subset A of $\mathscr{X}$.

Quantization is the problem of replacing data by an efficient and compact representation. For a fixed integer $k \geq 1$, the random variable X is represented by $q(X)$, where the so-called k-quantizer q maps $\mathscr{X}$ to a finite subset with at most k elements $c_1,\dots,c_\ell$, $\ell \leq k$. Every k-quantizer is characterized by its codebook $\mathbf{c} = \{c_1,\dots,c_\ell\}$ and the partition $S_1,\dots,S_\ell$ defined by $S_j = \{x \in \mathscr{X}, q(x) = c_j\}$. The error committed is given by the distortion $\Delta(q)$. For more information on quantization, we refer the reader to Gersho and Gray (1992), Graf and Luschgy (2000) and Linder (2002). A related problem consists in grouping data items in meaningful classes by minimizing the empirical distortion $\Delta_n(q)$ over all possible k-quantizers. The aim is to find a data-based quantizer q_n such that the clustering risk $\Delta(q_n)$ gets close to the optimal risk $\Delta^\star := \inf_q \Delta(q)$ as the size of the data set grows. It has been shown by Banerjee et al. (2005b) that the standard k-means clustering algorithm (see for instance Lloyd, 1982), where d is the Euclidean distance, generalizes to general Bregman divergences, which are dissimilarity measures defined for a strictly convex function ϕ by

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

where $D_y\phi$ denotes the Fréchet derivative of ϕ at y.

As for principal curves, they are parameterized curves in $\mathbb{R}^d$, i.e. continuous functions

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

where $I = [a,b]$ is a closed interval of the real line, passing "through the middle" of a probability distribution or a set of observations, as illustrated in Figure 1.

A principal curve for X is a parameterized curve $f(t)$ which is self-consistent, that is

$$f(t) = \mathbb{E}[X|t_f(X) = t], \quad t \in I. \tag{2}$$

Here, the so-called projection index $t_f(x)$ is defined by

$$t_f(x) = \max\left\{t \in I, \|x - f(t)\| = \min_{t'\in I} \|x - f(t')\|\right\}, \tag{3}$$

 ISSN: 2102-6238

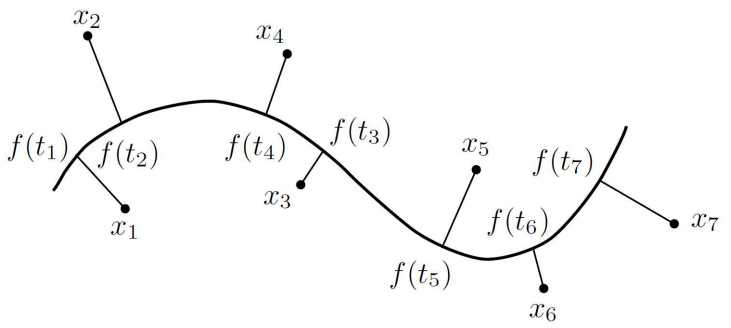


FIGURE 2. *The projection index t_f. For all i, t_i stands for $t_f(x_i)$.*

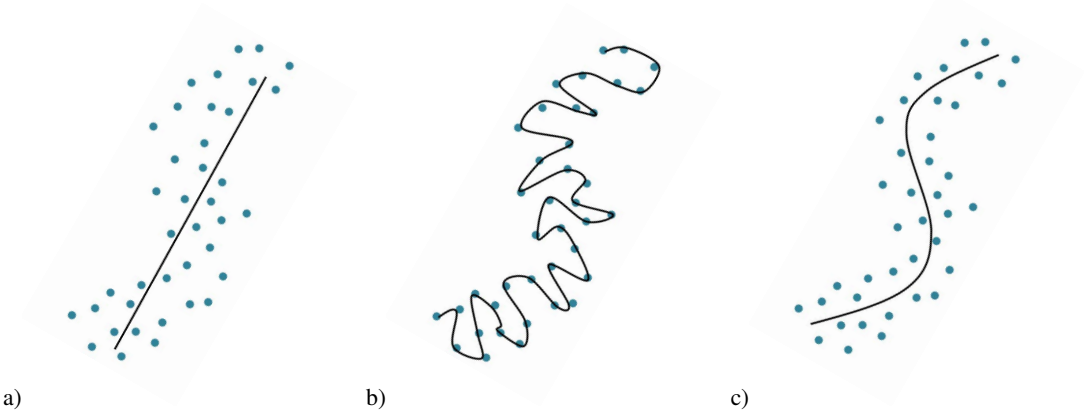


FIGURE 3. *Different results depending on the parameters of the curve. a) Too rough. b) Interpolation. c) Appropriate.*

so that $t_f(x)$ is the largest real number t minimizing the Euclidean distance between x and $f(t)$, as shown in Figure 2. The self-consistency property may be interpreted by saying that each point of the curve f is the mean of the observations projecting on f around this point. This original definition of a principal curve is due to Hastie and Stuetzle (1989). The implicit formulation may be avoided by considering the minimization of

$$\Delta(\varphi) = \mathbb{E}\|X - \varphi(X)\|^2, \quad \Delta_n(\varphi) = \frac{1}{n}\sum_{i=1}^{n} \|X_i - \varphi(X_i)\|^2,$$

as proposed by Kégl et al. (2000) (see also Sandilya and Kulkarni, 2002). Here, φ maps $\mathbb{R}^d$ to some one-dimensional structure.

In order to build a satisfactory principal curve, some parameters have to be chosen, as illustrated in Figure 3, to achieve a trade-off between closeness to the data and smoothness.

Note that principal curves may be seen as a generalization of Principal Component Analysis, searching for a curve instead of a direction of maximal variation.

For more details on principal curves, other definitions and applications, the reader may see for instance Fischer (2014) and references collected in that survey.

 ISSN: 2102-6238

The aim of this document is to present results pertaining to both situations described above. The paper is organized as follows. First, we will focus on theoretical properties of clustering and quantization with Bregman divergences as dissimilarity measures. Then, we will study model selection issues for principal curves. In both situations, simulations or real data examples are proposed as illustrations. To keep this article a reasonable length, the reader will be referred to the appropriate papers for the complete proofs, not presented here, of the different results.

2. Quantization with Bregman divergences

Bregman divergences are a broad class of dissimilarity measures indexed by strictly convex functions. Introduced by Bregman (1967), they are useful in a wide range of areas, among which statistical learning and data mining (Banerjee et al., 2005b; Cesa-Bianchi and Lugosi, 2006), computational geometry (Nielsen et al., 2007), natural sciences, speech processing and information theory (Gray et al., 1980). Squared Euclidean, Mahalanobis, Kullback-Leibler and L^2 distances are particular cases of Bregman divergences. In $\mathbb{R}^d$, a Bregman divergence d_ϕ has the form

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

where $\langle\cdot,\cdot\rangle$ denotes the standard inner product, and $\nabla\phi(y)$ the gradient of ϕ at y. For example, taking for ϕ the squared Euclidean norm gives back the squared Euclidean distance. The same definition is valid in Hilbert spaces, and generalizes to Banach spaces by setting

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

with $D_y\phi$ the Fréchet derivative of ϕ at y (Alber and Butnariu, 1997, Frigyik et al., 2008b; see also Jones and Byrne, 1990 and Csiszár, 1995). A Bregman divergence is not necessary a true metric, since it may be asymmetric or fail to satisfy the triangle inequality. However, Bregman divergences fulfill an interesting projection property which generalizes the Hilbert projection on a closed convex set, as shown in Bregman (1967) for the finite-dimensional setting and Alber and Butnariu (1997) for the functional case. Moreover, Banerjee et al. (2005b) have established that there exists a relation between finite-dimensional Bregman divergences and exponential families. Although they are not true metrics, Bregman divergences satisfy some properties, such as non-negativity and separation, convexity in the first argument and linearity (see Bregman, 1967, Nielsen et al., 2007 and Frigyik et al., 2008a for a complete description and proofs of these properties). Table 1 collects the most common examples of Bregman divergences. As Bregman divergences represent a natural tool to measure proximity between observations of complex nature and infinite-dimensional objects, such as curves or probability measures, we use them for quantization and clustering purposes. The results stated in this section are proved in Fischer (2010).

2.1. *Looking for an optimal quantizer*

Let E be either $\mathbb{R}^d$, or an infinite-dimensional Banach space, reflexive and separable, and $\mathscr{C}$ denote a measurable convex subset of E. Recall that the relative interior of $\mathscr{C}$, denoted

 ISSN: 2102-6238

TABLE 1. *Some examples of Bregman divergences. The matrix A is supposed to be positive definite. The notation $L^2(I,m)$ stands for the set of square integrable functions on an interval $I \subset \mathbb{R}$, with respect to the positive measure m, $L^2_{2\pi}(dt)$ for the set of 2π-periodic square integrable functions, $C^0([0,1])$ denotes the set of continuous functions on $[0,1]$, and $C^0_{2\pi}$ the set of 2π-periodic continuous functions.*

Bregman divergence	E	$\mathscr{C}$
Squared loss	$\mathbb{R}$	$\mathbb{R}$
Exponential loss	$\mathbb{R}$	$\mathbb{R}$
Norm-like	$\mathbb{R}$	$\mathbb{R}^+$
Generalized K-L (dim 1)	$\mathbb{R}$	$\mathbb{R}^+$
Logistic loss	$\mathbb{R}$	$[0,1]$
Itakura-Saito (dim 1)	$\mathbb{R}$	$(0,+\infty)$
Squared Euclidean distance	$\mathbb{R}^d$	$\mathbb{R}^d$
Mahalanobis distance	$\mathbb{R}^d$	$\mathbb{R}^d$
Kullback-Leibler (discrete)	$\mathbb{R}^d$	$(d-1)-$simplex
Generalized K-L (discrete)	$\mathbb{R}^d$	$(\mathbb{R}^+)^d$
Squared L^2 norm	$L^2(I,m)$	$L^2(I,m)$
Kullback-Leibler (continuous)	$L^2([0,1],dt)$	$\{x \in C^0([0,1]), \int_0^1 x(t)dt = 1\}$
Generalized K-L (continuous)	$L^2([0,1],dt)$	$\{x \in C^0([0,1]), x \geq 0\}$
Itakura-Saito (continuous)	$L^2_{2\pi}(dt)$	$\{x \in C^0_{2\pi}, x > 0\}$

Bregman divergence	$\phi(x)$	$d_\phi(x,y)$
Squared loss	x^2	$(x-y)^2$
Exponential loss	e^x	$e^x - e^y - (x-y)e^y$
Norm-like	x^α	$x^\alpha + (\alpha-1)y^\alpha - \alpha x y^{\alpha-1}$
Generalized K-L (dim 1)	$x\ln x$	$x\ln\frac{x}{y} - (x-y)$
Logistic loss	$x\ln x + (1-x)\ln(1-x)$	$x\ln\frac{x}{y} + (1-x)\ln\left(\frac{1-x}{1-y}\right)$
Itakura-Saito (dim 1)	$-\ln x$	$\frac{x}{y} - \ln\frac{x}{y} - 1$
Squared Euclidean distance	$\|x\|_2^2$	$\|x-y\|_2^2$
Mahalanobis distance	${}^t xAx$	${}^t(x-y)A(x-y)$
Kullback-Leibler (discrete)	$\sum_{\ell=1}^d x_\ell \ln x_\ell$	$\sum_{\ell=1}^d x_\ell \ln\frac{x_\ell}{y_\ell}$
Generalized K-L (discrete)	$\sum_{\ell=1}^d x_\ell \ln x_\ell$	$\sum_{\ell=1}^d x_\ell \ln\frac{x_\ell}{y_\ell} - \sum_{\ell=1}^d (x_\ell - y_\ell)$
Squared L^2 norm	$\int_I x^2(t)dm(t)$	$\|x-y\|_{L^2}^2$
Kullback-Leibler (continuous)	$\int_0^1 x(t)\ln x(t)dt$	$\int_0^1 x(t)\ln\frac{x(t)}{y(t)}dt$
Generalized K-L (continuous)	$\int_0^1 x(t)\ln x(t)dt$	$\int_0^1 x(t)\ln\frac{x(t)}{y(t)} + y(t) - x(t)dt$
Itakura-Saito (continuous)	$-\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln(x(\theta))d\theta$	$-\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\ln\frac{x(\theta)}{y(\theta)} - \frac{x(\theta)}{y(\theta)} + 1\right)d\theta$

 ISSN: 2102-6238

hereafter by $ri(\mathscr{C})$, is its interior with respect to the closed affine hull. We will write $\partial\mathscr{C}$ for the complement of $ri(\mathscr{C})$ in its closure $\overline{\mathscr{C}}$.

We are looking for the best possible quantizer for the random variable X taking its values in $\mathscr{C}$. Throughout the section, the following assumptions are made:

1. $\mathbb{E}\|X\| < +\infty$.
2. $\mathbb{E}X \in ri(\mathscr{C})$.
3. $\mathbb{E}|\phi(X)| < +\infty$ and, for all $c \in ri(\mathscr{C})$, $\mathbb{E}|D_c\phi(X)| < +\infty$.

This last requirement implies in particular that $\mathbb{E}d_\phi(X,c) < +\infty$ for all $c \in ri(\mathscr{C})$.

First, as in the Euclidean case (see, e.g., Linder, 2002), it is easy to show that among all quantizers with same codebook, the best one (with respect to the distortion $\Delta(q)$) is the nearest neighbor quantizer, whose partition $S_1,\dots,S_\ell$ is the Voronoi partition, i.e.,

$$S_1 = \{x \in \mathscr{C}, d_\phi(x,c_1) \leq d_\phi(x,c_p), p = 1,\dots,\ell\},$$
$$S_j = \{x \in \mathscr{C}, d_\phi(x,c_j) \leq d_\phi(x,c_p), p = 1,\dots,\ell\} \setminus \bigcup_{m=1}^{j-1} S_m, \quad j = 2,\dots\ell.$$

If an optimal quantizer exists, it is necessarily a nearest neighbor quantizer. Hence, in the sequel, we will always consider nearest neighbor quantizers and minimize the distortion over the codebook $\mathbf{c}$

$$\Delta(\mathbf{c}) = \mathbb{E}\min_{j=1,\dots,k} d_\phi(X,c_j), \quad \Delta_n(\mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\dots,k} d_\phi(X_i,c_j).$$

Conversely, given a partition $\{S_j\}_{j=1}^{\ell}$, with $\mu(S_j) > 0$ and $\mathbb{E}[X|X \in S_j] \in ri(\mathscr{C})$ for $j = 1,\dots,\ell$, the best quantizer is obtained by setting

$$c_j \in \arg\min_{c \in ri(\mathscr{C})} \mathbb{E}[d_\phi(X,c)|X \in S_j] \quad \text{for } j = 1,\dots,\ell.$$

Moreover, if S is a Borel subset of $\mathscr{C}$ with $\mu(S) > 0$ and $\mathbb{E}[X|X \in S] \in ri(\mathscr{C})$, the function $c \mapsto \mathbb{E}[d_\phi(X,c)|X \in S]$ reaches its infimum at a unique element of $ri(\mathscr{C})$, namely $\mathbb{E}[X|X \in S]$ (result due to Banerjee et al., 2005a in the finite-dimensional case). Thus, for every Bregman divergence, the minimizer is the conditional expectation, just like for the squared Euclidean distance. For example, when the distortion measure is an L^1 norm, it is the median instead of the expectation.

From an algorithmic point of view, the combination of the optimality of the conditional expectation and of the Voronoi partition shows that the k-means algorithm can be used to design an approximate minimizer in practice.

On the theoretical side, the existence of a minimum rests upon a compactness argument. Since E is reflexive, a closed and bounded convex subset of E is compact for the weak topology $\sigma(E,E')$, that is the coarsest topology on E making all continuous linear forms on E continuous. Moreover, every weakly lower semi-continuous function reaches its minimum on a weakly compact set. Thus, if we know in advance that $\mathbf{c}^\star$ is to be searched for in a closed and bounded convex set, an argument of continuity suffices to show the existence

 ISSN: 2102-6238

of $\mathbf{c}^\star$. In the sequel, $\mathscr{C}_R \subset ri(\mathscr{C})$ will denote a closed and bounded convex set of diameter $2R$. For example, $\mathscr{C}_R = B(0,R) = \{x \in E, \|x\| \leq R\}$ the closed ball of center 0 and radius R. A key fact is that $X \in \mathscr{C}_R$ implies, by Bregman projection (Alber and Butnariu, 1997), that $\mathbf{c}^\star \in \mathscr{C}_R$ if it exists.

Théorème 1. *Suppose that there exists $R > 0$ such that $\mathbb{P}(X \in \mathscr{C}_R) = 1$, and that for all $x \in \mathscr{C}$, $y \mapsto d_\phi(x,y)$ is weakly lower semi-continuous on $\mathscr{C}_R$. Then, there exists an optimal quantizer.*

Note that convex functions which are lower semi-continuous for the norm are weakly lower semi-continuous.

In the particular case where the convex set $\mathscr{C}$ lies in a finite-dimensional affine space, the result may be proved under weaker assumptions (see Fischer, 2010). Moreover, since the weak topology coincides with the norm topology in finite dimension, the term "weakly" in Theorem 1 can be dropped.

In fact, if we only have $\mathscr{C}_R \cap ri(\mathscr{C}) \neq \emptyset$ instead of $\mathscr{C}_R \subset ri(\mathscr{C})$, but ϕ is of Legendre type (see Rockafellar, 1970, and for the infinite-dimensional definition, Bauschke et al., 2001), it remains possible to use Bregman projection to obtain the same result.

In the particular case of a squared Hilbert distance, it can be shown that it is sufficient to look for an optimal quantizer on a ball. Hence the existence result follows directly.

Since the support of the empirical measure μ_n contains at most n points, it is included in a closed ball B_R. Thus, Theorem 1 implies the existence of a minimizer $\mathbf{c}_n^\star$ of the empirical distortion.

2.2. Convergence of the distortion

Suppose that there exists an optimal codebook $\mathbf{c}_n^\star$. We would like that $\Delta(\mathbf{c}_n^\star)$ gets close to the optimal distortion as the number n of observations grows. Assuming that $\mathbf{c}^\star$ exists, if $\mathbf{c}_n^\star$ and $\mathbf{c}^\star$ belong to $\mathscr{C}_R^k$,

$$\Delta(\mathbf{c}_n^\star) - \Delta^\star \leq 2 \sup_{\mathbf{c} \in \mathscr{C}_R^k} |\Delta_n(\mathbf{c}) - \Delta(\mathbf{c})|.$$

Yet, it can be proved under appropriate assumptions that $\sup_{\mathbf{c} \in \mathscr{C}_R^k} |\Delta_n(\mathbf{c}) - \Delta(\mathbf{c})|$ vanishes as n tends to infinity, so that the next theorem holds.

Théorème 2. *Assume that for all $x \in \mathscr{C}$, $y \mapsto d_\phi(x,y)$ is weakly lower semi-continuous, so that there exists a minimizer $\mathbf{c}_n^\star$ of the empirical distortion. If there exists $R > 0$ such that $\mathbb{P}(X \in \mathscr{C}_R) = 1$, and $M = M(\phi, R) \geq 0$ such that, for all $c \in \mathscr{C}_R$, $\|D_c\phi\| \leq M$, then*

$$\lim_{n \to +\infty} \Delta(\mathbf{c}_n^\star) = \Delta^\star \quad a.s., \qquad \lim_{n \to +\infty} \mathbb{E}\Delta(\mathbf{c}_n^\star) = \Delta^\star.$$

Note that these convergence results always hold when $\phi(\cdot) = \|\cdot\|^2$ (Biau et al., 2008). Besides, as above, assumptions could be relaxed in the finite-dimensional setting.

Let us now discuss some examples.

Example 2.1. 1. *Generalized K-L, dimension 1.* Here $E = \mathbb{R}$, $\mathscr{C} = \mathbb{R}^+$ and $d_\phi(x,y) = x \ln \frac{x}{y} - (x-y)$. Let $x \in \mathscr{C}$. Because the map $y \mapsto x \ln \frac{x}{y} - (x-y)$ is continuous and

 ISSN: 2102-6238

convex on $ri(\mathscr{C}) = (0, +\infty)$ (its second derivative is $\frac{x}{y^2} \geq 0$) and tends to $+\infty$ as y tends to 0 or $+\infty$, there exist optimal and empirically optimal quantizers. If $\mathbf{c}_n^\star$ is a minimizer of Δ_n, almost sure convergence of $\Delta(\mathbf{c}_n^\star)$ to $\Delta^\star$ is ensured.

2. *Exponential loss.* Let $E = \mathscr{C} = \mathbb{R}$ and $\phi(x) = e^x$, which yields $d_\phi(x,y) = e^x - e^y - (x-y)e^y$. The function $y \mapsto e^x - e^y - (x-y)e^y$ is continuous on $\mathbb{R}$. If $\mathbb{P}(|X| \leq R) = 1$, there exists an optimal quantizer, and since $\phi'(x) = e^x \leq e^R$ on $[-R,R]$, $\Delta(\mathbf{c}_n^\star)$ converges almost surely and in L^1 to $\Delta^\star$.
3. *Squared Euclidean distance.* In this particular case, existence of an optimal quantizer, almost sure and L^1 convergence of the distortion are guaranteed.
4. *Kullback-Leibler, discrete probability measures.* Here, $E = \mathbb{R}^d$, $\mathscr{C}$ is the $(d-1)$-simplex and $d_\phi(p,q) = \sum_{\ell=1}^d p_\ell \ln \frac{p_\ell}{q_\ell}$. The fact that the function $q = (q_1, \ldots, q_d) \mapsto \sum_{\ell=1}^d p_\ell \ln \frac{p_\ell}{q_\ell}$ is continuous and convex on $ri(\mathscr{C}) = \{(p_1, \ldots, p_d) \in (0,+\infty)^d, \sum_{\ell=1}^d p_\ell = 1\}$, and tends to $+\infty$ as one of the q_ℓ's tends to 0, ensures that there exists an optimal quantizer and we have almost sure convergence of the distortion.
5. *Squared L^2 distance.* Let $E = \mathscr{C} = L^2([0,1], dt)$, and $d_\phi(x,y) = \int_0^1 (x(t) - y(t))^2 dt$. This is a Hilbert norm, thus existence of a minimizer of the distortion and convergence are ensured.
6. *Generalized K-L.* Let $E = L^2([0,1], dt)$ and let $\mathscr{C}$ be the set of all continuous non-negative elements of E. Here $d_\phi(p,q) = \int_0^1 [p(t) \ln \frac{p(t)}{q(t)} + q(t) - p(t)] dt$. The map $q \mapsto d_\phi(p,q)$ is continuous and convex and therefore weakly semi-continuous. Assume that $\mathbb{P}(r \leq \|X\| \leq R) = 1$ $(r > 0)$. Then, there exists an optimal quantizer. Moreover, we have almost sure and L^1 convergence of the distortion.

As for rates of convergence, the following result is obtained.

Théorème 3. *Suppose that E is a type 2 Banach space with constant T_2, and that, for all $x \in \mathscr{C}$, $y \mapsto d_\phi(x,y)$ is weakly lower semi-continuous, which ensures the existence of an optimal codebook $\mathbf{c}_n^\star$. Assume that there exists $R > 0$ such that $\mathbb{P}(X \in \mathscr{C}_R) = 1$. If $|-\phi(c) + D_c\phi(c)|$ and $\|D_c\phi\|$ are uniformly bounded on $\mathscr{C}_R$ by $M_1 = M_1(\phi, R) \geq 0$ and $M_2 = M_2(\phi, R) \geq 0$ respectively, then*

$$\mathbb{E}\Delta(\mathbf{c}_n^\star) - \Delta^\star \leq \frac{4k}{\sqrt{n}} \left(M_1 + T_2 M_2 (\mathbb{E}\|X\|^2)^{1/2} \right),$$

and thus

$$\mathbb{E}\Delta(\mathbf{c}_n^\star) - \Delta^\star \leq \frac{4k}{\sqrt{n}} (M_1 + T_2 M_2 R).$$

Note that Theorem 3 yields dimension-free upper bounds.

Example 2.2. In this example, we give bounds obtained for some usual Bregman divergences. We assume throughout that there exists $R > 0$ such that $\mathbb{P}(\|X\| \leq R) = 1$.

1. *Squared loss.* For $\phi(x) = x^2$, $\mathbb{E}\Delta(\mathbf{c}_n^\star) - \Delta^\star \leq \frac{4k}{\sqrt{n}} \left(R^2 + 2R(\mathbb{E}|X|^2)^{1/2} \right) \leq \frac{12kR^2}{\sqrt{n}}$.
2. *Exponential loss.* For $\phi(x) = e^x$, $\mathbb{E}\Delta(\mathbf{c}_n^\star) - \Delta^\star \leq \frac{4k(2R-1)e^R}{\sqrt{n}}$.
3. *Squared Euclidean.* For $\phi(x) = \|x\|^2$, $\mathbb{E}\Delta(\mathbf{c}_n^\star) - \Delta^\star \leq \frac{12kR^2}{\sqrt{n}}$.

 ISSN: 2102-6238

4. *Mahalanobis.* For $\phi(x) = {}^t xAx$, A positive definite, $\mathbb{E}\Delta(\mathbf{c}_n^\star) - \Delta^\star \leq \frac{12k\|A\|R^2}{\sqrt{n}}$.

5. *Squared* L^2. When ϕ is a squared L^2 norm, $\mathbb{E}\Delta(\mathbf{c}_n^\star) - \Delta^\star \leq \frac{12kR^2}{\sqrt{n}}$.

Note that some Bregman divergences, typically Kullback-Leibler, involve a logarithm, which prevents $||D_c\phi||$ from being uniformly bounded on a ball B_R. In order to circumvent this difficulty, a possible solution is to consider a class of elements of E satisfying the following assumption:

– In dimension 1, $0 < r \leq X \leq R < +\infty \quad a.s.$
– In dimension d $(2 \leq d \leq +\infty)$, when the logarithm appears in a sum or an integral, $\sum_{\ell=1}^{d} \ln^2(x_\ell) \leq M(R)$ or $\int \ln^2(x(t))dt \leq M(R)$.

Several such conditions can be found in the literature on Kullback-Leibler divergence. For instance, Jordan et al. (2010), who develop an estimation method for the Kullback-Leibler divergence, require an envelope condition or boundedness from above and below.

As an illustration, let $d_\phi(x,y) = \int_0^1 x(t) \ln \frac{x(t)}{y(t)} dt$. Suppose that $\mathbb{P}(\|X\| \leq R) = 1$ for some $R > 0$ and that $\int_0^1 \ln^2(X(t))dt \leq R^2$. Assuming that the codebooks belong to the same function class as X, we obtain

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

2.3. Simulations

Now, we present some clustering results obtained with different Bregman divergences. These simulations have been carried out with the software R.

To assess the quality of the clustering, we use a correlation coefficient between partitions proposed by Strehl and Ghosh (2002), called normalized mutual information. Let S and S' be two partitions of the observations. Denoting by n_j (respectively n'_ℓ) the number of data points in S_j (respectively S'_ℓ) and by $n_{j,\ell}$ the number of points in S_j and S'_ℓ, normalized mutual information is given by

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

This indicator allows to compare partitions obtained by Bregman divergence clustering with an "expected partition": the closer to 1 the coefficient the better the result. Note that comparing the distortions obtained for several divergences does not provide a reliable indicator of the quality of the partitions, since the value of the distortion intrinsically depends on the divergence chosen: a larger error might be associated to a better partition.

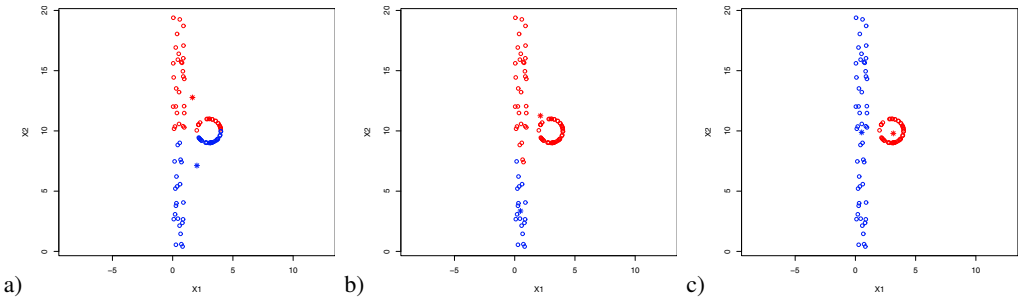
We present first examples in dimension 1, then in the plane $\mathbb{R}^2$, and finally, in the infinite-dimensional setting.

Gaussian, binomial and Poisson distributions As mentioned at the beginning of the section, there is a relationship between exponential families and Bregman divergences. As a first illustration, let us compare the partitions obtained for Gaussian, binomial and Poisson distributions, using the corresponding Bregman divergences, which are Euclidean distance,

 ISSN: 2102-6238

TABLE 2. *Normalized mutual information and number of cases where each divergence gives the best result (100 trials).*

	Euclidean	Logistic	Generalized K-L
Gaussian	**0.689** (**52**)	0.685 (42)	0.672 (35)
Binomial	0.791 (38)	**0.813** (**62**)	0.806 (57)
Poisson	0.702 (37)	0.728 (56)	**0.732** (**63**)



Euclidean: 0.302	Generalized K-L: 0.327	Logistic: 0.320	Itakura-Saito: **0.986**

FIGURE 4. *Clustering of uniform data on strip and circle ($k = 2$, $n = 100$). a) Euclidean. b) Generalized K-L. c) Itakura-Saito. Normalized mutual information table (50 trials).*

logistic loss, and generalized Kullback-Leibler divergence. For each distribution, 3 groups of 30 observations, centered in 10, 20 and 40 respectively, were generated. Setting the variance to 25 for the Gaussian distribution, and the number of trials to 100 for the binomial provides three models with similar variance. Negative realizations are discarded so that logistic loss and Kullback-Leibler divergence are well defined.

Table 2 presents average normalized mutual information over 100 trials, as well as the number of times where each divergence leads to the best partition (in the large sense: the sum of the values in a line is larger than 100). This example illustrates the fact that Euclidean distance is best suited to Gaussian, logistic loss to binomial, and Kullback-Leibler to Poisson data.

Strip and circle We are interested in clustering 50 observations uniformly distributed on a circle with center $(3, 10)$ and radius 1, and 50 observations on a rectangle of height 20, between the lines $x = 0$ et $x = 1$. The clustering results for different Bregman divergences and normalized mutual information over 50 trials are shown in Figure 4.

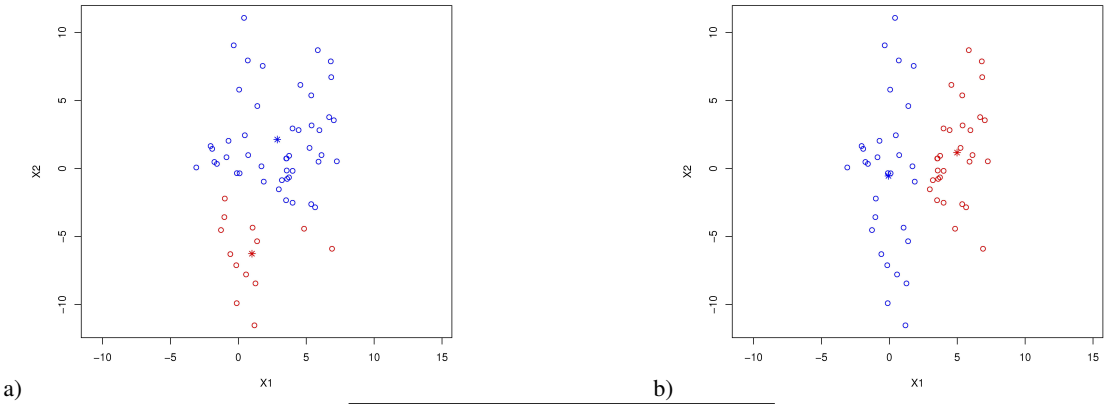
Note for instance that the Itakura-Saito divergence, which is neither symmetric nor convex in the second variable, allows to separate the strip and the circle, whereas the other Bregman divergences cut the data in a completely different manner.

Data on the simplex We simulated 45 observations on the 2-simplex from a Dirichlet distribution. Let us recall that a Dirichlet distribution with parameters (a_1, a_2, a_3), where

 ISSN: 2102-6238

TABLE 3. *Normalized mutual information (100 trials).*

Euclidean: 0.674	Kullback-Leibler: **0.714**	Logistic: 0.689	Itakura-Saito: 0.673



Euclidean: 0.445	Mahalanobis: **0.793**

FIGURE 5. *Clustering Gaussian data ($k = 2$, $n = 60$). a) Euclidean. b) Mahalanobis. Normalized mutual information (100 trials).*

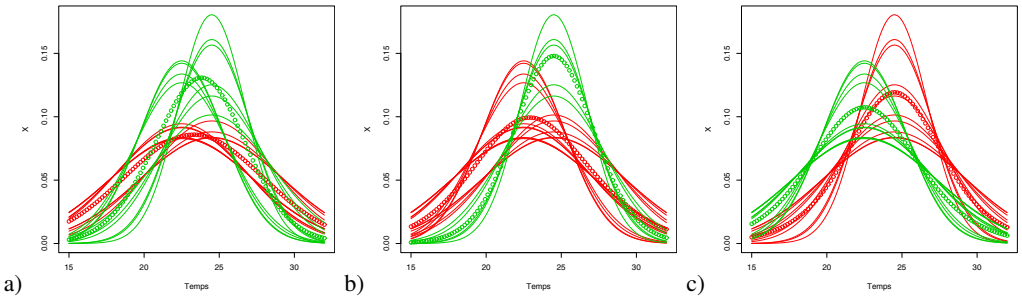
$a_i > 0$ for every $i = 1, \ldots, 3$, is given by

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

where $p_i > 0$ for $i = 1, \ldots, 3$, $p_1 + p_2 + p_3 = 1$ (proportions), and Γ denotes the function defined, for $x > 0$, by $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} \mathrm{d}t$. Here, 3 groups of 15 observations following respectively the Dirichlet distribution with parameters $(10, 10, 2)$, $(5, 5, 5)$ and $(2, 2, 10)$ were generated. Table 3 indicates that the Kullback-Leibler divergence is the most appropriate one. This result is consistent with the common use of this divergence in documents classification (Banerjee et al., 2005b). Indeed, our simulated observations may be interpreted as a very simple text classification problem based on 3 words or expressions, the distribution parameters corresponding to their average frequency.

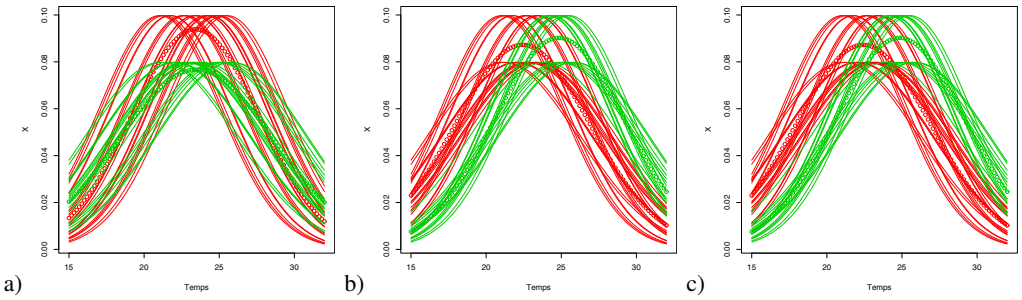
Euclidean and Mahalanobis distance Clustering the observations represented in Figure 5 with the square Euclidean distance or with the Mahalanobis distance with $A = \begin{pmatrix} 2 & 1 \\ 1 & 8 \end{pmatrix}^{-1}$ leads to very different groups. Indeed, these two ellipses are generated from Gaussian vectors with covariance matrix A^{-1}. We recover the fact that the best Mahalanobis distance is the one built on the inverse of the data covariance matrix. In practice, there exist methods allowing to estimate the covariance matrix in order to choose the right Mahalanobis distance (Art et al., 1982; Tarsitano, 2003).

 ISSN: 2102-6238



Kullback-Leibler: **0.910**	L^2: 0.799	Squared bias: 0.017

FIGURE 6. *Clustering Gaussian curves ($k = 2$, $n = 40$). a) Squared bias. b) L^2. c) Kullback-Leibler.*



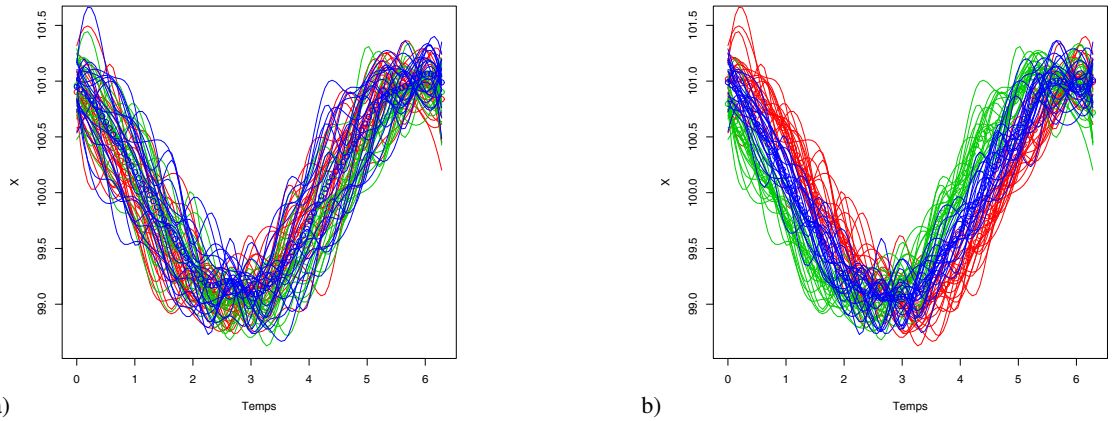
Kullback-Leibler: 0.017	L^2: 0.022	Squared bias: **1**

FIGURE 7. *Clustering Gaussian curves ($k = 2$, $n = 40$). a) Squared bias. b) L^2. c) Kullback-Leibler.*

Gaussian curves In the two next examples, we consider 40 Gaussian density curves. In the first case, there are two groups of 20 curves corresponding to Gaussian densities centered in 22.5 and 24.5 respectively, with a standard deviation chosen uniformly at random between 2 and 5. Results are presented in Figure 6. The two groups are most accurately recovered by the Kullback-Leibler divergence. In the second example, we have 20 Gaussian curves with standard deviation 4 and 20 curves with standard deviation 5, the mean being chosen uniformly between 21 and 26. This time, finding the two groups means clustering the curves with respect to the variance of the underlying normal distribution, whereas in the first example, the property characterizing a group was the mean. As shown in Figure 7, the squared bias provides the best result.

Noisy sinusoids In this last example, we consider observations building 3 groups of noisy sine waves, corresponding to 3 different phases. The variance of the Gaussian noise is set to 0.1. Results are visible in Figure 8. The squared L^2 distance and Itakura-Saito seem both to be accurate in order to cluster the sine waves with respect to their phase.

 ISSN: 2102-6238



L^2: **0.858**	Squared bias: 0.043	Itakura-Saito: **0.853**

FIGURE 8. *Clustering noisy sinusoids with phases* 0, $\pi/8$, $\pi/4$ ($k=3$, $n=45$). *a) Squared bias. b)* L^2. *Normalized mutual information (phases* 0, $\pi/24$, $\pi/12$*; 30 trials).*

3. Parameter selection for principal curves

In this section, let $\mathscr{X} = \mathbb{R}^d$. We assume that $\mathbb{E}\|X\|^2 < \infty$. Like a quantizer, defined by a codebook and a partition, a principal curve is in fact characterized by two objects, a parameterized curve $f : I \to \mathbb{R}^d$ and a map $\tau : \mathbb{R}^d \to I$. Playing the role of the Voronoi partition in quantization, there is a best choice for τ, which is the projection index t_f given by (3), so that principal curve estimation consists in minimizing the distortion over f

$$\Delta(f) = \mathbb{E}\min_{t\in I}\|X - f(t)\|^2, \quad \Delta_n(f) = \frac{1}{n}\sum_{i=1}^{n}\min_{t\in I}\|X_i - f(t)\|^2.$$

In the definition of Kégl et al. (2000), a principal curve of length L for X is a parameterized curve minimizing $\Delta(f)$ over curves of length at most L, whereas Sandilya and Kulkarni (2002) use a constraint on the turn of the curve. We define the length of a curve $f : I \to \mathbb{R}^d$ by

$$\mathscr{L}(f) = \sup\sum_{j=1}^{m}\|f(t_j) - f(t_{j-1})\|,$$

where the supremum is taken over all subdivisions $a = t_0 < t_1 < \cdots < t_m = b$, $m \geq 1$ (see, e.g., Kolmogorov and Fomin, 1975), whereas the turn of f is given by

$$\mathscr{K}(f) = \sup\sum_{j=1}^{m-1}\widehat{f(t_j)},$$

where $\widehat{f(t_j)}$ denotes the angle between the vectors $\overrightarrow{f(t_{j-1})f(t_j)}$ and $\overrightarrow{f(t_j)f(t_{j+1})}$, and, as above, the supremum is taken over all subdivisions $a = t_0 < t_1 < \cdots < t_m = b$, $m \geq 1$ (see

 ISSN: 2102-6238

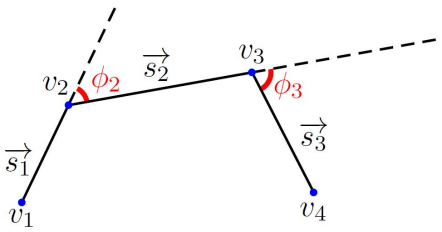


FIGURE 9. *Denoting by* $\overrightarrow{s_j}$ *the vector* $\overrightarrow{v_j v_{j+1}}$ *for all* $j = 1, \ldots, k$, *the angles involved in the definition of the turn are defined by* $\phi_{j+1} = (\overrightarrow{s_j}, \overrightarrow{s_{j+1}})$.

Alexandrov and Reshetnyak, 1989). In particular, the turn of a polygonal line with vertices $v_1, \ldots, v_{k+1}$ is just the sum of the angles at $v_2, \ldots, v_k$, as illustrated in Figure 9.

In this section, we study parameter selection methods, in order to construct a proper principal curve recovering accurately the shape of the data without interpolating. To this aim, we propose to use the approach of non-asymptotic model selection by penalization introduced by Birgé and Massart (1997) and Barron et al. (1999). First, we will consider a Gaussian framework, then the context of almost surely bounded random variables.

3.1. Length selection in a Gaussian framework

We investigate a Gaussian model selection method in order to choose the length of a principal curve. Proofs of the results presented in this subsection can be found in Fischer (2013). The context is similar to that of Caillerie and Michel (2011), who tackle model selection questions for graphs called simplicial complexes. In the subsection, the Euclidean space $\mathbb{R}^d$ is equipped with the inner product defined by

$$\langle \mathbf{u}, \mathbf{v} \rangle = \frac{1}{d} \sum_{j=1}^{d} u_j v_j. \tag{4}$$

The associated Euclidean norm is denoted by $\|\cdot\|$ and the associated distance by $d(\cdot;\cdot)$.

We assume that we observe random vectors $X_1, \ldots, X_n$ with values in $\mathbb{R}^d$ following the model

$$X_i = x_i^\star + \sigma \xi_i, \quad i = 1, \ldots, n, \tag{5}$$

where the $x_i^\star$ are unknown, the ξ_i are independent standard Gaussian vectors of $\mathbb{R}^d$ and $\sigma > 0$ is the noise level, supposed known. Denoting by $\mathbf{X} = {}^t({}^tX_1, \ldots, {}^tX_n)$ the (column) vector made of all coordinates of the random vectors X_i, $i = 1, \ldots, n$ and defining $\mathbf{x}^\star$ and ξ in the same way, the model (5) can be rewritten under the form

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

Let F and G be two fixed points of $\mathbb{R}^d$ and $\mathscr{L}$ a countable subset of $]0, +\infty[$. We introduce a countable collection $\{\mathscr{F}_\ell\}_{\ell \in \mathscr{L}}$, where each set $\mathscr{F}_\ell$ is a class of parameterized

 ISSN: 2102-6238

curves $f : I \to \mathbb{R}^d$ with length ℓ and endpoints F and G. Set $\lambda := \sqrt{\ell^2 - d(F;G)^2}$. We consider the criterion Δ'_n given by

$$\Delta'_n(f) = \frac{1}{n}\sum_{i=1}^{n} \min_{t\in I} \|X_i - f(t)\|^2 = \frac{1}{n}\sum_{i=1}^{n} \min_{x_i\in\Gamma_f} \|X_i - x_i\|^2,$$

where Γ_f denotes the range of the curve f. Due to the definition (4) of the norm $\|\cdot\|$, this is the empirical criterion $\Delta_n(f)$ normalized by the dimension d. Suppose that, for all $\ell \in \mathscr{L}$, $\hat{\mathbf{x}}_\ell := (\hat{x}_{1\ell}, \ldots, \hat{x}_{n\ell})$ minimizes

$$\frac{1}{n}\sum_{i=1}^{n} \|X_i - x_i\|^2$$

among all $\mathbf{x} \in \mathscr{C}_\ell := \bigcup_{f\in\mathscr{F}_\ell} (\Gamma_f)^n$. In order to determine the length ℓ, we will minimize a criterion of the type

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

where $\mathrm{pen} : \mathscr{L} \to \mathbb{R}^+$ is a penalty function, whose role is to prevent the choice of a too large ℓ. Observe that the classical asymptotic model selection criteria AIC (Akaike, 1973), BIC (Schwarz, 1978) or Mallows'C_p (Mallows, 1973), which involve the "number of parameters" to be estimated, are not suitable to design an appropriate penalty in this framework. However the non-asymptotic model selection theory developed by Birgé and Massart (2001) and Barron et al. (1999) allows us to derive the next theorem, based on results by Massart (2007) on Gaussian model selection for non linear models.

Théorème 4. *Assume that there are nonnegative weights* $\{w_\ell\}_{\ell\in\mathscr{L}}$ *such that* $\sum_{\ell\in\mathscr{L}} e^{-w_\ell} = \Sigma < \infty$, *and that, for every* $\ell \in \mathscr{L}$,

$$\sigma \le \frac{\lambda}{4\kappa}\left[\sqrt{\ln 2 + \frac{1}{d}\ln\left(\frac{\ell}{\lambda}\right)} + \sqrt{\pi}\right]^{-1}. \tag{6}$$

Then, there exist constants c_1 *and* c_2 *such that, for all* $\eta > 1$, *if*

$$\mathrm{pen}(\ell) \ge \eta\sigma^2\left[c_1\left(\ln\left(\frac{\ell^{1/d}\lambda^{1-1/d}}{\sigma}\right) + c_2\right) + \frac{4w_\ell}{nd}\right], \tag{7}$$

then, almost surely, there exists a minimizer $\hat{\ell}$ *of the penalized criterion*

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

Moreover, if $\tilde{\mathbf{x}} = \hat{\mathbf{x}}_{\hat{\ell}}$, *we have*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\|\tilde{x}_i - x_i^\star\|^2 \le c(\eta)\left[\inf_{\ell\in\mathscr{L}}\{D^2(\mathbf{x}^\star, \mathscr{C}_\ell) + \mathrm{pen}(\ell)\} + \frac{\sigma^2}{nd}(\Sigma + 1)\right],$$

where $D^2(\mathbf{x}^\star, \mathscr{C}_\ell) = \inf_{\mathbf{y}\in\mathscr{C}_\ell} \frac{1}{n}\sum_{i=1}^{n} \|y_i - x_i^\star\|^2$.

 ISSN: 2102-6238

Let us comment on the theorem.

The first remark is about the fact that Theorem 4 involves unknown constants. The proof indicates that $c_1 = 16\kappa^2$ and $c_2 = \pi - \ln(2\kappa\sqrt{\pi})$ could be chosen. However, these values are (likely too large) upper bounds. Furthermore, the variance noise σ has been supposed known and is involved in the penalty. Nevertheless, the noise level is generally unknown in practice. Note that it is possible to estimate σ separately and then proceed by plug-in. However, there is another solution to assess c_1, c_2 and σ, relying on the slope heuristics. This penalty calibration method introduced by Birgé and Massart (2007) precisely allows to tune a penalty known up to a multiplicative constant.

According to the formula binding ℓ and λ, the quantity $\ln(\ell^{1/d}\lambda^{1-1/d})$ in the penalty characterizes each model of curves with length ℓ. The other elements varying over the collection of models are the weights $\{w_\ell\}_{\ell\in\mathscr{L}}$. They should be large enough to ensure the finiteness of Σ, but not too large at the risk of overpenalizing. For linear models $\mathscr{C}_\ell$ with dimension D_ℓ, a possible choice for w_ℓ is $w_\ell = w(D_\ell)$ where $w(D) = cD + \ln|\{k \in \mathscr{L}, D_k = D\}|$ and $c > 0$ (see Massart, 2007). If there is no redundancy in the models dimension, this strategy amounts to choosing w_ℓ proportional to D_ℓ. By analogy, w_ℓ may here be chosen proportional to $\ln(\ell^{1/d}\lambda^{1-1/d})$. More formally, we set $w_\ell = c\ln\ell^{1/d}\lambda^{1-1/d}$, where the constant $c > 0$ is such that $\sum_{\ell\in\mathscr{L}} \frac{1}{\ell^{c/d}\lambda^{c(1-1/d)}} = \Sigma < +\infty$. Considering only the main term in the lower bound (7), this choice finally yields a penalty proportional to $\ln(\ell^{1/d}\lambda^{1-1/d})$, which may be calibrated in practice thanks to the slope heuristics.

Besides, condition (6) says that the noise level σ should not be too large with respect to λ. If $\lambda = \sqrt{\ell^2 - d(F;G)^2}$ is of the same order as σ, it is not possible to obtain a suitable principal curve with length ℓ.

Regarding the fact that the endpoints F and G of the principal curve are fixed, observe that several methods can be employed in practice to choose them from the data. A possible solution is to define F and G using the points that are farthest from each other in the minimum spanning tree of the data (or of some subset of the data), which can be constructed thanks to the algorithm of Kruskal (1956) or Prim (1957).

Finally, let us point out that the penalty shape obtained does not tend to 0 as n tends to infinity. This point is intrinsically related to the geometry of the problem, which is not made easier by increasing the size of the sample, since nothing has been specified about the repartition of the $x_i^\star$'s.

3.2. *Selecting parameters in a bounded framework*

Let $\|\cdot\|$ denote the standard Euclidean norm again. Assume that

$$\mathbb{P}(X \in \mathscr{C}) = 1, \tag{8}$$

where $\mathscr{C}$ is a convex compact subset of $\mathbb{R}^d$, with diameter δ. By Lemma 1 in Kégl (1999), requirement (8) implies that, for any given positive length L, there exists a parameterized curve $f^\star$ with length at most L and support in $\mathscr{C}$ minimizing $\Delta(f)$. It follows from Proposition 1 in Sandilya and Kulkarni (2002) that this remains true when replacing the length by the turn. We still denote the minimizer by $f^\star$ in this case, since there will be no ambiguity. In the sequel, we restrict ourselves to parameterized curves whose support is included in $\mathscr{C}$.

 ISSN: 2102-6238

Contrary to the Gaussian framework discussed in the previous subsection, here results will be about the curve f itself and not about the range of the curve. Moreover, this time, penalty shapes tending to 0 as the sample size grows to infinity will be obtained. From the technical standpoint, note that these two different contexts lead to the use of quite different tools. The results stated in the present subsection are proved in Biau and Fischer (2012).

Principal curves with bounded length Let $\mathscr{L}$ be a countable subset of $]0,L]$ and $\mathbb{Q}$ a grid over $\mathscr{C}$, that is $\mathbb{Q}=\mathscr{C}\cap\Gamma$, where Γ is a lattice of $\mathbb{R}^d$. For every $k\geq 1$ and $\ell\in\mathscr{L}$, the model $\mathscr{F}_{k,\ell}$ is defined as the collection of all polygonal lines with k segments, with length at most ℓ, and with vertices belonging to $\mathbb{Q}$. We note that each model $\mathscr{F}_{k,\ell}$ as well as the family of models $\{\mathscr{F}_{k,\ell}\}_{k\geq 1,\ell\in\mathscr{L}}$ are countable. For $k\geq 1$ and $\ell\in\mathscr{L}$, let

$$\hat{f}_{k,\ell}\in\operatorname*{arg\,min}_{f\in\mathscr{F}_{k,\ell}}\Delta_n(f)$$

be a curve achieving the minimum of the empirical criterion $\Delta_n(f)$ over the class $\mathscr{F}_{k,\ell}$. Our goal is to select the best principal curve $\tilde{f}$ among the collection $\{\hat{f}_{k,\ell}\}_{k\geq 1,\ell\in\mathscr{L}}$. The model selection approach will allow us to control the loss

$$\mathscr{D}(f^\star,\tilde{f})=\Delta(\tilde{f})-\Delta(f^\star)$$

between the target $f^\star$ and the selected curve $\tilde{f}$. Let $\mathrm{pen}:\mathbb{N}^\star\times\mathscr{L}\to\mathbb{R}^+$ be some penalty function and denote by $(\hat{k},\hat{\ell})$ a pair of minimizers of the criterion

$$\mathrm{crit}(k,\ell)=\Delta_n(\hat{f}_{k,\ell})+\mathrm{pen}(k,\ell).$$

In order to obtain the desired principal curve $\tilde{f}=\hat{f}_{\hat{k},\hat{\ell}}$, an adequate penalty $\mathrm{pen}(k,\ell)$ has to be designed, which can be achieved by establishing an upper bound on the quantity

$$\mathbb{E}\sup_{f\in\mathscr{F}_{k,\ell}}\left(\Delta(f)-\Delta_n(f)\right).$$

Then, the following result can be proved.

Théorème 5. *Consider nonnegative weights* $\{x_{k,\ell}\}_{k\geq 1,\ell\in\mathscr{L}}$ *such that* $\sum_{k\geq 1,\ell\in\mathscr{L}}e^{-x_{k,\ell}}=\Sigma<\infty$, *and a penalty function* $\mathrm{pen}:\mathbb{N}^\star\times\mathscr{L}\to\mathbb{R}^+$. *Let* $\tilde{f}=\hat{f}_{\hat{k},\hat{\ell}}$. *There exist nonnegative constants* $c_0,\ldots,c_2$, *depending on the dimension* d *and the diameter* δ *of the convex set* $\mathscr{C}$, *such that, if for all* $(k,\ell)\in\mathbb{N}^\star\times\mathscr{L}$,

$$\mathrm{pen}(k,\ell)\geq\frac{1}{\sqrt{n}}\left[c_1\sqrt{k}+c_2\max\left(\frac{\ell}{\sqrt{k}},\sqrt{k\ln\frac{\ell}{k}}\right)\ell+c_0\right]+\delta^2\sqrt{\frac{x_{k,\ell}}{2n}},\tag{9}$$

then

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

where $\mathscr{D}(f^\star,\mathscr{F}_{k,\ell})=\inf_{f\in\mathscr{F}_{k,\ell}}\mathscr{D}(f^\star,f)$.

 ISSN: 2102-6238

Let us give some comments.

Firstly, we see that the penalty shape involves a term proportional to $\sqrt{k/n}$, whereas the quantity $\ell/\sqrt{kn}$ or $\sqrt{k/n\ln \ell/k}$ suggests that k and ℓ should be cleverly chosen relatively to each other. This penalty form, which vanishes at the rate $1/\sqrt{n}$, seems relevant insofar as the number k of segments and the length ℓ of the curves measure the complexity of the models.

As in the Gaussian framework, the proof provides possible values for the constants $c_0,\dots,c_2$, but these values are not very helpful since they are upper bounds which are probably far from being tight. Besides, the proof also reveals that $c_1 = c_1'\delta^2$, $c_2 = c_2'\delta$ and $c_0 = c_0'\delta^2$, where c_0', c_1' and c_2' are constants without dimension, so that the penalty is in fact homogeneous to a squared length, just like the criterion $\Delta_n(f)$.

As for the weights, if the cardinality of the collection of models is not larger than n^2 (this will be the case in all our practical examples), we may set $x_{k,\ell} = 2\ln n$ for every (k,ℓ). This choice does not affect the penalty shape, though modifying the rate, and leads to $\Sigma = 1$ in the risk bound.

Principal curves with bounded turn The same kind of result can be obtained in the context of curves with bounded turn, using the fact that a curve with bounded turn also has bounded length. Indeed, the following lemma holds (see, e.g., Alexandrov and Reshetnyak (1989, Chapter 5)):

Lemma 3.1. *Let f be a curve with turn κ and let δ be the diameter of $\mathscr{C}$. Then $\mathscr{L}(f) \le \delta\zeta(\kappa)$, where the function ζ is defined by*

$$\zeta(x) = \begin{cases} \dfrac{1}{\cos(x/2)} & \text{if} \quad 0 \le x \le \dfrac{\pi}{2} \\ 2\sin(x/2) & \text{if} \quad \dfrac{\pi}{2} \le x \le \dfrac{2\pi}{3} \\ \dfrac{x}{2} - \dfrac{\pi}{3} + \sqrt{3} & \text{if} \quad x \ge \dfrac{2\pi}{3}. \end{cases}$$

Let $\mathscr{K}$ be a countable subset of $[0,K]$ and $\{\mathscr{F}_{k,\kappa}\}_{k\ge 1,\kappa\in\mathscr{K}}$ a countable collection of models, where each $\mathscr{F}_{k,\kappa}$ consists of polygonal lines with k segments, turn at most κ, and vertices belonging to some grid $\mathbb{Q}$ over $\mathscr{C}$. For $k \ge 1$ and $\kappa \in \mathscr{K}$, we define

$$\hat{f}_{k,\kappa} \in \operatorname*{arg\,min}_{f\in\mathscr{F}_{k,\kappa}} \Delta_n(f)$$

and

$$\operatorname{crit}(k,\kappa) = \Delta_n(\hat{f}_{k,\kappa}) + \operatorname{pen}(k,\kappa).$$

As before, let $\tilde{f} = \hat{f}_{\hat{k},\hat{\kappa}}$, where $(\hat{k},\hat{\kappa})$ is a minimizer of $\operatorname{crit}(k,\kappa)$. The version of Theorem 5 for principal curves with bounded turn states as follows.

Théorème 6. *Consider nonnegative weights $\{x_{k,\kappa}\}_{k\ge 1,\kappa\in\mathscr{K}}$ such that $\sum_{k\ge 1,\kappa\in\mathscr{K}} e^{-x_{k,\kappa}} = \Sigma < \infty$, and a penalty function $\operatorname{pen} : \mathbb{N}^\star \times \mathscr{K} \to \mathbb{R}^+$. Let $\tilde{f} = \hat{f}_{\hat{k},\hat{\kappa}}$. There exist nonnegative constants $c_0,\dots,c_2$, depending only on the dimension d, such that, if for all $(k,\kappa) \in \mathbb{N}^\star \times \mathscr{K}$,*

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

 ISSN: 2102-6238

then

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

where $\mathscr{D}(f^\star,\mathscr{F}_{k,\kappa})=\inf_{f\in\mathscr{F}_{k,\kappa}}\mathscr{D}(f^\star,f)$.

The resulting penalty shape is similar to (9). The length ℓ is replaced by $\zeta(\kappa)$, increasing function of the turn κ.

3.3. Some illustrations

This section presents some real data illustrations, carried out with the software MATLAB. Our purpose here is either length selection (Theorem 4), or simultaneous choice of length and number of segments (Theorem 5). In order to assess the constants appearing in the theorems, we use the slope heuristics proposed by Birgé and Massart (2007) (see also Lebarbier, 2005, Arlot and Massart, 2009, Lerasle, 2012, Saumard, 2013, and the overview by Baudry et al., 2012, who have implemented the method in the package CAPUSHE).

In short, the slope heuristics consists in observing that the empirical contrast is proportional to the penalty shape for complex models and in using the slope of this line to assess the constant. Here, when dealing with two parameters, we use a bivariate version of this heuristics, where constants are chosen via a bivariate ordinary least square regression. We let w_ℓ be proportional to $\ln \ell^{1/d}\lambda^{1-1/d}$ and $x_{k,\ell}=2\ln n$.

Three examples will be presented: we show first an application of Theorem 4 to GPS tracks data and then applications of Theorem 5 to character recognition and seismic data.

GPS track We present an application of principal curve to mapping. Indeed, Brunsdon (2007) has shown that principal curves may be useful in that area, in order to estimate paths from GPS tracks. More specifically, principal curves are a means to compute an average path from GPS data registered by several people moving on a given street.

The place chosen in this example is the "Labyrinth" of the Jardin des Plantes in Paris. Here, the slope heuristics step was applied via the package CAPUSHE. The result is visible in Figure 10. The figure gives first an air photography of the place and the corresponding GPS track data points. Then, the resulting principal curve is shown both on the data cloud and as an overlay on the photography, which allows to assess the performance of the method. We see that the Labyrinth is quite well recovered, with a very smooth curve.

NIST database digits The data set used here is part of NIST Special Database 19 (`http://www.nist.gov/srd/nistsd19.cfm`), which contains handwritten characters from 3600 writers. The data consists in binary images scanned at 11.8 dots per millimeter (300 dpi), which uniformly fill the area corresponding to the thickness of the pen stroke. Skeletonization, which consists in reducing foreground regions in such an image without affecting the general shape of the handwritten character, often constitutes a preliminary step to perform character recognition (see, e.g., Deutsch, 1968 and Alcorn and Hoggar, 1969). The principal curve estimation method was applied to two NIST database digits, 2 and 5. We observe in Figure

 ISSN: 2102-6238

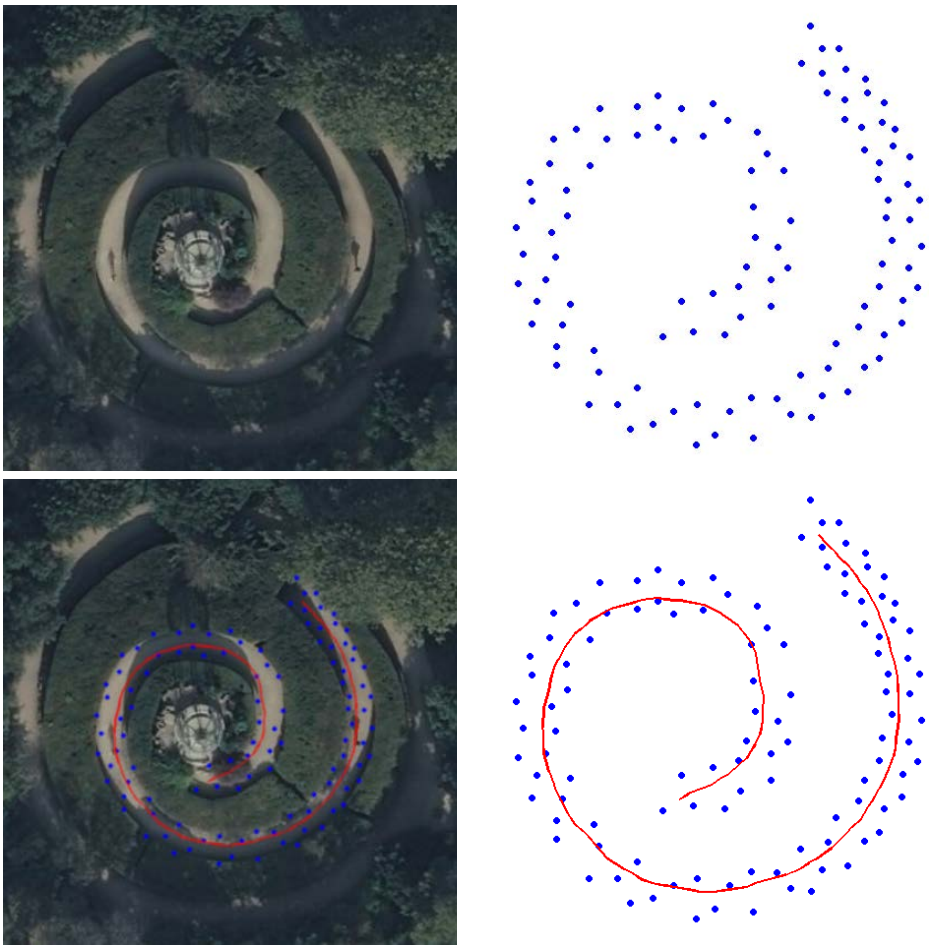


FIGURE 10. *Principal curve fitted for the Labyrinth data.*

11 that both results are satisfactory. Note that R^2 coefficients equal to 0.99 were obtained in the least square regression used to apply the slope heuristics.

Seismic data In this last experiment, principal curves are used in order to recover lithospheric plates borders using seismic data. As shown in Figure 12, seismic impacts correspond to the borders of lithosperic plates. Available on the USGS (United States Geological Survey) website (`http://earthquake.usgs.gov/research/data/centennial.php`), the data set is part of the "Centennial Catalog", listing the major earthquakes registered since 1900 (Engdahl and Villaseñor, 2002).

We focus on two regions indicated on the world map in Figure 13. The principal curve results, shown in Figure 14, confirm the good performance of the method.

 ISSN: 2102-6238

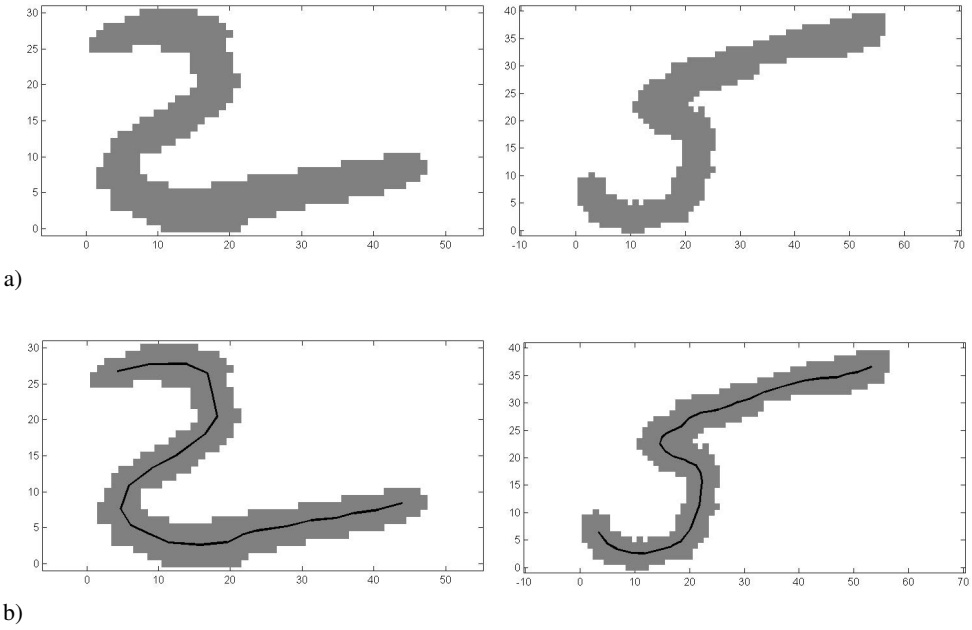


FIGURE 11. *a) Two NIST database handwritten digits. b) Principal curves selected:* $\hat{k} = 23$, $\hat{\ell} = 80$; $\hat{k} = 38$, $\hat{\ell} = 82$.

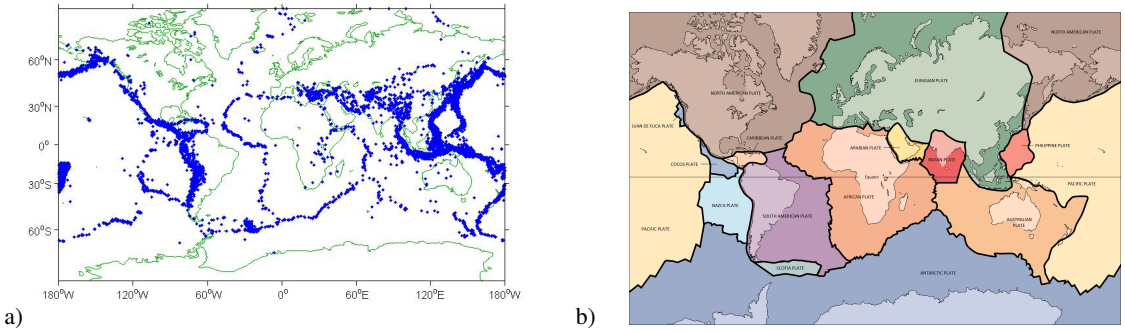


FIGURE 12. *a) Earthquake impacts and b) lithospheric plate borders (UGS map).*

 ISSN: 2102-6238

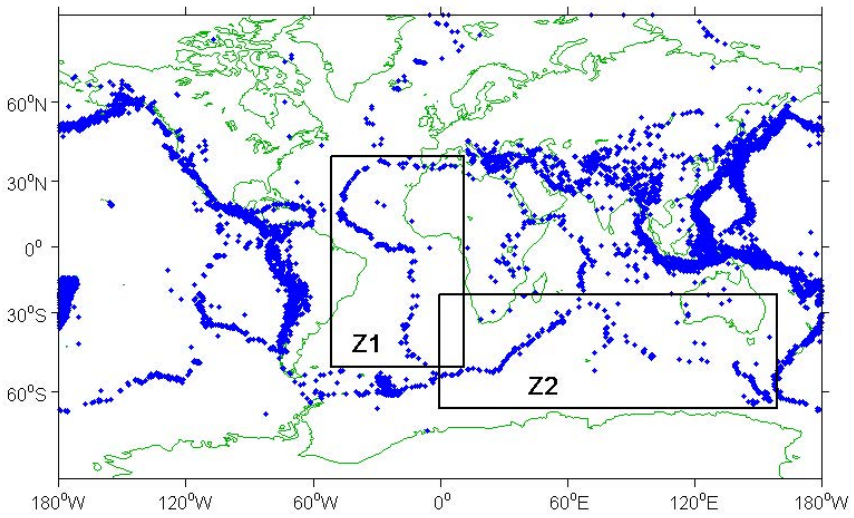
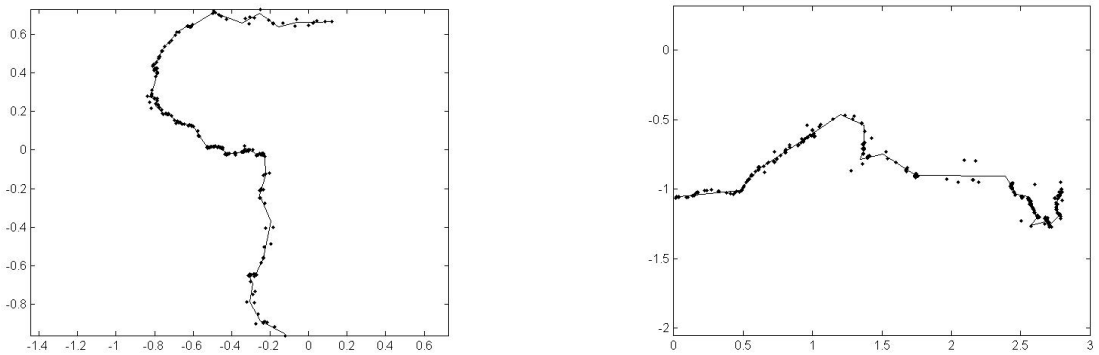


FIGURE 13. *Localization of the two considered seismic zones* ***Z1*** *(about 60°S 50°W to 40°N 0°) and* ***Z2*** *(about 65°S 0° to 25°S 160°E).*



a) b)

FIGURE 14. *Selected principal curves for the seismic zones. a)* ***Z1*** *($n = 252$) $\hat{k} = 55$, $\hat{\ell} = 31$. b)* ***Z2*** *($n = 322$) $\hat{k} = 22$, $\hat{\ell} = 38$.*

 ISSN: 2102-6238

Acknowledgment

I would like to thank the Editor and the two referees for all their interesting and insightful comments.

Je remercie bien sincèrement la Société Française de Statistique, et en particulier Jean-Jacques Droesbeke et l'ensemble du jury du Prix Marie-Jeanne Laurent-Duhamel 2014. Félicitations à mon co-récipiendaire Christophe Ley et merci à lui pour sa bonne humeur. Je ne saurais oublier Gérard et tous ceux avec qui j'ai eu le plaisir de discuter de ces jolies questions : de tout coeur, merci à chacun !

Références

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281.

Alber, Y. and Butnariu, D. (1997). Convergence of Bregman projection methods for solving consistent convex feasibility problems in reflexive Banach spaces. *Journal of Optimization Theory and Applications*, 92 :33–61.

Alcorn, T. M. and Hoggar, C. W. (1969). Preprocessing of data for character recognition. *Marconi Review*, pages 61–81.

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

Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10 :245–279.

Art, D., Gnanadesikan, R., and Kettenring, J. R. (1982). Data-based metrics for cluster analysis. *Utilitas Mathematica*, 21A :75–99.

Banerjee, A., Guo, X., and Wang, H. (2005a). On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005b). Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6 :1705–1749.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113 :301–413.

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics : overview and implementation. *Statistics and Computing*, 22 :455–470.

Bauschke, H. H., Borwein, J. M., and Combettes, P. L. (2001). Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces. *Communications in Contemporary Mathematics*, 3 :615–647.

Biau, G., Devroye, L., and Lugosi, G. (2008). On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54 :781–790.

Biau, G. and Fischer, A. (2012). Parameter selection for principal curves. *IEEE Transactions on Information Theory*, 58 :1924–1939.

Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In Pollard, D., Torgersen, E., and Yang, G., editors, *Festschrift for Lucien Le Cam : Research Papers in Probability and Statistics*, pages 55–87. Springer, New York.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3 :203–268.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138 :33–73.

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*.

 ISSN: 2102-6238

Caillerie, C. and Michel, B. (2011). Model selection for simplicial approximation. *Foundations of Computational Mathematics*, 11 :707–731.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press, New York.

Csiszár, I. (1995). Generalized projections for non-negative functions. *Acta Mathematica Hungarica*, 68 :161–185.

Deutsch, E. S. (1968). Preprocessing for character recognition. In *Proceedings of the IEE–NPL Conference on Pattern Recognition*, pages 179–190.

Engdahl, E. R. and Villaseñor, A. (2002). Global seismicity : 1900–1999. In Lee, W., Kanamori, H., Jennings, P., and Kisslinger, C., editors, *International Handbook of Earthquake and Engineering Seismology*, pages 665–690. Academic Press, London.

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

Fischer, A. (2013). Selecting the length of a principal curve within a Gaussian model. *Electronic Journal of Statistics*, 7 :342–363.

Fischer, A. (2014). Deux méthodes d'apprentissage non supervisé : synthèse sur la méthode des centres mobiles et présentation des courbes principales. *Journal de la Société Française de Statistique*, 155 :2–35.

Frigyik, B. A., Srivastava, S., and Gupta, M. R. (2008a). An introduction to functional derivatives. Technical Report UWEETR-2008-0001, Department of Electrical Engineering, University of Washington, Seattle.

Frigyik, B. A., Srivastava, S., and Gupta, M. R. (2008b). Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 54 :5130–5139.

Gersho, A. and Gray, R. M. (1992). *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Norwell.

Graf, S. and Luschgy, H. (2000). *Foundations of Quantization for Probability Distributions*. Lecture Notes in Mathematics. Springer-Verlag, Berlin, Heidelberg.

Gray, R. M., Buzo, A., Gray, A. H., and Matsuyama, Y. (1980). Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28 :367–376.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84 :502–516.

Jones, L. and Byrne, C. (1990). General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis. *IEEE Transactions on Information Theory*, 36.

Jordan, M. I., Nguyen, X., and Wainwright, M. J. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56 :5847–5861.

Kolmogorov, A. N. and Fomin, S. V. (1975). *Introductory Real Analysis*. Dover Publications, Mineola.

Kruskal, J. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. In *Proceedings of the American Mathematical Society*, volume 7, pages 48–50.

Kégl, B. (1999). *Principal Curves : Learning, Design, and Applications*. PhD thesis, Concordia University, Montréal, Québec, Canada.

Kégl, B., Krzyżak, A., Linder, T., and Zeger, K. (2000). Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 :281–297.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal Processing*, 85 :717–736.

Lerasle, M. (2012). Optimal model selection in density estimation. *Annales de l'Institut Henri Poincaré*, 48 :884–908.

Linder, T. (2002). Learning-theoretic methods in vector quantization. In Györfi, L., editor, *Principles of Nonparametric Learning*. Springer-Verlag, Wien.

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

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15 :661–675.

Massart, P. (2007). *Concentration Inequalities and Model Selection*. Ecole d'Eté de Probabilités de Saint-Flour XXXIII – 2003, Lecture Notes in Mathematics. Springer, Berlin, Heidelberg.

Nielsen, F., Boissonnat, J., and Nock, R. (2007). Bregman Voronoi diagrams : properties, algorithms and applications. Technical Report 6154, INRIA.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell System Technology Journal*, 36 :1389–1401.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press, Princeton, New Jersey.

Sandilya, S. and Kulkarni, S. R. (2002). Principal curves with bounded turn. *IEEE Transactions on Information*

 ISSN: 2102-6238

Theory, 48 :2789–2793.
Saumard, A. (2013). The slope heuristics in heteroscedastic regression. *Electronic Journal of Statistics*, 7 :1184–1223.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6 :33–73.
Strehl, A. and Ghosh, J. (2002). Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3 :583–617.
Tarsitano, A. (2003). Mahalanobis metrics for *k*-means algorithms. In *Atti del Convegno intermedio SIS, Napoli*.

 ISSN: 2102-6238