Parameter Selection for Principal Curves

Gérard Biau
Université Pierre et Marie Curie[1] & Ecole Normale Supérieure, France[2]
gerard.biau@upmc.fr

Aurélie Fischer[3]
Université Pierre et Marie Curie, France
aurelie.fischer@upmc.fr

       Abstract – Principal curves are nonlinear generalizations of the notion of first principal component. Roughly, a principal curve is a parameterized curve in $\mathbb{R}^d$ which passes through the "middle" of a data cloud drawn from some unknown probability distribution. Depending on the definition, a principal curve relies on some unknown parameters (number of segments, length, turn...) which have to be properly chosen to recover the shape of the data without interpolating. In the present paper, we consider the principal curve problem from an empirical risk minimization perspective and address the parameter selection issue using the point of view of model selection via penalization. We offer oracle inequalities and implement the proposed approaches to recover the hidden structures in both simulated and real-life data.

Index terms – Principal curves, parameter selection, model selection, oracle inequality, penalty calibration, slope heuristics.

2010 Mathematics Subject Classification: 62G08, 62G05.

1 Introduction

1.1 Principal curves

Statisticians use various methods in order to sum up information and represent the data by simpler quantities. Among these methods, Principal Component Analysis

---

[3]Corresponding author.

(PCA) aims at determining the maximal variance axes of a data cloud, as a means to represent the observations in a compact manner revealing as well as possible their variability (see, e.g., Mardia, Kent and Bibby [34]). This technique, initiated at the beginning of the last century by Pearson [37] and Spearman [41], and further developed by Hotelling [26], is certainly one of the most famous and most widely used procedure of multivariate analysis. Whether in the context of dimension reduction or feature extraction, PCA often provides a first important insight in the data structure.

However, in a number of situations, it may be of interest to summarize information in a nonlinear manner instead of representing the data by straight lines. This approach leads to the notion of principal curve, which can be thought of as a nonlinear generalization of the first principal component. Roughly, the purpose is to search for a curve passing through the middle of the observations, as illustrated in Figure 1. Principal curves have a broad range of applications in many different areas, such as physics (Hastie and Stuetzle [25], Friedsam and Oren [22]), character and speech recognition (Kégl and Krzyżak [28], Reinhard and Niranjan [38]), mapping and geology (Brunsdon [9], Stanford and Raftery [42], Banfield and Raftery [4], Einbeck, Tutz and Evers [19, 20]), natural sciences (De'ath [13], Corkeron, Anthony and Martin [12], Einbeck, Tutz and Evers [19]) and medicine (Wong and Chung [45], Caffo, Crainiceanu, Deng and Hendrix [10]).

The definition of a principal curve typically depends of the principal component property one wants to generalize. Most of the time, this definition is first stated for an $\mathbb{R}^d$-valued random variable $\mathbf{X} = (X_1, \dots, X_d)$ with known distribution, and then adapted to the practical situation where one observes independent draws $\mathbf{X}_1, \dots, \mathbf{X}_n$ distributed as $\mathbf{X}$.
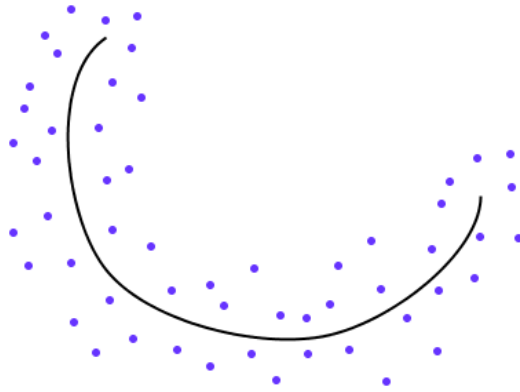

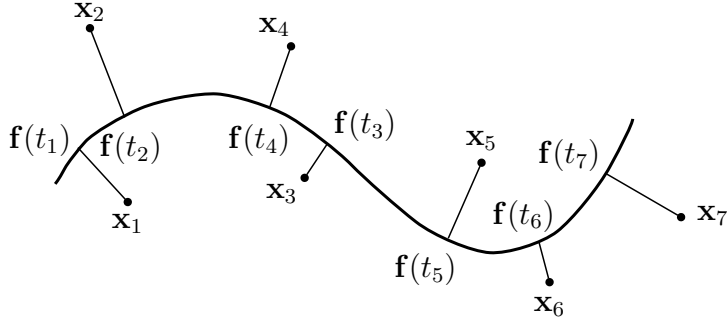Figure 1: An example of principal curve.

Figure 2: The projection index $t_{\mathbf{f}}$. For all i, t_i stands for $t_{\mathbf{f}}(\mathbf{x}_i)$.

The original definition of a principal curve goes back to Hastie and Stuetzle [25] and relies on the self-consistency property of principal components. In words, a smooth (infinitely differentiable) parameterized curve $\mathbf{f}(t) = (f_1(t), \ldots, f_d(t))$ is a principal curve for $\mathbf{X}$ if $\mathbf{f}$ does not intersect itself, if it has finite length inside any bounded subset of $\mathbb{R}^d$, and if it is self-consistent. This last requirement means that

$$\mathbf{f}(t) = \mathbb{E}[\mathbf{X}|t_{\mathbf{f}}(\mathbf{X}) = t], \tag{1}$$

where the so-called projection index $t_{\mathbf{f}}(\mathbf{x})$ is the largest real number t minimizing the squared Euclidean distance between $\mathbf{x}$ and $\mathbf{f}(t)$, as depicted in Figure 2. More formally,

$$t_{\mathbf{f}}(\mathbf{x}) = \sup \left\{ t \; : \; \|\mathbf{x} - \mathbf{f}(t)\| = \inf_{t'} \|\mathbf{x} - \mathbf{f}(t')\| \right\}.$$

The self-consistency property may be interpreted by saying that each point of the curve $\mathbf{f}$ is the mean of the observations projecting on $\mathbf{f}$ around this point. Hastie and Stuetzle discuss in [25] an iterative algorithm, alternating between a projection and a conditional expectation step, which yields an approximate principal curve. As this approach exhibits different types of bias, Banfield and Raftery [4] and Chang and Ghosh [11] propose a modification of the algorithm, whereas Tibshirani, tackling the model bias problem, adopts in [43] a semiparametric strategy and defines principal curves in terms of a mixture model. For more references on principal curves and related points of view, we refer the reader to Verbeek, Vlassis and Kröse [44] (k-segments algorithm), Delicado [14] (principal curves of oriented points), Einbeck, Tutz and Evers [20] (local principal curves) and Genovese, Perone-Pacifico, Verdinelli and Wasserman [23], who recently discussed a closely related approach, called nonparametric filament estimation.

3

In the present paper, we will adopt the principal curve definition of Kégl, Krzyżak, Linder and Zeger [29], which is slightly different from the original one. The main advantage of this definition, which is recalled in the next paragraph, is that it avoids the implicit conditional expectation requirement (1) and, consequently, turns out to be more easily amenable to mathematical analysis.

1.2 Constrained principal curves

In the definition of Kégl, Krzyżak, Linder and Zeger [29] (**KKLZ** hereafter), a principal curve of length (at most) L for $\mathbf{X}$ is a parameterized curve minimizing the least-square criterion

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

over a collection $\mathcal{F}_L$ of curves of length not larger than some prespecified positive L. We note that, in this context, a principal curve always exists provided $\mathbb{E}\|\mathbf{X}\|^2 < \infty$, but that it may not necessarily be unique. In practice, as the distribution of $\mathbf{X}$ is unknown, $\Delta(\mathbf{f})$ is replaced by its empirical counterpart

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

based on a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of independent random variables distributed as $\mathbf{X}$. Considering the minimum $\hat{\mathbf{f}}_{k,n}$ of $\Delta_n(\mathbf{f})$ over the subclass $\mathcal{F}_{k,L} \subset \mathcal{F}_L$ of all polygonal lines $\mathbf{f}_{k,n}$ with k segments and length not larger than L, Kégl, Krzyżak, Linder and Zeger [29] prove that, whenever $\mathbf{X}$ is almost surely bounded, and for the choice $k \propto n^{1/3}$,

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

As the task of finding a polygonal line with k segments and length at most L minimizing $\Delta_n(\mathbf{f})$ is computationally difficult, **KKLZ** propose an approximate iterative algorithm that they call the Polygonal Line Algorithm. This algorithm is initialized using the smallest segment included in the first principal component containing all projected data points. Then, at each step, a vertex—and thus, a segment—is added to the current polygonal line, and the vertices are updated in a cyclic manner during an inner loop alternating between a projection and an optimization step. Performing the projection step is similar to constructing a Voronoi partition, with respect to both the vertices and segments. To optimize a vertex, a local version of $\Delta_n(\mathbf{f})$ is used, involving only the data projecting to this vertex and to the adjacent segments. The criterion is penalized to avoid sharp angles, which in turn amounts to penalizing the length of the curve.

Working out the angle penalty in the Polygonal Line Algorithm, Sandilya and Kulkarni (**SK** hereafter) propose in [39] a closely related definition, by imposing a constraint on the turn (Alexandrov and Reshetnyak [2]) of the curve $\mathbf{f}$. This approach consists in replacing the class $\mathcal{F}_L$ by $\mathcal{F}_K$, where K stands for the maximal turn. Thus, denoting by $\mathcal{F}_{k,K} \subset \mathcal{F}_K$ the subclass of all polygonal lines $\mathbf{f}_{k,n}$ with k segments and turn not larger than K, **SK** prove that, whenever $\mathbf{X}$ is almost surely bounded, and for the choice $k \propto n^{1/3}$,

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

Whether in the **KKLZ** definition or in the **SK** one, selecting the various smoothness parameters (the number k of segments, the curve length ℓ, the turn κ) is an essential issue, as illustrated in Figure 3. A good choice of these parameters is critical, since a principal curve obtained with a poor class will be too rough, whereas a class containing too many curves may lead to severe interpolation problems. In practice, the Polygonal Line Algorithm stops when k is larger than a certain threshold, chosen heuristically and tuned after carrying out several experiments. The stopping condition involves the number n of observations and the actual value of the criterion Δ_n. However, to our knowledge, this empirical procedure is not supported by any theoretical argument and leads to variable results, depending on the data set. Besides, note that assessing the complexity of Hastie and Stuetzle [25] principal curve estimates by cross-validation has often been observed to fail. As put forward by Duchamp and Stuetzle [16], these principal curves are saddle points of the distance between a random vector and a curve, and therefore cross-validation is not a well-suited technique in the principal curve framework.
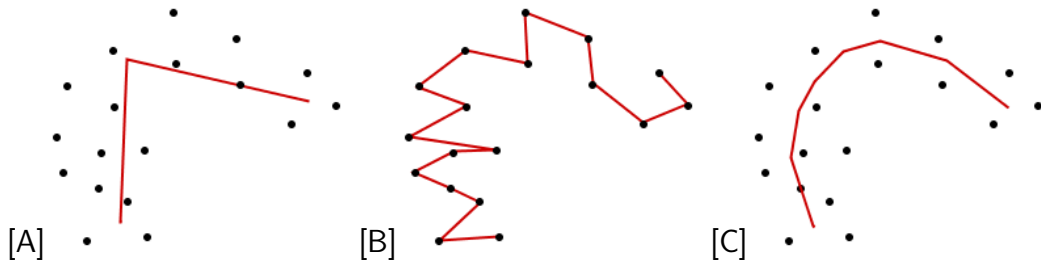

Figure 3: Principal curves fitted with [A] a too small number k of segments, [B] a too large k and [C] an appropriate one.

As far as we know, the issue of an automatic (i.e., data-dependent) choice of the parameters k, ℓ and κ has not been addressed in the literature. Thus, to fill the gap, we propose in the present contribution to focus on this question both from a theoretical and practical point of view. Our approach will strongly rely on the

model selection theory by penalization introduced by Birgé and Massart [7] and Barron, Birgé and Massart [5], as well as on a recent penalty calibration approach proposed by Birgé and Massart [8] and Arlot and Massart [3].

The paper is organized as follows. First, we consider in Section 2 principal curves with bounded length and show that the polygonal line obtained by minimizing some appropriate penalized criterion satisfies an oracle-type inequality. Section 3 provides a similar result in the context of principal curves with bounded turn. Our theoretical findings are illustrated on both simulated and real data sets in Section 4. For the sake of clarity, proofs are collected in Section 5.

2 Principal curves with bounded length

Let $\| \cdot \|$ be the standard Euclidean norm over $\mathbb{R}^d$. A parameterized curve in $\mathbb{R}^d$ is a continuous function

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

where $I = [a, b]$ is a closed interval of the real line. The length of $\mathbf{f}$ is defined by

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

where the supremum is taken over all subdivisions $a = t_0 < t_1 < \cdots < t_m = b$, $m \geq 1$ (see, e.g., Kolmogorov and Fomin [30]). Throughout the document, it is assumed that $\mathbb{E}\|\mathbf{X}\|^2 < \infty$ and that

$$\mathbb{P}\left\{\mathbf{X} \in \mathcal{C}\right\} = 1, \tag{2}$$

where $\mathcal{C}$ is a convex compact subset of $\mathbb{R}^d$, with diameter δ. By Lemma 1 in Kégl [27], the requirement (2) implies that, for any given positive length L, there exists a principal curve for $\mathbf{X}$ with length at most L in $\mathcal{C}$, that is a (non necessarily unique) parameterized curve $\mathbf{f}^\star$ with length not larger than L and support in $\mathcal{C}$ achieving the minimum of $\mathbb{E}\left[\inf_{t \in I} \|\mathbf{X} - \mathbf{f}(t)\|^2\right]$. Consequently, in the sequel, we will restrict ourselves to curves whose support is included in $\mathcal{C}$ and denote by $\mathcal{F}$ the set of all parameterized curves $\mathbf{f} = (f_1, \ldots, f_d)$ belonging to $\mathcal{C}$.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a sample of independent random variables distributed as $\mathbf{X}$, and consider the contrast

$$\Delta(\mathbf{f}, \mathbf{x}) = \inf_{t \in I} \|\mathbf{x} - \mathbf{f}(t)\|^2, \quad \mathbf{f} \in \mathcal{F}, \mathbf{x} \in \mathbb{R}^d.$$

The associated empirical risk based on the sample $\mathbf{X}_1, \dots \mathbf{X}_n$ is defined as

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

For some prespecified length $L > 0$, we set

$$\mathbf{f}^\star \in \underset{\mathbf{f} \in \mathcal{F}, \mathscr{L}(\mathbf{f}) \leq L}{\arg\min} \ \mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})].$$

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

$$\hat{\mathbf{f}}_{k,\ell} \in \underset{\mathbf{f} \in \mathcal{F}_{k,\ell}}{\arg\min} \Delta_n(\mathbf{f})$$

be a curve achieving the minimum of the empirical criterion $\Delta_n(\mathbf{f})$ over the polygonal line class $\mathcal{F}_{k,\ell}$.

At this stage of the procedure, we have at hand a family of estimates $\{\hat{\mathbf{f}}_{k,\ell}\}_{k \geq 1, \ell \in \mathcal{L}}$ and our goal is to select the best principal curve $\tilde{\mathbf{f}}$ among this collection. To this aim, we make use of the model selection approach of Barron, Birgé and Massart [5], which allows to assess the adjustment quality by controlling the loss

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

between the target $\mathbf{f}^\star$ and the selected curve $\tilde{\mathbf{f}}$. (For a comprehensive introduction to the area of model selection, the reader is referred to the monograph of Massart [35].) More formally, let pen $: \mathbb{N}^* \times \mathcal{L} \to \mathbb{R}^+$ be some penalty function and denote by $(\hat{k}, \hat{\ell})$ a pair of minimizers of the criterion

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

In order to obtain the desired principal curve $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k}, \hat{\ell}}$, we have to design an adequate penalty $\mathrm{pen}(k, \ell)$. This is done in the following theorem, which is an adaptation of a general model selection result of Massart [35, Theorem 8.1]. However, for the sake of completeness, it is proved in its full length in Section 5.

Theorem 2.1. *Consider a family of nonnegative weights* $\{x_{k,\ell}\}_{k \geq 1, \ell \in \mathcal{L}}$ *such that*

$$\sum_{k \geq 1, \ell \in \mathcal{L}} e^{-x_{k,\ell}} = \Sigma < \infty,$$

and a penalty function $\text{pen} : \mathbb{N}^* \times \mathcal{L} \to \mathbb{R}^+$. *Let* $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k}, \hat{\ell}}$. *If for all* $(k, \ell) \in \mathbb{N}^* \times \mathcal{L}$,

$$\text{pen}(k, \ell) \geq \mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k,\ell}} \left(\mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})] - \Delta_n(\mathbf{f}) \right) \right] + \delta^2 \sqrt{\frac{x_{k,\ell}}{2n}},$$

then

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

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

Theorem 2.1 offers a nonasymptotic bound, expressing the fact that the expected loss of the final estimate $\tilde{\mathbf{f}}$ is close to the minimal loss over all $k \geq 1$ and $\ell \in \mathcal{L}$, up to a term tending to 0. Thus, in order to apply this theorem to the principal curve problem, we now have to find an upper bound on the quantity

$$\mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k,\ell}} \left(\mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})] - \Delta_n(\mathbf{f}) \right) \right]. \tag{3}$$

This is achieved by Proposition 2.1 below, which is proved by showing that the expected maximal deviation (3) may be bounded by a Rademacher average (see Bartlett, Boucheron, and Lugosi [6] and Koltchinskii [31]) and by resorting to a Dudley integral (Dudley [17]).

Proposition 2.1. *Let* $\mathcal{F}_{k,\ell}$ *be the set of all polygonal lines with* k *segments, length at most* ℓ, *and vertices in a grid* $\mathcal{Q} \subset \mathcal{C}$. *Then there exist nonnegative constants* $a_0, \ldots, a_2$, *depending on the maximal length* L, *the dimension* d *and the diameter* δ *of the convex set* $\mathcal{C}$, *such that*

$$\mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k,\ell}} \left(\mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})] - \Delta_n(\mathbf{f}) \right) \right] \leq \frac{1}{\sqrt{n}} \left[a_1 \sqrt{k} + a_2 \ell + a_0 \right].$$

Finally, by combining Theorem 2.1 and Proposition 2.1, we are in a position to state the main result of this section.

Theorem 2.2. *Consider a family of nonnegative weights* $\{x_{k,\ell}\}_{k \geq 1, \ell \in \mathcal{L}}$ *such that*

$$\sum_{k \geq 1, \ell \in \mathcal{L}} e^{-x_{k,\ell}} = \Sigma < \infty,$$

and a penalty function $\text{pen} : \mathbb{N}^* \times \mathcal{L} \to \mathbb{R}^+$. *Let* $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k}, \hat{\ell}}$. *There exist nonnegative constants* $c_0, \ldots, c_2$, *depending on the maximal length* L, *the dimension* d *and the diameter* δ *of the convex set* $\mathcal{C}$, *such that, if for all* $(k, \ell) \in \mathbb{N}^* \times \mathcal{L}$,

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

then

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

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

Some comments are in order.

Firstly, we see that the penalty shape involves a term proportional to $\sqrt{k/n}$ and a term proportional to $\ell/\sqrt{n}$. 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.

Observe next that the proof of Proposition 2.1 provides possible values for the constants $c_0, \ldots, c_2$. However, these values are not very helpful since they are upper bounds which are probably far from being tight. Nevertheless, the proof also reveals that $c_1 = c_1' \delta^2$, $c_2 = c_2' \delta$ 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 $\Delta_n(\mathbf{f})$ is.

Finally, an important practical issue is how to choose the weights $\{x_{k,\ell}\}_{k \geq 1, \ell \in \mathcal{L}}$. These weights should be large enough to ensure the finiteness of Σ, but not too large at the risk of overpenalizing. If the cardinality of the collection of models is not larger than n^2 (this will be the case in all our practical examples), we may set $x_{k,\ell} = 2 \ln n$ for every (k, ℓ). This choice does not affect the penalty shape, though modifying the rate, and leads to $\Sigma = 1$ in the risk bound.

Remark 2.1. Clearly, when the length ℓ of polygonal lines is fixed, and the aim is to select the number k of segments, the dominant term reflecting the complexity of the models in the penalty is $\sqrt{k/n}$. In this particular context, the weights may be taken equal to $\ln n$, or, by analogy with the Gaussian linear model selection framework, proportional to k. Indeed, in this framework, each model S_m, $m \in \mathcal{M}$, has dimension D_m and an interesting choice for x_m is then $x_m = x(D_m)$, where $x(D) = cD + \ln |\{m \in \mathcal{M} : D_m = D\}|$ and $c > 0$. When there is no redundancy in the models dimension, this strategy amounts to choosing x_m proportional to D_m. In our problem, this means setting $x_k = ck$ for every k, where the constant $c > 0$ ensures that $\sum_{k \geq 1} e^{-ck} = \Sigma < \infty$. Thus, in this somewhat restrictive situation, the penalty is of the order $\sqrt{k/n}$.

3 Principal curves with bounded turn

As it was already mentioned in the Introduction, Sandilya and Kulkarni [39] (**sk**) suggest an alternative approach for principal curves, based on the control of the turn. Recall that the turn $\mathscr{K}(\mathbf{f})$ of a curve $\mathbf{f} : I \to \mathbb{R}^d$, $I = [a, b]$, is given by

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

where $\widehat{f(t_j)}$ denotes the angle between the vectors $\overrightarrow{f(t_{j-1})f(t_j)}$ and $\overrightarrow{f(t_j)f(t_{j+1})}$, and the supremum is taken over all subdivisions $a = t_0 < t_1 < \cdots < t_m = b$, $m \geq 1$ (Alexandrov and Reshetnyak [2]). Thus, the turn of a polygonal line with vertices $v_1, \ldots, v_{k+1}$ is just the sum of the angles at $v_2, \ldots, v_k$ (see Figure 4 for an illustration).
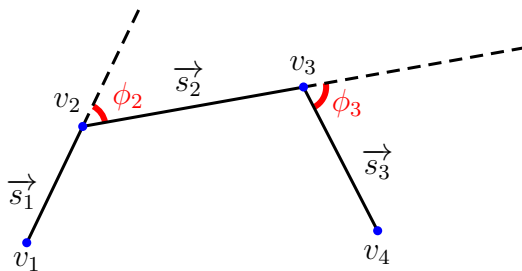

Figure 4: Denoting by $\overrightarrow{s_j}$ the vector $\overrightarrow{v_j v_{j+1}}$ for all $j = 1, \ldots, k$, the angles involved in the definition of the turn are defined by $\phi_{j+1} = (\overrightarrow{s_j}, \overrightarrow{s_{j+1}})$.

As a logical continuation to Section 2, we propose in the present section to analyse the **sk** definition from a model selection point of view. To this aim, we use the fact that a curve with bounded turn also has bounded length, as shown in Lemma 3.1 below.

We still assume that $\mathbb{P}\{\mathbf{X} \in \mathcal{C}\} = 1$, where $\mathcal{C}$ is a convex compact subset of $\mathbb{R}^d$ with diameter δ. By Proposition 1 in **sk**, this requirement ensures the existence of a curve $\mathbf{f}^\star$ with bounded turn minimizing the criterion $\mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})]$. More formally, for some prespecified turn $K \geq 0$, we set

$$\mathbf{f}^\star \in \operatorname*{arg\,min}_{\mathbf{f} \in \mathcal{F}, \mathscr{K}(\mathbf{f}) \leq K} \mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})],$$

where $\mathscr{K}(\mathbf{f})$ denotes the turn of $\mathbf{f}$. Proceeding as in Section 2, we let $\mathcal{K}$ be a countable subset of $[0, K]$ and define a countable collection of models $\{\mathcal{F}_{k,\kappa}\}_{k \geq 1, \kappa \in \mathcal{K}}$

as follows. Each $\mathcal{F}_{k,\kappa}$ consists of polygonal lines with k segments, with turn at most κ, and with vertices belonging to some grid $\mathcal{Q}$ over $\mathcal{C}$. For $k \geq 1$ and $\kappa \in \mathcal{K}$, define

$$\hat{\mathbf{f}}_{k,\kappa} \in \underset{\mathbf{f} \in \mathcal{F}_{k,\kappa}}{\arg\min} \, \Delta_n(\mathbf{f})$$

to be a polygonal line minimizing the empirical criterion $\Delta_n(\mathbf{f})$ over $\mathcal{F}_{k,\kappa}$. We wish to design an appropriate penalty function pen $: \mathbb{N}^* \times \mathcal{K} \to \mathbb{R}^+$ and minimize the criterion

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

in order to obtain a suitable principal curve. As before, we let $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k},\hat{\kappa}}$, where $(\hat{k}, \hat{\kappa})$ is a minimizer of the penalized criterion $\mathrm{crit}(k, \kappa)$, and intend to control the loss $\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) = \mathbb{E}[\Delta(\tilde{\mathbf{f}}, \mathbf{X}) - \Delta(\mathbf{f}^\star, \mathbf{X})]$.

To get a result of the form of Theorem 2.2, we already know that it suffices to find an upper bound on the quantity

$$\mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k,\kappa}} \left(\mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})] - \Delta_n(\mathbf{f}) \right) \right].$$

As a first step towards this direction, we will need the following lemma, which establishes an interesting link between the length of a curve and its turn. For a proof of this result, we refer the reader to Alexandrov and Reshetnyak [2, Chapter 5].

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

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

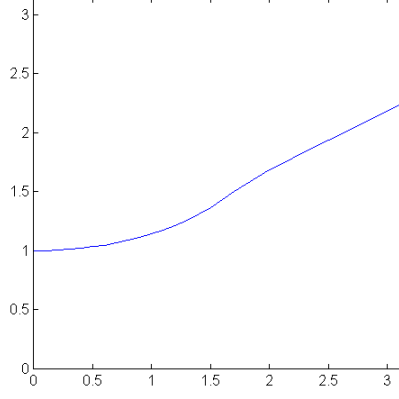The graph of the function ζ is shown in Figure 5.

Figure 5: Graph of the function ζ.

Thanks to this result, the approach developed in Section 2 adapts to the new context. Proposition 3.1 below is the counterpart of Proposition 2.1.

Proposition 3.1. *Let $\mathcal{F}_{k,\kappa}$ be the set of all polygonal lines with k segments, turn at most κ, and vertices in a grid $\mathcal{Q} \subset \mathcal{C}$, and let δ be the diameter of the convex set $\mathcal{C}$. Then there exist nonnegative constants $a_0, \ldots, a_4$, depending only on the dimension d, such that*

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\kappa}} \left(\mathbb{E}[\Delta(\mathbf{f},\mathbf{X})] - \Delta_n(\mathbf{f})\right)\right]$$
$$\leq \delta^2\left[a_1\sqrt{k} + a_2\sqrt{\zeta(\kappa)} + a_3\frac{\zeta(\kappa)}{\sqrt{k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}<1\}} + a_4\sqrt{k\ln\frac{\zeta(\kappa)}{k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}\geq1\}} + a_0\right].$$

Putting finally Theorem 2.1 and Proposition 3.1 together, we obtain:

Theorem 3.1. *Consider a family of nonnegative weights $\{x_{k,\kappa}\}_{k\geq1,\kappa\in\mathcal{K}}$ such that*

$$\sum_{k\geq1,\kappa\in\mathcal{K}} e^{-x_{k,\kappa}} = \Sigma < \infty,$$

and a penalty function pen $: \mathbb{N}^* \times \mathcal{K} \to \mathbb{R}^+$. *Let $\tilde{\mathbf{f}} = \hat{\mathbf{f}}_{\hat{k},\hat{\kappa}}$. There exist nonnegative constants $c_0, \ldots, c_2$, depending only on the dimension d, such that, if for all $(k,\kappa) \in \mathbb{N}^* \times \mathcal{K}$,*

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

12

then

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

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

The expression of the penalty shape involves a term of the order $\sqrt{k/n}$—just like in the case of curves with bounded length—, whereas the length ℓ is replaced by $\zeta(\kappa)$, which is an increasing function of the turn κ. This is relevant, since the number of segments k and the turn κ characterize the complexity of the models. Moreover, the additive term $\max \left[\zeta(\kappa)/\sqrt{kn}, \sqrt{k \ln \zeta(\kappa)/kn} \right]$ shows that k and κ should be cleverly chosen relatively to each other in order to get a nice principal curve. Roughly, a greater curvature implies more segments.

4 Experimental results

This section presents some simulations and real data experiments, carried out with the software MATLAB, to illustrate the model selection procedure suggested by Theorem 2.2. The penalty shapes in the theorem involve constants which have to be practically determined. To this end, a possible route is to use the so-called slope heuristics, introduced by Birgé and Massart [8] and further developed by Arlot and Massart [3] (see also Lerasle [33] and Saumard [40]). In short, this calibration method allows to tune a penalty known up to some multiplicative constant. The slope heuristics assumes that the empirical contrast decreases when the complexity of the models increases, which is clearly the case in our principal curve context. The procedure is based on the fact that the graph of the empirical contrast as a function of the penalty shape decreases strongly at the beginning and more slowly later, with a linear trend. At the end, the heuristics specifies that the desired constant is equal to $-2s$, where s is the slope of this line.

Hence, in the sequel, the number k of segments and the length ℓ of the principal curve are chosen according to the following strategy, denoted hereafter by **MS**:

Algorithm **MS**

---

1. For each number k of segments, $k = 1, \ldots, 80$, and for a range of values of the length ℓ, compute $\hat{\mathbf{f}}_{k,\ell}$ by minimizing the empirical criterion $\Delta_n(\mathbf{f})$ and record

$$\Delta_n(\hat{\mathbf{f}}_{k,\ell}) = \frac{1}{n} \sum_{i=1}^{n} \Delta(\hat{\mathbf{f}}_{k,\ell}, \mathbf{X}_i).$$

2. Set $x_{k,\ell} = 2 \ln n$ and consider a penalty of the form

$$\mathrm{pen}(k, \ell) = c_1 \sqrt{k} + c_2 \ell.$$

3. Select the constants $\hat{c}_1$ and $\hat{c}_2$ using a bivariate version of the slope heuristics.

4. Retain the curve $\hat{\mathbf{f}}_{\hat{k},\hat{\ell}}$ obtained by minimizing the penalized criterion
$$\mathrm{crit}(k, \ell) = \Delta_n(\hat{\mathbf{f}}_{k,\ell}) - 2(\hat{c}_1 \sqrt{k} + \hat{c}_2 \ell).$$

Throughout this experimental section, the maximal values of the parameters have been chosen to be reasonably large without increasing the computation time uselessly. The maximal length and the step defining the range of values of ℓ depend on the scale of the considered data set.

The minimization of the criterion $\Delta_n(\mathbf{f})$ (step 1 of the algorithm) is achieved through a MATLAB optimization routine.

The weights $x_{k,\ell}$ were all set to $2 \ln n$. We realize that this choice is somewhat arbitrary. However, as mentioned in the discussion after Theorem 2.2, as soon as the number of models is not larger than n^2—which is clearly the case in our examples—, this is a convenient choice, which, moreover, does not modify the penalty shape. Besides, the calibration of c_0 is a challenging question which has been given little consideration in the literature so far, even in the standard slope heuristics context. Note that a possible route to take the constant term into account was proposed by Lebarbier [32]. Here, in our bivariate framework, we deal, to a first approximation, with a penalty of the form $c_1 \sqrt{k} + c_2 \ell$.

14

Assessing the values of the constants via the slope heuristics rests upon the assumption that, for large values of k and ℓ, $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ behaves like $c_1\sqrt{k} + c_2\ell$. The constants $\hat{c}_1$ and $\hat{c}_2$ are then chosen via an ordinary least square regression and we compute the corresponding R^2 coefficient to measure the quality of the regression. We also tried a robust regression, whose results were observed to be very similar, and thus, are not reported here.

Finally, the results of the algorithm **MS** were systematically compared to the outputs of the Polygonal Line Algorithm of Kégl, Krzyżak, Linder, and Zeger [29]. In short, this procedure optimizes the vertices of the curve one after the other, using a local version of the criterion $\Delta_n(\mathbf{f})$, which relies on a local angle penalty. To our knowledge, this heuristic technique is not supported by any theoretical result. However, it is known to perform well and should in our context be understood as a benchmark.

4.1 Simulated data

In this first series of experiments, we considered two-dimensional data distributed with some noise around a reference curve. More formally, observations were generated from the model

$$\mathbf{X} = \mathbf{Y} + \varepsilon,$$

where $\mathbf{Y}$ is uniformly distributed over some planar curve $\mathbf{f}$ and ε is a bivariate Gaussian noise, independent of $\mathbf{Y}$. Even if the generative curve $\mathbf{f}$ is not a principal curve stricto sensu—because of the model bias—, this Gaussian model is considered as a benchmark for simulations in the literature on principal curves.

The union of the generative curve and the estimated curve can be seen as a self-intersecting polygon, the area of which may be used to quantitatively asses how far the estimated curve is from the true one. In the sequel, we compute for each simulated example an error criterion corresponding to the average area over 20 trials, normalized with respect to the scale of the data.

In a first example, we let $\mathbf{f}$ be a half-circle with radius 1. The noise variance is set to 0.004 and the number n of observations to 100 (see Figure 6).
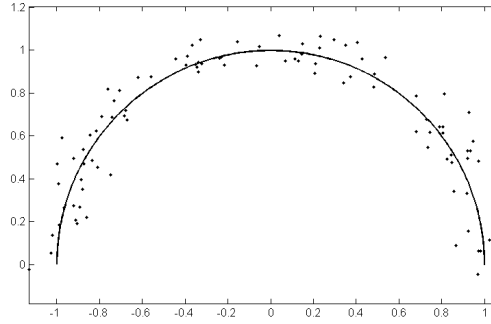
Figure 6: 100 observations distributed around a half-circle with radius 1.

Recall that the algorithm **MS** computes the criterion $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ for a table of values of $\sqrt{k}$ and ℓ and selects the best constants according to a bivariate slope heuristics. Figure 7 shows the contour plot of $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ as a function of $\sqrt{k}$ and ℓ, which supports the idea that this function is linear in $\sqrt{k}$ and ℓ when k and ℓ become large. The irregularities reflect the fact that the criterion $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ is not decreasing continuously when increasing the parameters, though decreasing on the whole. This phenomenon, which also appears in the Polygonal Line Algorithm (**PL** hereafter), is due to a convergence problem related to the optimization function.
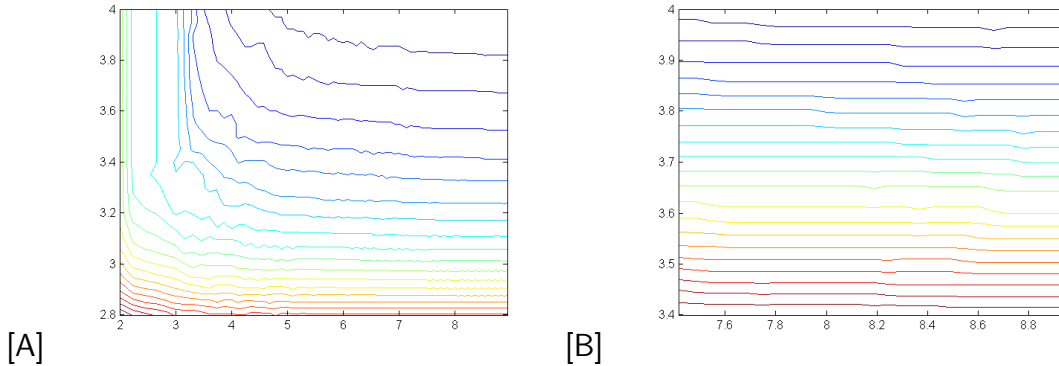

[A]   [B]

Figure 7: [A] Contour plot of $\Delta_n(\hat{\mathbf{f}}_{k,\ell})$ as a function of $\sqrt{k}$ and ℓ for the half-circle data (n=100). [B] Zoom in the zone of large values of k and ℓ.

Both algorithms were applied to the data set. The resulting principal curves are visible in Figure 8. For comparison purposes, Figure 9 also show some curves obtained by minimizing $\Delta_n(\mathbf{f})$ for other values of k and ℓ. The average R^2 over 20 trials corresponding to the regression used in **MS** is 0.98 and the error criterion equals 0.030.
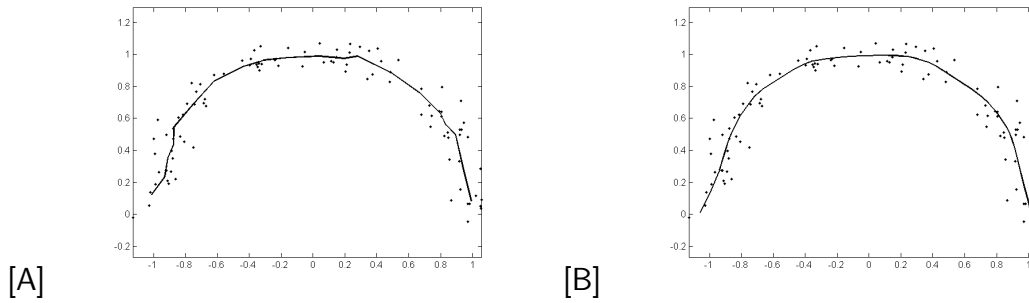
Figure 8: Selected principal curves for the half-circle data (n=100). [A] Method **MS**: $\hat{k} = 20$, $\hat{\ell} = 3$. [B] **PL** algorithm.

It can be noted that the outputs of both algorithms have approximately the same quality, despite a few irregularities on the **MS** principal curve, not visible on the **PL** result.
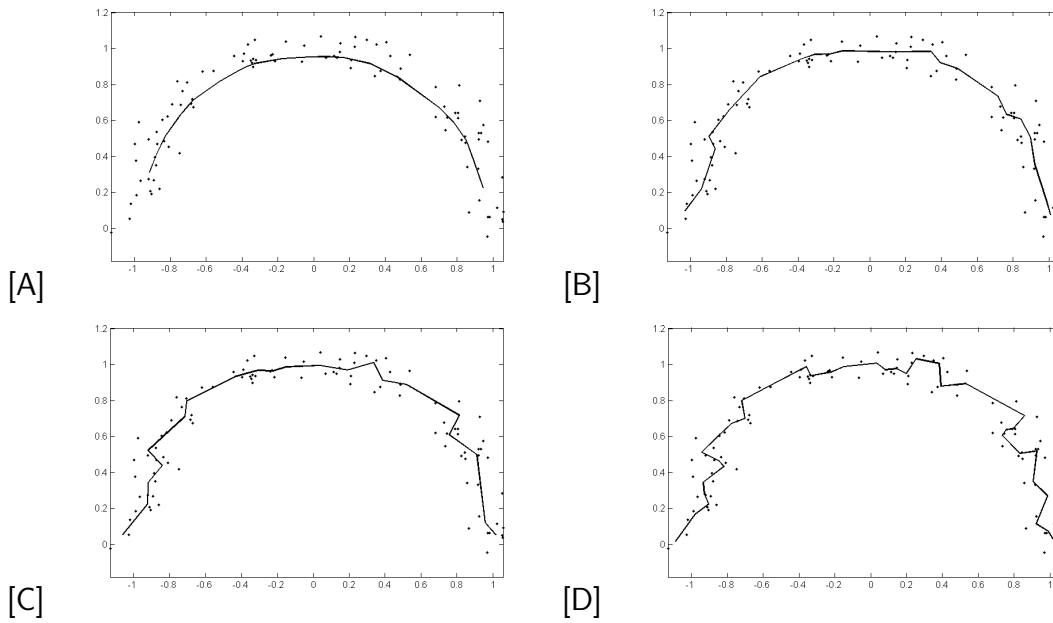

Figure 9: Method **MS**: Examples of principal curves for some prespecified values of k and ℓ (n=100). [A] $k = 20$, $\ell = 2.5$. [B] $k = 20$, $\ell = 3.1$. [C] $k = 20$, $\ell = 3.4$. [D] $k = 35$, $\ell = 4$.

The methods **MS** and **PL** were also tested on a larger sample $n = 250$. We observed that both principal curves obtained with this sample size are very accurate.

In a second set of numerical examples, we took handwritten-type digits as generative curves, with noise variance 0.04. As depicted in Figure 10, 150 observations were sampled around the digit 2 and the digit 3 and 250 observations around the digit 5.
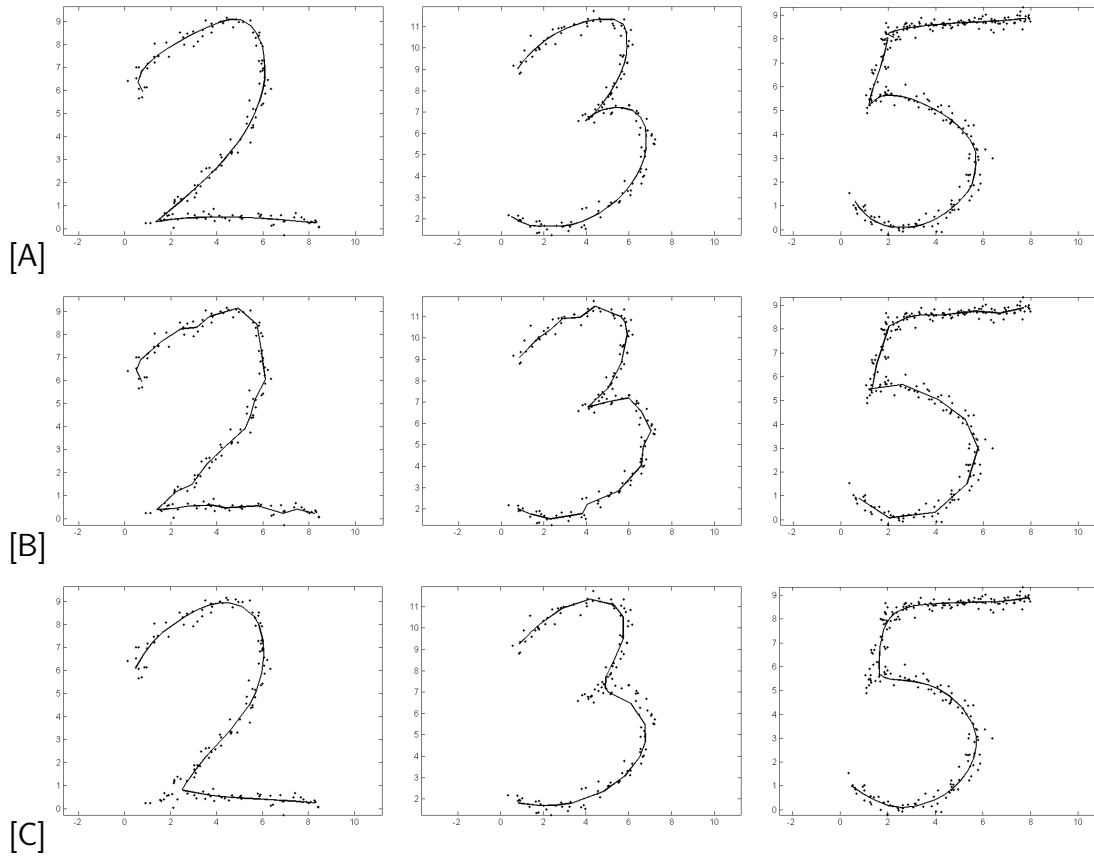

Figure 10: [A] Observations sampled around the digit 2 (n=150), 3 (n=150) and 5 (n=250). [B] Principal curves selected by the method **MS**: $\hat{k} = 27$, $\hat{\ell} = 24$; $\hat{k} = 23$, $\hat{\ell} = 23$; $\hat{k} = 17$, $\hat{\ell} = 21$. [C] **PL** principal curves.

With respect to the digit 2 data, the **MS** principal curve follows the observations more closely than what is expected. On the other hand, the **PL** output looks smoother, but a comparison with the generative curve shows that the loop at the top and the angle at the bottom of the digit 2 are not recovered precisely. For the digit 3, we note again that the algorithm **MS** slightly overfits the data, whereas the smoother curve **PL** misses the angle. The same comment holds for the digit 5, but to a lesser degree. On this last example, both algorithms performed quite similarly and the resulting principal curves are visually satisfactory. For these simulated digits, the average R^2 coefficients equal respectively 0.87, 0.91 and 0.93 and the error areas 0.032, 0.026 and 0.021.

This small simulation study reveals, as expected, that a good automatic choice of the parameters $\hat{k}$ and $\hat{\ell}$ is crucial to obtain a suitable principal curve. On the whole, the visual quality of **MS** is fully acceptable, even if the principal curves fitted by this algorithm often follow the data quite closely, in particular when the sample size is not very large. In return, the global shape of the digit is better recovered than using **PL**.

4.2 Real data sets

4.2.1 NIST database digits

The first real-life data set used in this second series of experiments originated from NIST Special Database 19 (`http://www.nist.gov/srd/nistsd19.cfm`), containing handwritten characters from 3600 writers. The data consists in binary images scanned at 11.8 dots per millimeter (300 dpi), which uniformly fill the area corresponding to the thickness of the pen stroke. 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 [15] and Alcorn and Hoggar [1]).

Algorithms **MS** and **PL** were applied to the three NIST database digits visible in Figure 11.
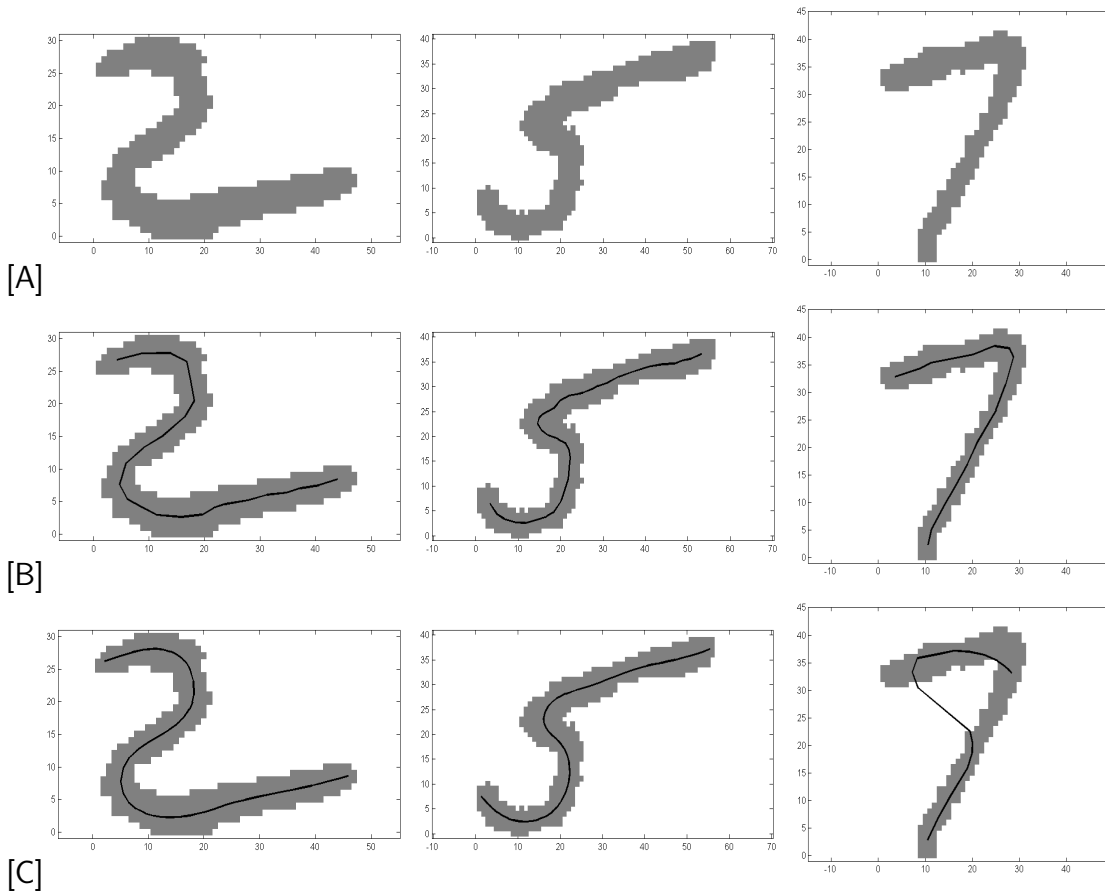
Figure 11: [A] Three NIST database handwritten digits. [B] Principal curves selected by the method MS: $\hat{k} = 23$, $\hat{\ell} = 80$; $\hat{k} = 38$, $\hat{\ell} = 82$; $\hat{k} = 15$, $\hat{\ell} = 66$. [C] PL principal curves.

We observe that both results for the digit 2 are similar and completely satisfactory. Regarding the digit 5, **MS** seems to better recover the overall shape. Finally, the principal curve fitted by **MS** for the digit 7 is suitable, whereas the output of **PL** looks definitely not satisfactory. R^2 coefficients equal to 0.99, 0.99 and 0.95 respectively were obtained in the **MS** regression step.

As a general conclusion on these NIST digit data sets, we found that **MS** performs well. Here, the algorithm does not seem to overfit, probably because the sample size is large enough.

4.2.2 Seismic data

Together with satellite images, the localization of earthquakes is an essential source of information in geology for the study of seismic faults, whether in accretion or subduction regions. As an illustration, Figure 12 depicts seismic impacts in the world—the map is drawn using Miller's projection—, as well as a world map from the USGS (United States Geological Survey) showing the various lithospheric plates. The data set, which can be downloaded on the USGS website (`http://earthquake.usgs.gov/research/data/centennial.php`), is part of the "Centennial Catalog", listing the major earthquakes registered since 1900 (Engdahl et Villaseñor [21]). In this subsection, we employ algorithm **MS** as a means to recover the borders of lithospheric plates using the earthquake localization data of Figure 12. Again, the **PL** output is given as a benchmark.

We decided to focus on two particularly representative seismic active zones. The first one (**Z1** hereafter) is located in the Atlantic Ocean, to the west of the African continent (about 60°S 50°W to 40°N 0°), and the second one (**Z2** hereafter) extends from the south of Africa to the south of Australia (about 65°S 0° to 25°S 160°E). The localization of these two regions on the world map is visible in Figure 13. The results for **Z1** are shown in Figure 14 and for **Z2** in Figure 15.
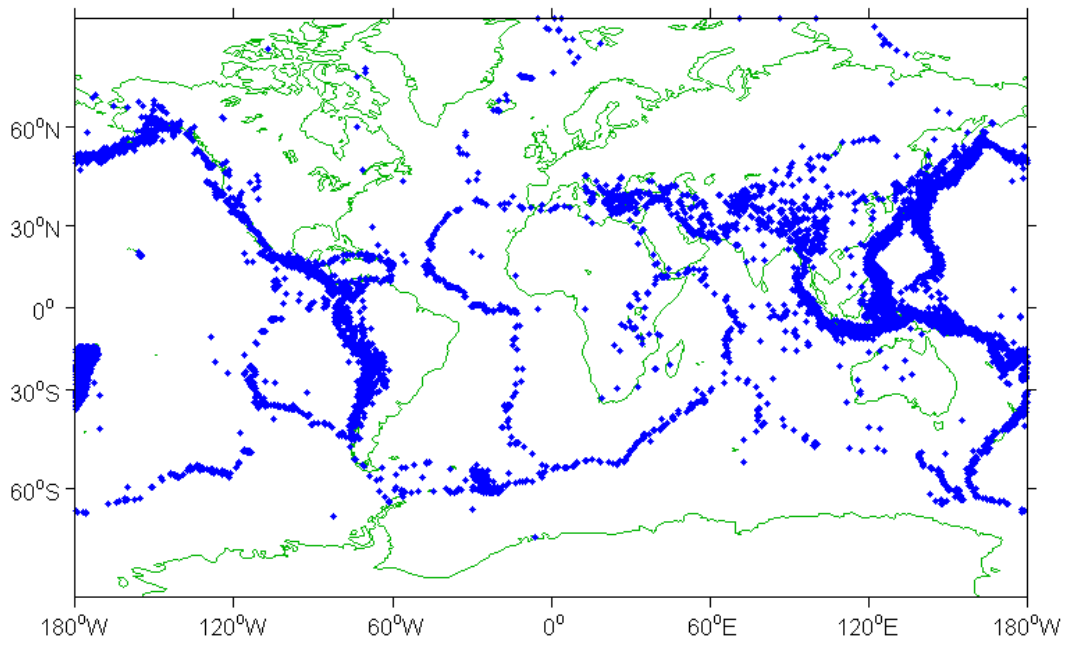
[A]
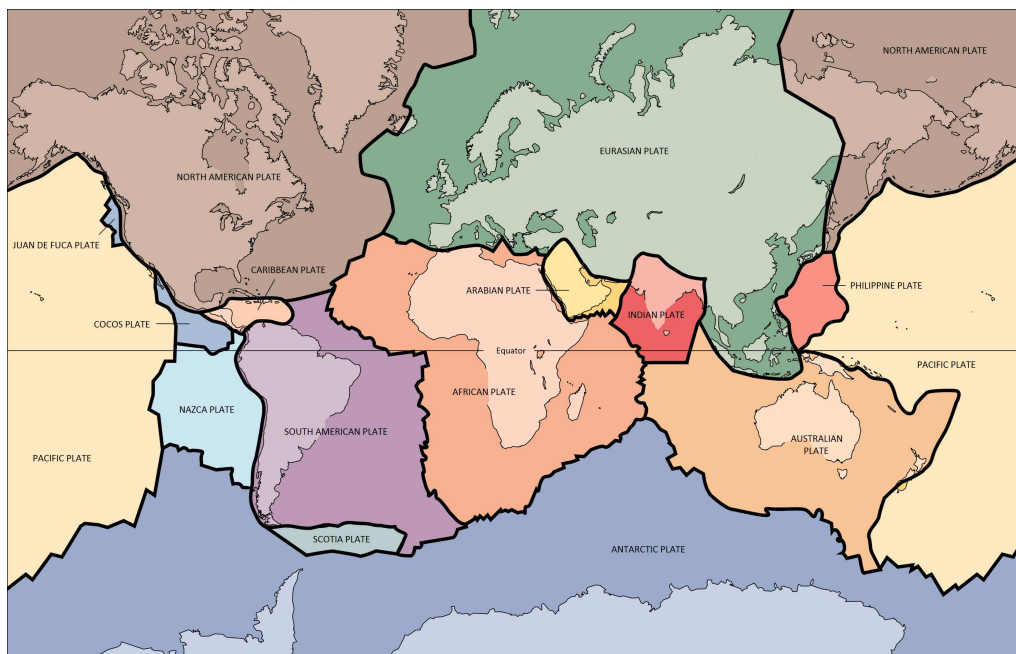
[B]

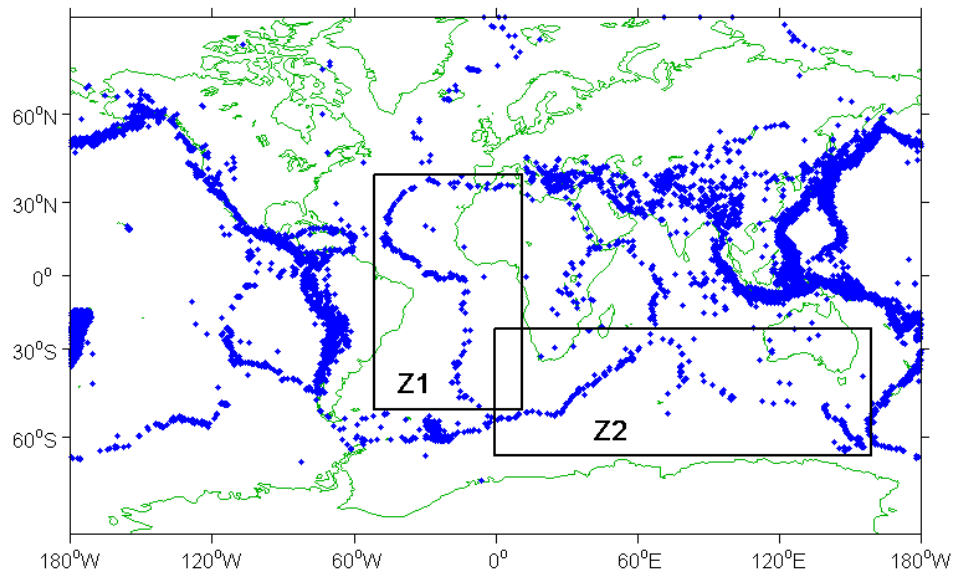Figure 12: [A] Earthquake impacts and [B] lithospheric plate borders.

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).
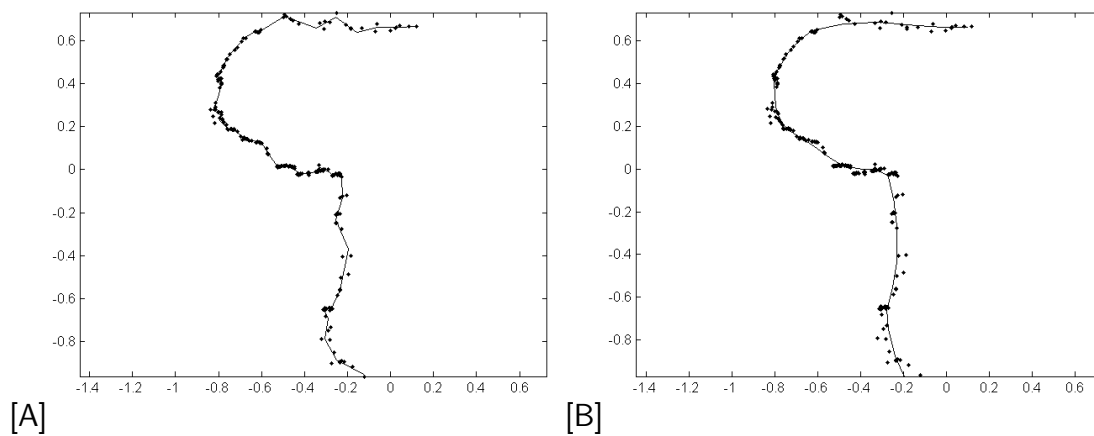

[A]

[B]

Figure 14: Selected principal curves for the seismic zone **Z1** (n=252). [A] Method **MS**: $\hat{k} = 55$, $\hat{\ell} = 31$. [B] **PL** principal curve.
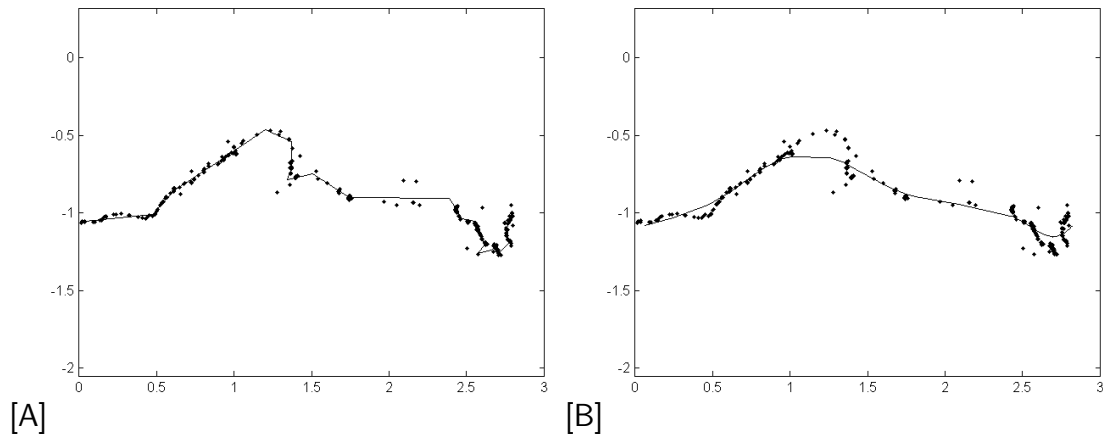
Figure 15: Selected principal curves for the seismic zone **Z2** (n=322). [A] Method
MS: $\hat{k} = 22$, $\hat{\ell} = 38$. [B] **PL** principal curve.

In Figure 14, we see, for the seismic zone **Z1**, that the method **MS** again yields a principal curve following the data points quite closely. On the contrary, the **PL** algorithm provides a smoother curve, which at first sight seems a better result. However, the border of the lithospheric plate is probably more likely to look like the more irregular **MS** principal curve, as suggested by Figure 12 [B]. The same observation holds for **Z2** (Figure 15). Moreover, in this case, the **PL** output does not recover the shape of the plate border, which certainly passes through the most northern points and not several degrees south. Apparently, the local penalty on the angles leads here to overpenalization. Thus, on this seismic data set, **MS** results seem to be more relevant.

It is noteworthy that using this type of earthquake data to draw faults could be especially useful to locate some faults which cannot be easily spotted and necessitate monitoring for seismic risk prevention. With this respect, Harding and Berghoff [24], employing a method based on airborne laser mapping, study for instance seismic hazards in a zone densely covered by vegetation, located in the Puget Lowland of Washington State, USA. Using a principal curve approach to solve this kind of problems is undoubtedly an interesting project for future research.

24

5 Proofs

5.1 Proof of Theorem 2.1

Theorem 2.1 is an adaptation of Theorem 8.1 in Massart [35]. We first recall the following lemma, which is a consequence of McDiarmid's inequality [36] (see Massart [35, Theorem 5.3]).

Lemma 5.1. *If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent random variables and $\mathcal{G}$ is a finite or countable class of real-valued functions such that $a \leq g \leq b$ for every function $g \in \mathcal{G}$, then, setting $Z = \sup_{g \in \mathcal{G}} \sum_{i=1}^{n} (g(\mathbf{X}_i) - \mathbb{E}[g(\mathbf{X}_i)])$, we have, for every $\varepsilon \geq 0$,*

$$\mathbb{P}\{Z - \mathbb{E}[Z] \geq \varepsilon\} \leq \exp\left(-\frac{2\varepsilon^2}{n(b-a)^2}\right).$$

Proof of the theorem. Let $\bar{\Delta}_n(\mathbf{f}) = \Delta_n(\mathbf{f}) - \mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})]$ denote the centered empirical process. For all $k \geq 1$ and $\ell \in \mathcal{L}$, for any $\mathbf{f}_{k,\ell} \in \mathcal{F}_{k,\ell}$, we have, by definition of $\tilde{\mathbf{f}}$,

$$\Delta_n(\tilde{\mathbf{f}}) + \text{pen}(\hat{k}, \hat{\ell}) \leq \Delta_n(\mathbf{f}_{k,\ell}) + \text{pen}(k, \ell).$$

Equivalently,

$$\Delta_n(\tilde{\mathbf{f}}) - \Delta_n(\mathbf{f}_{k,\ell}) \leq \text{pen}(k, \ell) - \text{pen}(\hat{k}, \hat{\ell}).$$

Since $\Delta_n(\tilde{\mathbf{f}}) = \mathbb{E}[\Delta(\tilde{\mathbf{f}}, \mathbf{X})] + \bar{\Delta}_n(\tilde{\mathbf{f}})$ and $\Delta_n(\mathbf{f}_{k,\ell}) = \mathbb{E}[\Delta(\mathbf{f}_{k,\ell}, \mathbf{X})] + \bar{\Delta}_n(\mathbf{f}_{k,\ell})$, this inequality becomes

$$\mathbb{E}[\Delta(\tilde{\mathbf{f}}, \mathbf{X})] - \mathbb{E}[\Delta(\mathbf{f}_{k,\ell}, \mathbf{X})] \leq \bar{\Delta}_n(\mathbf{f}_{k,\ell}) - \bar{\Delta}_n(\tilde{\mathbf{f}}) + \text{pen}(k, \ell) - \text{pen}(\hat{k}, \hat{\ell}). \quad (4)$$

Moreover, for every $\mathbf{f} \in \mathcal{F}$,

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

so that

$$\mathbb{E}[\Delta(\tilde{\mathbf{f}}, \mathbf{X})] - \mathbb{E}[\Delta(\mathbf{f}_{k,\ell}, \mathbf{X})] = \mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) - \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}). \quad (5)$$

Therefore, combining (4) and (5),

$$\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) \leq \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) - \bar{\Delta}_n(\tilde{\mathbf{f}}) + \text{pen}(k, \ell) - \text{pen}(\hat{k}, \hat{\ell}). \quad (6)$$

Consider now a family of nonnegative weights $\{x_{k,\ell}\}_{k \geq 1, \ell \in \mathcal{L}}$ such that

$$\sum_{k \geq 1, \ell \in \mathcal{L}} e^{-x_{k,\ell}} = \Sigma < \infty,$$

and let $z > 0$. Applying Lemma 5.1, we get, for all $k' \geq 1$, $\ell' \in \mathcal{L}$ and $\varepsilon \geq 0$,

$$\mathbb{P}\left\{ \sup_{\mathbf{f} \in \mathcal{F}_{k',\ell'}} (-\bar{\Delta}_n(\mathbf{f})) \geq \mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k',\ell'}} (-\bar{\Delta}_n(\mathbf{f})) \right] + \varepsilon \right\} \leq \exp\left(-\frac{2n\varepsilon^2}{\delta^4} \right).$$

This may be rewritten, for $\varepsilon = \delta^2 \sqrt{\frac{x_{k',\ell'}+z}{2n}}$,

$$\mathbb{P}\left\{ \sup_{\mathbf{f} \in \mathcal{F}_{k',\ell'}} (-\bar{\Delta}_n(\mathbf{f})) \geq \mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k',\ell'}} (-\bar{\Delta}_n(\mathbf{f})) \right] + \delta^2 \sqrt{\frac{x_{k',\ell'}+z}{2n}} \right\} \leq e^{-x_{k',\ell'}-z}.$$

Setting $E_{k',\ell'} = \mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k',\ell'}} (-\bar{\Delta}_n(\mathbf{f})) \right]$, we thus have, for all $k' \geq 1$ and $\ell' \in \mathcal{L}$,

$$\sup_{\mathbf{f} \in \mathcal{F}_{k',\ell'}} (-\bar{\Delta}_n(\mathbf{f})) \leq E_{k',\ell'} + \delta^2 \sqrt{\frac{x_{k',\ell'}+z}{2n}},$$

except on a set of probability not larger than Σe^{-z}. Then, inequality (6) implies

$$\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) \leq \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + E_{\hat{k},\hat{\ell}} + \delta^2 \sqrt{\frac{x_{\hat{k},\hat{\ell}}+z}{2n}} - \text{pen}(\hat{k},\hat{\ell}) + \text{pen}(k,\ell)$$

$$\leq \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + E_{\hat{k},\hat{\ell}} + \delta^2 \sqrt{\frac{x_{\hat{k},\hat{\ell}}}{2n}} - \text{pen}(\hat{k},\hat{\ell}) + \text{pen}(k,\ell) + \delta^2 \sqrt{\frac{z}{2n}},$$

except on a set of probability not larger than Σe^{-z}. Consequently, if for all $k' \geq 1$ and $\ell' \in \mathcal{L}$,

$$\text{pen}(k',\ell') \geq E_{k',\ell'} + \delta^2 \sqrt{\frac{x_{k',\ell'}}{2n}},$$

then

$$\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) \leq \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + \text{pen}(k,\ell) + \delta^2 \sqrt{\frac{z}{2n}},$$

except on a set of probability not larger than Σe^{-z}. Put differently,

$$\mathbb{P}\left\{ \delta^{-2}\sqrt{2n}[\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) - \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + \text{pen}(k,\ell)] \geq \sqrt{z} \right\} \leq \Sigma e^{-z},$$

or, letting $z = u^2$,

$$\mathbb{P}\left\{ [\delta^{-2}\sqrt{2n}[\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) - \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + \text{pen}(k,\ell)] \geq u \right\} \leq \Sigma e^{-u^2}.$$

Recalling that $\int_0^\infty e^{-u^2} du = \frac{\sqrt{\pi}}{2}$ and letting $g_+ = \max(g, 0)$, we obtain

$$\mathbb{E}\left[(\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}}) - \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \bar{\Delta}_n(\mathbf{f}_{k,\ell}) + \text{pen}(k,\ell))_+ \right] \leq \frac{\delta^2 \Sigma}{2^{3/2}} \sqrt{\frac{\pi}{n}}.$$

Hence, as $\mathbb{E}[\bar{\Delta}_n(\mathbf{f}_{k,\ell})] = 0$,

$$\mathbb{E}[\mathcal{D}(\mathbf{f}^\star, \tilde{\mathbf{f}})] \leq \mathcal{D}(\mathbf{f}^\star, \mathbf{f}_{k,\ell}) + \mathrm{pen}(k, \ell) + \frac{\delta^2 \Sigma}{2^{3/2}} \sqrt{\frac{\pi}{n}}.$$

Since this is true for all k and ℓ, we finally get

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

where $\mathcal{D}(\mathbf{f}^\star, \mathcal{F}_{k,\ell}) = \inf_{\mathbf{f} \in \mathcal{F}_{k,\ell}} \mathcal{D}(\mathbf{f}^\star, \mathbf{f})$. This concludes the proof of Theorem 2.1.

5.2 Proof of Proposition 2.1

The first step consists in proving that the quantity

$$\mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k,\ell}} (\mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})] - \Delta_n(\mathbf{f})) \right]$$

may be upper bounded by means of the Rademacher average

$$\mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k,\ell}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \Delta(\mathbf{f}, \mathbf{X}_i) \right],$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent Rademacher random variables, defined by $\mathbb{P}\{\varepsilon_i = 1\} = \mathbb{P}\{\varepsilon_i = -1\} = 1/2$, independent of $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Let $\mathbf{X}'_1, \ldots, \mathbf{X}'_n$ be independent copies of $\mathbf{X}_1, \ldots, \mathbf{X}_n$, also independent of $\varepsilon_1, \ldots, \varepsilon_n$. A symmetrization argument yields

$$\mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k,\ell}} (\mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})] - \Delta_n(\mathbf{f})) \right]$$

$$= \mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k,\ell}} \left(\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \Delta(\mathbf{f}, \mathbf{X}'_i) \Big| \mathbf{X}_1, \ldots, \mathbf{X}_n \right] - \frac{1}{n} \sum_{i=1}^{n} \Delta(\mathbf{f}, \mathbf{X}_i) \right) \right]$$

$$\leq \mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k,\ell}} \frac{1}{n} \sum_{i=1}^{n} \left(\Delta(\mathbf{f}, \mathbf{X}'_i) - \Delta(\mathbf{f}, \mathbf{X}_i) \right) \right]$$

$$= \mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k,\ell}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(\Delta(\mathbf{f}, \mathbf{X}'_i) - \Delta(\mathbf{f}, \mathbf{X}_i) \right) \right]$$

$$\leq \mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k,\ell}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \Delta(\mathbf{f}, \mathbf{X}'_i) \right] + \mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k,\ell}} \frac{1}{n} \sum_{i=1}^{n} (-\varepsilon_i) \Delta(\mathbf{f}, \mathbf{X}_i) \right]$$

$$= 2\mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_{k,\ell}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \Delta(\mathbf{f}, \mathbf{X}_i) \right].$$

Next, the Rademacher average

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta(\mathbf{f},\mathbf{X}_i)\right]$$

may be bounded by resorting to a Dudley integral. More precisely, let

$$S_{k,\ell}=\{\Delta(\mathbf{f},\cdot),\mathbf{f}\in\mathcal{F}_{k,\ell}\}$$

be a subset of the continuous functions from $\mathcal{C}$ to $\mathbb{R}^+$, endowed with the sup-norm $\|\cdot\|_\infty$, and denote by $\mathcal{N}(S_{k,\ell},\|\cdot\|_\infty,\varepsilon)$ the covering number of $S_{k,\ell}$, i.e., the minimal number of closed balls of radius ε needed to cover $S_{k,\ell}$. According to Dudley [18], there exists an absolute constant $c>0$ such that, for all $\mathbf{X}_1,\ldots,\mathbf{X}_n$,

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\ell}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta(\mathbf{f},\mathbf{X}_i)\right]\leq\frac{c}{\sqrt{n}}\int_0^\infty\sqrt{\ln\mathcal{N}(S_{k,\ell},\|\cdot\|_\infty,\varepsilon)}\mathrm{d}\varepsilon.$$

To evaluate the covering number of $S_{k,\ell}$, we may use Lemma 2 in Kégl [27], which ensures that

$$\mathcal{N}(S_{k,\ell},\|\cdot\|_\infty,\varepsilon)\leq 2^{\ell\delta/\varepsilon+3k+1}V_d^{k+1}\left[\frac{\delta^2\sqrt{d}}{\varepsilon}+\sqrt{d}\right]^d\left[\frac{\ell\delta\sqrt{d}}{k\varepsilon}+3\sqrt{d}\right]^{kd},$$

where V_d denotes the volume of the d-dimensional unit ball. Observe that

$$\ln\mathcal{N}(S_{k,\ell},\|\cdot\|_\infty,\varepsilon)$$
$$\leq\left(\frac{\ell\delta}{\varepsilon}+3k+1\right)\ln 2+(k+1)\ln V_d+d\ln\left(\frac{\delta^2\sqrt{d}}{\varepsilon}+\sqrt{d}\right)+kd\ln\left(\frac{\ell\delta\sqrt{d}}{k\varepsilon}+3\sqrt{d}\right)$$
$$=\frac{\ell\delta}{\varepsilon}\ln 2+(3k+1)\ln 2+(k+1)\ln V_d+d(k+1)\ln\sqrt{d}+d\ln\left(\frac{\delta^2}{\varepsilon}+1\right)+kd\ln 3$$
$$+kd\ln\left(\frac{\ell\delta}{3k\varepsilon}+1\right)$$
$$=\frac{\ell\delta}{\varepsilon}\ln 2+d\ln\left(\frac{\delta^2}{\varepsilon}+1\right)+kd\ln\left(\frac{\ell\delta}{3k\varepsilon}+1\right)+kd\ln 3+(3k+1)\ln 2$$
$$+(k+1)(\ln V_d+\frac{d}{2}\ln d).$$

Hence, recalling that the support of $\mathbf{f}$ is included in a set $\mathcal{C}$ with diameter δ, we obtain

$$\int_0^\infty\sqrt{\ln\mathcal{N}(S_{k,\ell},\|\cdot\|_\infty,\varepsilon)}\mathrm{d}\varepsilon=\int_0^{\delta^2}\sqrt{\ln\mathcal{N}(S_{k,\ell},\|\cdot\|_\infty,\varepsilon)}\mathrm{d}\varepsilon$$
$$\leq\sqrt{\ell\delta\ln 2}I_1+\sqrt{d}I_2+\sqrt{kd}I_3+\delta^2 A(k,d),$$

28

where $I_1 = \int_0^{\delta^2} \frac{1}{\sqrt{\varepsilon}} d\varepsilon$, $I_2 = \int_0^{\delta^2} \sqrt{\ln\left(\frac{\delta^2}{\varepsilon} + 1\right)} d\varepsilon$, $I_3 = \int_0^{\delta^2} \sqrt{\ln\left(\frac{\ell\delta}{3k\varepsilon} + 1\right)} d\varepsilon$, and

$$A(k, d) = \left[kd\ln 3 + (3k + 1)\ln 2 + (k + 1)(\ln V_d + \frac{d}{2}\ln d) \right]^{1/2}.$$

Control of I_1. Clearly,

$$I_1 = \int_0^{\delta^2} \frac{1}{\sqrt{\varepsilon}} d\varepsilon = 2\delta.$$

Control of I_2. We have

$$I_2 \leq \int_0^{\delta^2} \sqrt{\ln\left(\frac{2\delta^2}{\varepsilon}\right)} d\varepsilon$$

$$= 2\delta^2 \int_0^{1/2} \sqrt{\ln\frac{1}{u}} du$$

$$\leq \delta^2(\sqrt{\ln 2} + \sqrt{\pi}).$$

Control of I_3. Let $M = \max(3k, L/\delta)$. Clearly, for all $\ell \in \mathcal{L}$, $\delta \geq \frac{\ell}{M}$, and then $\delta^2 \geq \frac{\ell\delta}{M}$. Let us cut up the integral I_3 and write

$$I_3 = \int_0^{\delta^2} \sqrt{\ln\left(\frac{\ell\delta}{3k\varepsilon} + 1\right)} d\varepsilon$$

$$= \int_0^{\ell\delta/M} \sqrt{\ln\left(\frac{\ell\delta}{3k\varepsilon} + 1\right)} d\varepsilon + \int_{\ell\delta/M}^{\delta^2} \sqrt{\ln\left(\frac{\ell\delta}{3k\varepsilon} + 1\right)} d\varepsilon. \qquad (7)$$

Observe, since $\varepsilon \leq \frac{\ell\delta}{M}$, that $\frac{\ell\delta}{3k\varepsilon} \geq 1$. Consequently,

$$\int_0^{\ell\delta/M} \sqrt{\ln\left(\frac{\ell\delta}{3k\varepsilon} + 1\right)} d\varepsilon \leq \int_0^{\ell\delta/M} \sqrt{\ln\left(\frac{2\ell\delta}{3k\varepsilon}\right)} d\varepsilon$$

$$= \frac{2\ell\delta}{3k} \int_0^{3k/2M} \sqrt{\ln\frac{1}{u}} du$$

$$\leq \frac{\ell\delta}{M}\left(\sqrt{\ln\left(\frac{2M}{3k}\right)} + \sqrt{\pi}\right).$$

The second integral in equality (7) may be bounded using the fact that the integrand is a decreasing function of ε:

$$\int_{\ell\delta/M}^{\delta^2} \sqrt{\ln\left(\frac{\ell\delta}{3k\varepsilon} + 1\right)} d\varepsilon \leq \left(\delta^2 - \frac{\ell\delta}{M}\right)\sqrt{\ln\left(\frac{M}{3k} + 1\right)}$$

$$\leq \left(\delta^2 - \frac{\ell\delta}{M}\right)\sqrt{\ln\left(\frac{2M}{3k}\right)}.$$

As a result,

$$I_3 \leq \delta^2 \sqrt{\ln\left(\frac{2M}{3k}\right)} + \frac{\ell\delta}{M}\sqrt{\pi}.$$

Thus,

$$\int_0^\infty \sqrt{\ln\mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon$$

$$\leq 2\delta\sqrt{\delta\ell\ln 2} + \sqrt{d}\delta^2(\sqrt{\ln 2} + \sqrt{\pi}) + \frac{\ell\delta}{M}\sqrt{kd\pi} + \delta^2\sqrt{kd\ln\left(\frac{2M}{3k}\right)} + \delta^2 A(k,d)$$

$$= 2\delta\sqrt{\delta\ell\ln 2} + \frac{\ell\delta}{M}\sqrt{kd\pi} + \sqrt{k}\delta^2\left[d\ln\left(\frac{2M}{3k}\right) + d\ln 3 + \frac{d}{2}\ln d + \ln V_d + 3\ln 2\right]^{1/2} + a_0',$$

where a_0' is a nonnegative constant. Recalling that $M \geq 3k$, we have $1/M \leq 1/(3\sqrt{k})$. Hence, $\ell\delta\sqrt{kd\pi}/M \leq \ell\delta\sqrt{d\pi}/3$. Since $\sqrt{\ell} \leq \max(1,\ell)$, we finally obtain

$$\int_0^\infty \sqrt{\ln\mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon \leq a_1\sqrt{k} + a_2\ell + a_0,$$

where the nonnegative constants $a_0, \ldots, a_2$ only depend on the maximal length L, the dimension d and the diameter δ of the convex set $\mathcal{C}$.

5.3 Proof of Proposition 3.1

Let

$$S_{k,\kappa} = \{\Delta(\mathbf{f}, \cdot), \mathbf{f} \in \mathcal{F}_{k,\kappa}\}$$

be a subset of the continuous functions from $\mathcal{C}$ to $\mathbb{R}^+$, endowed with the sup-norm $\|\cdot\|_\infty$. Starting as in the proof of Proposition 2.1, we know that, for all $\mathbf{X}_1, \ldots, \mathbf{X}_n$,

$$\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_{k,\kappa}}\left(\mathbb{E}[\Delta(\mathbf{f}, \mathbf{X})] - \Delta_n(\mathbf{f})\right)\right] \leq \frac{c}{\sqrt{n}}\int_0^\infty \sqrt{\ln\mathcal{N}(S_{k,\kappa}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon,$$

for some absolute constant $c > 0$. Now, according to Lemma 5 in Sandilya and Kulkarni [39], we may write, for each $\varepsilon > 0$,

$$\ln \mathcal{N}(S_{k,\kappa}, \|\cdot\|_\infty, \varepsilon)$$

$$\leq \left(\frac{\zeta(\kappa)\delta^2}{\varepsilon} + 2k + 1 \right) \ln 2 + (k+1)\ln V_d + d\ln\left(\frac{\delta^2\sqrt{d}}{\varepsilon} + \sqrt{d} \right)$$

$$+ kd\ln\left(\frac{\zeta(\kappa)\delta^2\sqrt{d}}{k\varepsilon} + 3\sqrt{d} \right)$$

$$= \frac{\zeta(k)\delta^2}{\varepsilon}\ln 2 + (2k+1)\ln 2 + (k+1)\ln V_d + d(k+1)\ln\sqrt{d} + d\ln\left(\frac{\delta^2}{\varepsilon} + 1 \right)$$

$$+ kd\ln 3 + kd\ln\left(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1 \right)$$

$$= \frac{\zeta(k)\delta^2}{\varepsilon}\ln 2 + d\ln\left(\frac{\delta^2}{\varepsilon} + 1 \right) + kd\ln\left(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1 \right) + kd\ln 3 + (2k+1)\ln 2$$

$$+ (k+1)(\ln V_d + \frac{d}{2}\ln d).$$

Consequently,

$$\int_0^\infty \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon = \int_0^{\delta^2} \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)}\mathrm{d}\varepsilon$$

$$\leq \delta\sqrt{\zeta(\kappa)\ln 2}I_1 + \sqrt{d}I_2 + \sqrt{kd}I_3 + \delta^2 A(k,d),$$

where $I_1 = \int_0^{\delta^2} \frac{1}{\sqrt{\varepsilon}}\mathrm{d}\varepsilon$, $I_2 = \int_0^{\delta^2} \sqrt{\ln\left(\frac{\delta^2}{\varepsilon} + 1 \right)}\mathrm{d}\varepsilon$, $I_3 = \int_0^{\delta^2} \sqrt{\ln\left(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1 \right)}\mathrm{d}\varepsilon$,
and

$$A(k,d) = \delta^2 \left[kd\ln 3 + (2k+1)\ln 2 + (k+1)(\ln V_d + \frac{d}{2}\ln d) \right]^{1/2}.$$

Control of I_1. We clearly have

$$I_1 = \int_0^{\delta^2} \frac{1}{\sqrt{\varepsilon}}\mathrm{d}\varepsilon = 2\delta.$$

Control of I_2. We have

$$I_2 \leq \int_0^{\delta^2} \sqrt{\ln\left(\frac{2\delta^2}{\varepsilon} \right)}\mathrm{d}\varepsilon$$

$$= 2\delta^2 \int_0^{1/2} \sqrt{\ln\frac{1}{u}}\mathrm{d}u$$

$$\leq \delta^2(\sqrt{\ln 2} + \sqrt{\pi}).$$

Control of I_3. Assume first that $\dfrac{\zeta(\kappa)}{3k} \geq 1$. Then

$$
\begin{aligned}
I_3 &= \int_0^{\delta^2} \sqrt{\ln\left(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\right)}\, d\varepsilon \\
&\leq \int_0^{\delta^2} \sqrt{\ln\left(\frac{2\zeta(\kappa)\delta^2}{3k\varepsilon}\right)}\, d\varepsilon \\
&= \frac{2\zeta(\kappa)\delta^2}{3k} \int_0^{3k/2\zeta(\kappa)} \sqrt{\ln\frac{1}{u}}\, du \\
&\leq \delta^2\left(\sqrt{\ln\frac{2\zeta(\kappa)}{3k}} + \sqrt{\pi}\right).
\end{aligned}
$$

On the other hand, if $\dfrac{\zeta(\kappa)}{3k} < 1$, we cut up I_3 into two pieces and write

$$
\begin{aligned}
I_3 &= \int_0^{\delta^2} \sqrt{\ln\left(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\right)}\, d\varepsilon \\
&= \int_0^{\zeta(\kappa)\delta^2/3k} \sqrt{\ln\left(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\right)}\, d\varepsilon + \int_{\zeta(\kappa)\delta^2/3k}^{\delta^2} \sqrt{\ln\left(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\right)}\, d\varepsilon. \quad (8)
\end{aligned}
$$

The first integral is bounded by using the inequality $\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} \geq 1$ for all $\varepsilon \in \,]0, \frac{\zeta(\kappa)\delta^2}{3k}]$. We obtain

$$
\begin{aligned}
\int_0^{\zeta(\kappa)\delta^2/3k} \sqrt{\ln\left(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\right)}\, d\varepsilon &\leq \int_0^{\zeta(\kappa)\delta^2/3k} \sqrt{\ln\left(\frac{2\zeta(\kappa)\delta^2}{3k\varepsilon}\right)}\, d\varepsilon \\
&= \frac{2\zeta(\kappa)\delta^2}{3k} \int_0^{1/2} \sqrt{\ln\frac{1}{u}}\, du \\
&\leq \frac{\zeta(\kappa)\delta^2}{3k}(\sqrt{\ln 2} + \sqrt{\pi}).
\end{aligned}
$$

With respect to the second integral in (8), we note that the function under the integral is decreasing in ε, so that

$$
\int_{\zeta(\kappa)\delta^2/3k}^{\delta^2} \sqrt{\ln\left(\frac{\zeta(\kappa)\delta^2}{3k\varepsilon} + 1\right)}\, d\varepsilon \leq \left(\delta^2 - \frac{\zeta(\kappa)\delta^2}{3k}\right)\sqrt{\ln 2}.
$$

Thus, we have

$$
I_3 \leq \begin{cases} \delta^2\left(\sqrt{\ln\dfrac{\zeta(\kappa)}{3k}} + \sqrt{\pi} + \sqrt{\ln 2}\right) & \text{if} \quad \dfrac{\zeta(\kappa)}{3k} \geq 1 \\[2ex] \delta^2\left(\dfrac{\zeta(\kappa)}{3k}\sqrt{\pi} + \sqrt{\ln 2}\right) & \text{if} \quad \dfrac{\zeta(\kappa)}{3k} < 1. \end{cases}
$$

Hence, collecting the different results,

$$
\int_0^\infty \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)} \mathrm{d}\varepsilon
$$

$$
\leq 2\delta^2 \sqrt{\zeta(\kappa)\ln 2} + \sqrt{d}\delta^2(\sqrt{\ln 2} + \sqrt{\pi}) + \delta^2\sqrt{kd}\left(\sqrt{\ln \frac{\zeta(\kappa)}{3k}} + \sqrt{\pi} + \sqrt{\ln 2}\right)\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}\geq 1\}}
$$

$$
+ \delta^2\sqrt{kd}\left(\frac{\zeta(\kappa)}{3k}\sqrt{\pi} + \sqrt{\ln 2}\right)\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}<1\}} + \delta^2 A(k,d)
$$

$$
\leq \delta^2\left(2\sqrt{\zeta(\kappa)\ln 2} + \frac{\zeta(\kappa)}{3\sqrt{k}}\sqrt{\pi d}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}<1\}} + \sqrt{kd\ln\frac{\zeta(\kappa)}{3k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}\geq 1\}}\right.
$$

$$
\left. + \sqrt{k}\left[\sqrt{d}(\sqrt{\pi} + \sqrt{\ln 2}) + \left(d\ln 3 + \frac{d}{2}\ln d + \ln V_d + 2\ln 2\right)^{1/2} + a_0\right]\right),
$$

where a_0 is a nonnegative constant. Finally,

$$
\int_0^\infty \sqrt{\ln \mathcal{N}(S_{k,\ell}, \|\cdot\|_\infty, \varepsilon)} \mathrm{d}\varepsilon
$$

$$
\leq \delta^2\left(a_1\sqrt{k} + a_2\sqrt{\zeta(\kappa)} + a_3\frac{\zeta(\kappa)}{\sqrt{k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}<1\}} + a_4\sqrt{k\ln\frac{\zeta(\kappa)}{k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}\geq 1\}} + a_0\right),
$$

where the nonnegative constants $a_0, \ldots, a_4$ only depend on the dimension d.

References

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

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

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

[4] J. D. Banfield and A. E. Raftery. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *Journal of the American Statistical Association*, 87:7–16, 1992.

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

[6] P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2001.

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

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

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

[10] B. S. Caffo, C. M. Crainiceanu, L. Deng, and C. W. Hendrix. A case study in pharmacologic colon imaging using principal curves in single photon emission computed tomography. *Journal of the American Statistical Association*, 103:1470–1480, 2008.

[11] K. Chang and J. Ghosh. Principal curves for nonlinear feature extraction and classification. *SPIE Applications of Artificial Neural Networks in Image Processing III*, 3307:120–129, 1998.

[12] P. J. Corkeron, P. Anthony, and R. Martin. Ranging and diving behaviour of two 'offshore' bottlenose dolphins, Tursiops sp., off eastern Australia. *Journal of the Marine Biological Association of the United Kingdom*, 84:465–468, 2004.

[13] G. De'ath. Principal curves: a new technique for indirect and direct gradient analysis. *Ecology*, 80:2237–2253, 1999.

[14] P. Delicado. Another look at principal curves and surfaces. *Journal of Multivariate Analysis*, 77:84–116, 2001.

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

[16] T. Duchamp and W. Stuetzle. Extremal properties of principal curves in the plane. *The Annals of Statistics*, 24:1511–1520, 1996.

[17] R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1:290–330, 1967.

[18] R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 1999.

[19] J. Einbeck, G. Tutz, and L. Evers. Exploring multivariate data structures with local principal curves. In C. Weihs and W. Gaul, editors, *Classification – The Ubiquitous Challenge, Proceedings of the 28th Annual Conference of the Gesellschaft für Klassifikation, University of Dortmund*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 256–263. Springer, Berlin, Heidelberg, 2005.

[20] J. Einbeck, G. Tutz, and L. Evers. Local principal curves. *Statistics and Computing*, 15:301–313, 2005.

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

[22] H. Friedsam and W. A. Oren. The application of the principal curve analysis technique to smooth beamlines. In *Proceedings of the 1st International Workshop on Accelerator Alignment*, 1989.

[23] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. The geometry of nonparametric filament estimation. 2010. Available at `http://arxiv.org/PS_cache/arxiv/pdf/1003/1003.5536v2.pdf`.

[24] D. J. Harding and G. S. Berghoff. Fault scarp detection beneath dense vegetation cover: airborne lidar mapping of the Seattle fault zone, Bainbridge Island, Washington State. In *Proceedings of the American Society of Photogrammetry and Remote Sensing Annual Conference*, 2000.

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

[26] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 498–520, 1933.

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

[28] B. Kégl and A. Krzyżak. Piecewise linear skeletonization using principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:59–74, 2002.

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

[30] A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis.* Dover Publications, Mineola, 1975.

[31] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47:1902–1914, 2001.

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

[33] M. Lerasle. Optimal model selection in density estimation. *Annales de l'Institut Henri Poincaré.* In press.

[34] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis.* Academic Press, London, 1979.

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

[36] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press, Cambridge, 1989.

[37] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[38] K. Reinhard and M. Niranjan. Parametric subspace modeling of speech transitions. *Speech Communication*, 27:19–42, 1999.

[39] S. Sandilya and S. R. Kulkarni. Principal curves with bounded turn. *IEEE Transactions on Information Theory*, 48:2789–2793, 2002.

[40] A. Saumard. The slope heuristics in heteroscedastic regression. 2010. Available at http://hal.archives-ouvertes.fr/docs/00/51/23/06/PDF/Slope-Heuristics-Regression.pdf.

[41] C. Spearman. General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.

[42] D. C. Stanford and A. E. Raftery. Finding curvilinear features in spatial point patterns: principal curve clustering with noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:2237–2253, 2000.

[43] R. Tibshirani. Principal curves revisited. *Statistics and Computing*, 2:183–190, 1992.

[44] J. J. Verbeek, N. Vlassis, and B. Kröse. A soft *k*-segments algorithm for principal curves. In *Proceedings of International Conference on Artificial Neural Networks 2001*, pages 450–456, 2001.

[45] W. C. K. Wong and A. C. S. Chung. Principal curves to extract vessels in 3D angiograms. In *Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'08)*, pages 1–8, 2008.