Selecting the length of a principal curve within a Gaussian model

Aurélie Fischer

Laboratoire MAP5, Université Paris Descartes
45 rue des Saints-Pères, 75006 Paris, France

aurelie.fischer@parisdescartes.fr

Abstract – Principal curves are parameterized curves passing "through the middle" of a data cloud. These objects constitute a way of generalization of the notion of first principal component in Principal Component Analysis. Several definitions of principal curve have been proposed, one of which can be expressed as a least-square minimization problem. In the present paper, adopting this definition, we study a Gaussian model selection method for choosing the length of the principal curve, in order to avoid interpolation, and obtain a related oracle-type inequality. The proposed method is practically implemented and illustrated on cartography problems.

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

AMS 2010 Mathematics Subject Classification: 62G08, 62G05.

1 Introduction

Principal curves can be thought of as a nonlinear generalization of Principal Component Analysis. Instead of searching for the first principal component of a data cloud, the purpose is to design a curve passing "through the middle" of the observations, as illustrated in Figure 1. Principal curves have many applications in various areas, such as physics (Hastie and Stuetzle [21], Friedsam and Oren [19]), character and speech recognition (Kégl and Krzyżak [22], Reinhard and Niranjan [30]), but also mapping and geology (Brunsdon [10], Stanford and Raftery [34], Banfield and Raftery [3], Einbeck, Tutz and Evers [17, 18]), natural sciences (De'ath [14], Corkeron, Anthony and Martin [13], Einbeck, Tutz and Evers [17]) and medicine (Wong and Chung [38], Caffo, Crainiceanu, Deng and Hendrix [11]).

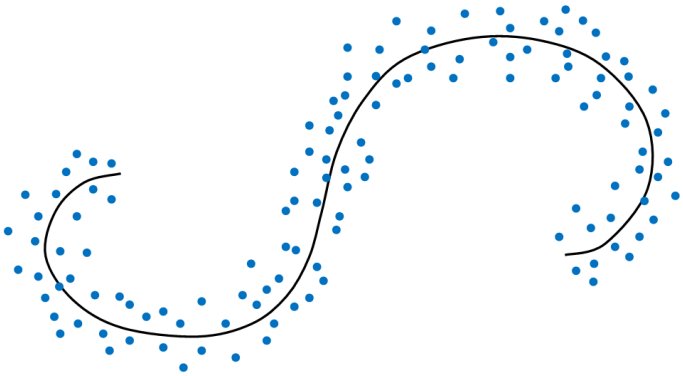


Figure 1: An example of principal curve.

These curves are parameterized curves in $\mathbb{R}^d$, i.e. continuous functions

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

where $I = [a, b]$ is a closed interval of the real line. The original definition of a principal curve, due to Hastie and Stuetzle [21], relies on the self-consistency property of principal components. 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, has finite length inside any bounded subset of $\mathbb{R}^d$, and is self-consistent, which means that

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

Here, the so-called projection index $t_{\mathbf{f}}(\mathbf{x})$ is defined by

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

where $\| \cdot \|$ denotes the standard Euclidean norm of $\mathbb{R}^d$. So, $t_{\mathbf{f}}(\mathbf{x})$ is the largest real number t minimizing the Euclidean distance between $\mathbf{x}$ and $\mathbf{f}(t)$, as shown in Figure 2. 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.
A number of other points of view, more or less related to this original definition, have been proposed thereafter. Tibshirani [35], keeping the self-consistency property, adopts a semiparametric approach and defines principal curves in terms of a mixture model, whereas Delicado [15] generalizes another property of principal components, leading to the notion of "principal curves of oriented points". The definitions of Verbeek, Vlassis and Kröse [36] and Einbeck, Tutz and Evers [18] are based on local principal components put together and the "locally defined principal curves" of Ozertem and Erdogmus [28] correspond to the ridge of a density function. Recently, Genovese, Perone-Pacifico, Verdinelli and Wasserman [20] discussed a closely related problem, called nonparametric filament estimation.

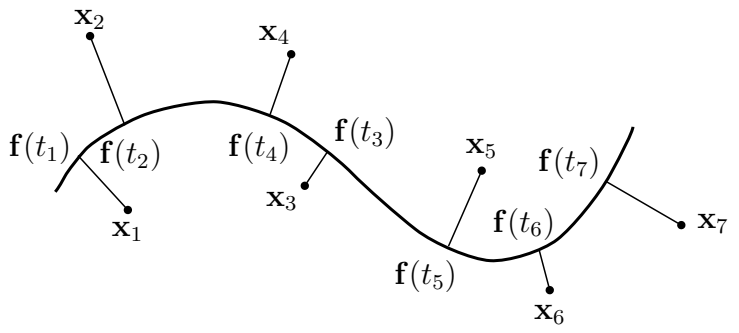


Figure 2: The projection index $t_{\mathbf{f}}$. For all i, t_i stands for $t_{\mathbf{f}}(\mathbf{x}_i)$.

In this paper, we will adopt the principal curve definition of Kégl, Krzyżak, Linder and Zeger [23], which is closely related to the original one, but presents the advantage of avoiding the implicit formulation. Instead, this definition takes the form of an empirical risk minimization problem, which is easier to handle than (1). In the definition of Kégl, Krzyżak, Linder and Zeger [23], a principal curve of length L for $\mathbf{X}$ is a parameterized curve minimizing the least-square type criterion

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

over all curves of length not larger than $L > 0$. Such a principal curve always exists provided $\mathbb{E}\|\mathbf{X}\|^2 < \infty$, but it may not necessarily be unique. Note that Sandilya and Kulkarni [31] have proposed a similar definition, using a constraint on the turn instead of the length of the curve.

In practice, the distribution of $\mathbf{X}$ is unknown, and we have at hand a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of independent random variables distributed as $\mathbf{X}$. In this situation, $\Delta(\mathbf{f})$ is replaced by its empirical counterpart

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

In order to construct a satisfactory principal curve, a good choice of the length is crucial. Indeed, a principal curve constrained to have a too small length will not be able to capture the shape of the data, whereas a too long curve may lead to interpolation problems, as illustrated in Figure 3. In the present contribution, we propose to study the length selection problem, using the approach of non-asymptotic model selection by penalization introduced by Birgé and Massart [7] and Barron, Birgé and Massart [4]. To this end, we will consider a Gaussian framework. A related point of view in the context of almost surely bounded random variables is discussed in Biau and Fischer [6].

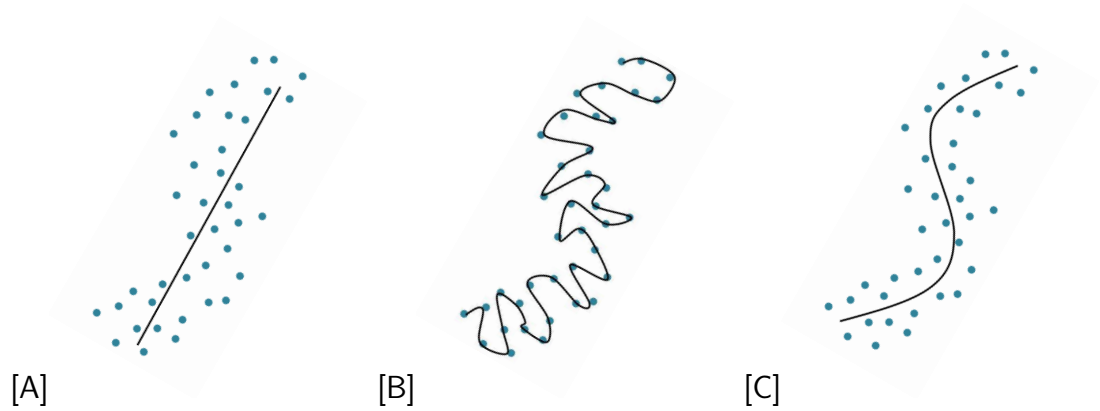


Figure 3: Principal curves fitted with [A] a too small length, [B] a too large length and [C] an appropriate one.

The rest of the paper is organized as follows. In Section 2, we consider the problem of choosing the length of a principal curve using a Gaussian model selection approach, and show that the curve obtained by minimizing some appropriate penalized criterion satisfies an oracle-type inequality. Section 3 presents some experimental results in the context of cartography. Proofs are collected in Section 4 for the sake of clarity.

2 Length selection

We investigate a Gaussian model selection method in order to choose the length of a principal curve. Our context is similar to that of Caillerie and Michel [12], who tackle model selection questions for graphs called simplicial complexes. In the sequel, the Euclidean space $\mathbb{R}^d$ is equipped with the inner product defined by

$$\langle \mathbf{u}, \mathbf{v} \rangle = \frac{1}{d} \sum_{j=1}^{d} u_j v_j, \tag{2}$$

and $\| \cdot \|$ denotes the associated Euclidean norm.

We assume that we observe random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with values in $\mathbb{R}^d$ following the model

$$\mathbf{X}_i = \mathbf{x}_i^\star + \sigma \boldsymbol{\xi}_i, \quad i = 1, \ldots, n, \tag{3}$$

where the $\mathbf{x}_i^\star$ are unknown, the $\boldsymbol{\xi}_i$ are independent standard Gaussian vectors of $\mathbb{R}^d$ and $\sigma > 0$ stands for the noise level, which is supposed known. Let us denote by $\overrightarrow{\mathbf{X}} = {}^t({}^t\mathbf{X}_1, \ldots, {}^t\mathbf{X}_n)$ the (column) vector made of all coordinates of the random vectors $\mathbf{X}_i$, $i = 1, \ldots, n$. Defining $\overrightarrow{\mathbf{x}^\star}$ and $\overrightarrow{\boldsymbol{\xi}}$ in the same way, the model (3) can be rewritten under the form

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

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

curves $\mathbf{f} : I \to \mathbb{R}^d$ with length ℓ and endpoints F and G. Our aim is to select the length ℓ. To do this, we consider the criterion Δ'_n defined by

$$\begin{aligned}\Delta'_n(\mathbf{f}) &= \frac{1}{n}\sum_{i=1}^{n} \inf_{t\in I} \|\mathbf{X}_i - \mathbf{f}(t)\|^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \inf_{\mathbf{x}_i\in\Gamma_\mathbf{f}} \|\mathbf{X}_i - \mathbf{x}_i\|^2,\end{aligned}$$

where $\Gamma_\mathbf{f}$ denotes the range of the curve $\mathbf{f}$. Due to the definition of the norm $\|\cdot\|$ chosen above (2), this is the empirical criterion $\Delta_n(\mathbf{f})$ normalized by the dimension d. Suppose that, for all $\ell \in \mathcal{L}$, $\overrightarrow{\mathbf{x}}_\ell^{(n)} := (\hat{\mathbf{x}}_{1\ell}, \ldots, \hat{\mathbf{x}}_{n\ell})$ minimizes

$$\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - \mathbf{x}_i\|^2$$

among all $\overrightarrow{\mathbf{x}} \in \mathcal{C}_\ell := \bigcup_{\mathbf{f}\in\mathcal{F}_\ell} (\Gamma_\mathbf{f})^n$. In order to determine the length ℓ, our purpose is to minimize in ℓ a criterion of the type

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

where $\mathrm{pen} : \mathcal{L} \to \mathbb{R}^+$ is a penalty function, which should avoid the selection of a too large ℓ. Our goal is to design an appropriate penalty. Observe that the classical asymptotic model selection criteria AIC (Akaike [1]), BIC (Schwarz [33]) or Mallows'C_p (Mallows [26]), which involve the "number of parameters" to be estimated, are not suitable in this framework. Therefore, our approach will rely on the non-asymptotic model selection theory developed by Birgé and Massart [8] and Barron, Birgé and Massart [4].

When the considered models are linear subspaces, the penalty can be chosen proportional to the dimension of the model, according to Birgé and Massart [8]. Here, the models $\mathcal{C}_\ell$ are not linear subspaces of $\mathbb{R}^{nd}$ and the dimension must be replaced by another quantity. In order to measure the complexity of these nonlinear models, we will use metric entropy. The metric entropy of a set S is given by

$$\mathcal{H}(S, \|\cdot\|, \varepsilon) = \ln \mathcal{N}(S, \|\cdot\|, \varepsilon),$$

were the covering number $\mathcal{N}(S, \|\cdot\|, \varepsilon)$ is the minimal number of balls with radius ε for the norm $\|\cdot\|$ needed to cover S.

Our approach is based on a general model selection theorem for nonlinear Gaussian models (Massart [27]). Let us denote by $\|\cdot\|_{nd}$ the normalized norm of $\mathbb{R}^{nd}$, defined by the inner product $\langle \overrightarrow{\mathbf{u}}, \overrightarrow{\mathbf{v}} \rangle_{nd} = \frac{1}{nd}\sum_{i=1}^{nd} u_i v_i$. For every $\ell \in \mathcal{L}$, let φ_ℓ be a function such that $\varphi_\ell \geq \phi_\ell$, where ϕ_ℓ is given by

$$\phi_\ell(u) = \kappa \int_0^u \sqrt{\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)}\mathrm{d}\varepsilon, \tag{4}$$

with κ an absolute constant. We define d_ℓ by the equation

$$\varphi_\ell\left(2\sigma\frac{\sqrt{d_\ell}}{\sqrt{nd}}\right)=\frac{\sigma d_\ell}{\sqrt{nd}}.$$

Assume that there exists a family of weights $\{w_\ell\}_{\ell\in\mathcal{L}}$ satisfying

$$\sum_{\ell\in\mathcal{L}}e^{-w_\ell}=\Sigma<\infty.$$

Under these assumptions and with this notation, Theorem 4.18 in Massart [27] can be written in the following manner:

Theorem 2.1. *Let $\eta>1$ and*

$$\mathrm{pen}(\ell)\geq\eta\frac{\sigma^2}{nd}\left(\sqrt{d_\ell}+\sqrt{2w_\ell}\right)^2.$$

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

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

Moreover, writing $\tilde{\mathbf{x}}_i:=\hat{\mathbf{x}}_{i\hat{\ell}}$ for all i=1,...,n, we have

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

where $d^2(\overrightarrow{\mathbf{x}^\star},\mathcal{C}_\ell)=\inf_{\overrightarrow{\mathbf{y}}\in\mathcal{C}_\ell}\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{y}_i-\mathbf{x}_i^\star\|^2$.

This result establishes, for a penalty $\mathrm{pen}(\ell)$ which is large enough, an oracle-type inequality in expectation for the $\tilde{\mathbf{x}}_i$, $i=1,\ldots,n$. Provided a control of the Dudley integral (4) (Dudley [16]), this theorem will apply in our context and allow us to select the length ℓ of the curve. To assess this integral, we will need some technical lemmas, which are proved in Section 4.

The first step consists in controlling the metric entropy of the classes $\mathcal{C}_\ell$, $\ell\in\mathcal{L}$. Note that, for all $\ell\in\mathcal{L}$, $\bigcup_{\mathbf{f}\in\mathcal{F}_\ell}\Gamma_{\mathbf{f}}$ corresponds to an ellipsoid of $\mathbb{R}^d$, as stated in the next lemma. In the sequel, this ellipsoid will be denoted by $\mathcal{E}_\ell$.

Lemma 2.1. *Every parameterized curve of $\mathbb{R}^d$ with endpoints F and G and length ℓ ($\ell>FG$), is included in an ellipsoid $\mathcal{E}_\ell$ with first principal axis of length ℓ, the other axes having length $\lambda=\sqrt{\ell^2-FG^2}$.*

In particular, in $\mathbb{R}^2$, $\mathcal{E}_\ell$ is an ellipse with foci F and G (see Figure 4), and in $\mathbb{R}^3$, it is a ellipsoid of revolution around the axis passing through these two points.

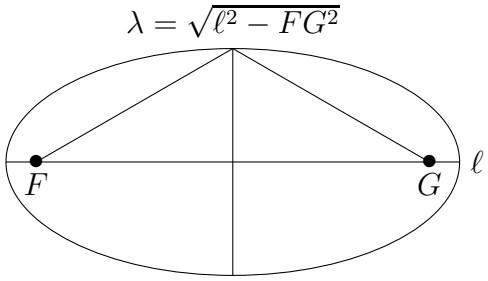


Figure 4: In the plane $\mathbb{R}^2$, ellipse $\mathcal{E}_\ell$ with foci F and G and axes ℓ and λ.

We obtain then the following upper bound for $\mathcal{N}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)$, $\ell \in \mathcal{L}$.

Lemma 2.2. *Suppose that $\ell \geq \lambda \geq \varepsilon$. The covering number of $\mathcal{C}_\ell$ for the normalized norm $\|\cdot\|_{nd}$ of $\mathbb{R}^{nd}$ satisfies*

$$\mathcal{N}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^{nd} (\ell\lambda^{d-1})^n.$$

Bounding the integral

$$\phi_\ell(u) = \kappa \int_0^u \sqrt{\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)}\mathrm{d}\varepsilon$$

for all $\ell \in \mathcal{L}$, we can then define an adequate function φ_ℓ .

Lemma 2.3. *The function φ_ℓ given by*

$$\varphi_\ell(r) = \begin{cases} \kappa r\sqrt{nd}\left(\sqrt{\ln\left(\dfrac{2\ell^{1/d}\lambda^{1-1/d}}{r}\right)} + \sqrt{\pi}\right) & \text{if} \quad r \leq \lambda \\ \varphi_\ell(\lambda) + (r-\lambda)\varphi_\ell'(\lambda) & \text{if} \quad r \geq \lambda \end{cases}$$

satisfies, for all r,

$$\varphi_\ell(r) \geq \phi_\ell(r).$$

Finally, in order to apply Theorem 2.1, we have to assess d_ℓ, defined by the equation

$$\varphi_\ell\left(\frac{2\sigma\sqrt{d_\ell}}{\sqrt{nd}}\right) = \frac{\sigma d_\ell}{\sqrt{nd}},$$

which is the purpose of the next lemma.

Lemma 2.4. *Let φ_ℓ be given by Lemma 2.3. Suppose that*

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

Then, equation

$$\varphi_\ell\left(\frac{2\sigma\sqrt{d_\ell}}{\sqrt{nd}}\right) = \frac{\sigma d_\ell}{\sqrt{nd}}$$

admits a solution d_ℓ satisfying

$$d_\ell \leq 8\kappa^2 nd \left(\ln \left(\frac{\ell^{1/d} \lambda^{1-1/d}}{2\sigma\kappa\sqrt{\pi}} \right) + \pi \right).$$

We are now in a position to state the main result of this section.

Theorem 2.2. *Assume that there exists a family of weights $\{w_\ell\}_{\ell \in \mathcal{L}}$ such that*

$$\sum_{\ell \in \mathcal{L}} e^{-w_\ell} = \Sigma < \infty,$$

and that, for every $\ell \in \mathcal{L}$,

$$\sigma \leq \frac{\lambda}{4\kappa} \left[\sqrt{\ln 2 + \frac{1}{d} \ln \left(\frac{\ell}{\lambda} \right)} + \sqrt{\pi} \right]^{-1}. \tag{5}$$

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

$$\mathrm{pen}(\ell) \geq \eta\sigma^2 \left[c_1 \left(\ln \left(\frac{\ell^{1/d} \lambda^{1-1/d}}{\sigma} \right) + c_2 \right) + \frac{4w_\ell}{nd} \right], \tag{6}$$

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

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

Moreover, if $\tilde{\mathbf{x}}_i = \hat{\mathbf{x}}_{i\hat{\ell}}$ for all $i = 1, \ldots, n$, we have

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

where $d^2(\overrightarrow{\mathbf{x}^\star}, \mathcal{C}_\ell) = \inf_{\overrightarrow{\mathbf{y}} \in \mathcal{C}_\ell} \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{x}_i^\star\|^2$.

Let us now comment on the theorem.

The first remark is about the fact that Theorem 2.2 involves unknown constants. Lemma 2.4 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. In fact, the expression (6) does not provide a penalty function directly, but gives its shape instead. 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 [9] (see also Arlot and Massart [2], Lerasle [25] and Saumard [32]) 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 $\{w_\ell\}_{\ell\in\mathcal{L}}$. For linear models $\mathcal{C}_\ell$ with dimension D_ℓ, a possible choice for w_ℓ is $w_\ell = w(D_\ell)$ where $w(D) = cD + \ln|\{k \in \mathcal{L}, D_k = D\}|$ and $c > 0$ (see Massart [27]). 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 $\sum_{\ell\in\mathcal{L}} \frac{1}{\ell^{c/d}\lambda^{c(1-1/d)}} = \Sigma < +\infty$. 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.

In addition, observe that condition (5) says that the noise level σ should not be too large with respect to λ. In other words, if $\lambda = \sqrt{\ell^2 - FG^2}$ is of the same order as σ, it is not possible to obtain a suitable principal curve with length ℓ.

Finally, let us point out that due to the exponent n in the covering number in Lemma 2.2—a comment is given in Section 4 after the proof of the lemma (Remark 4.1)—, 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. Indeed, its resolution is not made easier by increasing the size of the sample, since nothing has been specified about the repartition of the $\mathbf{x}_i^\star$'s. A possible direction for future research could consist in dealing with the framework in which these points are distributed along the curve following a uniform distribution.

3 Experimental results

In this section, we propose to illustrate the length selection method practically. The experiments presented here are carried out with the software MATLAB. As announced in Section 2, the penalty given in Theorem 2.2 will be calibrated thanks to the slope heuristics. Two strategies may be used : the dimension jump method consists in identifying an abrupt jump in the models complexity, whereas the other solution is to observe that the empirical contrast is proportional to the penalty shape for complex models and use the slope of this line to assess the constant.

In this practical implementation, we considered polygonal lines, which present the advantage that projecting on the curve reduces to projection on a segment. However, the method described below could probably be replaced by a more sophisticated technique dealing with smooth curves. In the sequel, the maximal number k of segments is taken large enough, to ensure that the only parameter reflecting the complexity of the curve is the length. Then, the length ℓ of the principal curve is chosen as follows:

1. For a range of values of the length ℓ, compute $\hat{\mathbf{f}}_\ell$ by minimizing the empirical criterion $\Delta_n(\mathbf{f})$ and record

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

2. Let w_ℓ be proportional to $\ln \ell^{1/d}\lambda^{1-1/d}$ and consider a penalty of the form

$$\mathrm{pen}(k,\ell) = c\ln(\ell^{1/d}\lambda^{1-1/d}).$$

3. Select the constant $\hat{c}$ using the slope heuristics.

4. Retain the curve $\hat{\mathbf{f}}_{\hat{\ell}}$ obtained by minimizing the penalized criterion

$$\mathrm{crit}(\ell) = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}_i - \hat{\mathbf{x}}_{i\ell}\|^2 - 2\hat{c}\ln(\ell^{1/d}\lambda^{1-1/d}).$$

In step 1 of the algorithm, the criterion $\Delta_n(\mathbf{f})$ is minimized by using a MATLAB optimization routine.

The weights w_ℓ were chosen as suggested in the discussion after Theorem 2.2. This is a convenient choice, which does not modify the penalty shape.

To apply the slope heuristics in step 3, we employ the MATLAB package CAPUSHE, implemented by Baudry, Maugis and Michel [5]. We tried both the dimension jump and the slope estimation methods, which results were found to be very similar.

Finally, recall that the endpoints F and G of the principal curve have been assumed to be fixed. From a practical point of view, several methods can be employed to choose these two points from the observations. A possible solution is to define F and G with the aid of the points that are farthest from each other in the minimum spanning tree of the data (or of some subset of the data). Figure 5 gives some examples of such trees, which can be constructed using Kruskal's algorithm [24] or Prim's one [29].

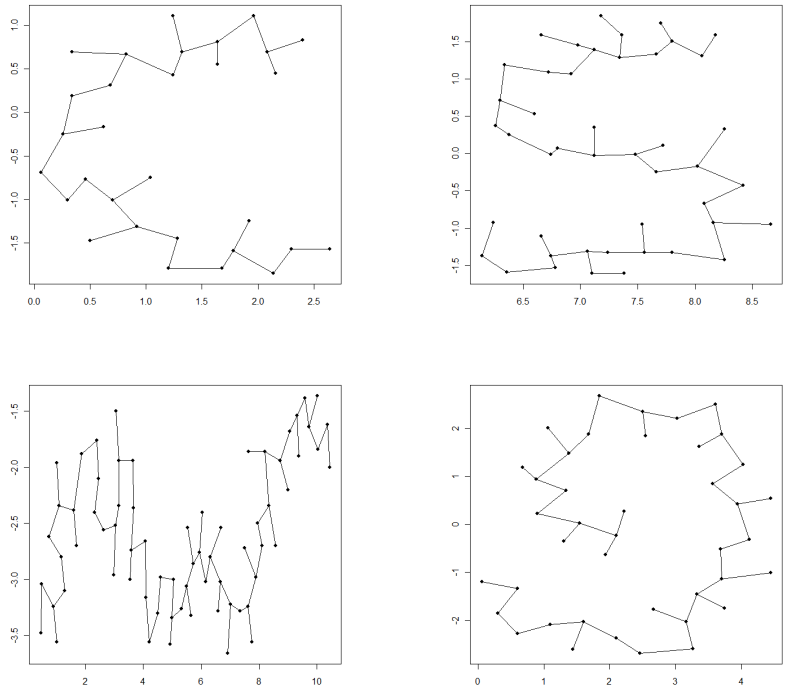


Figure 5: Some examples of minimum spanning trees.

We will now present two examples of applications to mapping. Indeed, Brunsdon [10] has shown that principal curves may be useful in that area, in order to estimate paths from GPS tracks. More specifically, principal curves can serve as a means to compute an average path from GPS data registered by several people moving on a given street.

The results obtained for some hairpin street located in France and the Labyrinth of the Jardin des Plantes in Paris are visible in Figure 6 and 7 respectively. Each figure gives first an air photography of the place and the corresponding GPS tracks 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. Moreover, the principal curves fitted using our model selection approach (denoted by LS in the sequel) can be compared for both data sets to those obtained with a benchmark algorithm. Indeed, Figure 8 gives the outputs of the Polygonal Line Algorithm, which is based on a local control of the curvature and was proposed by Kégl, Krzyżak, Linder and Zeger [23] (PL hereafter).

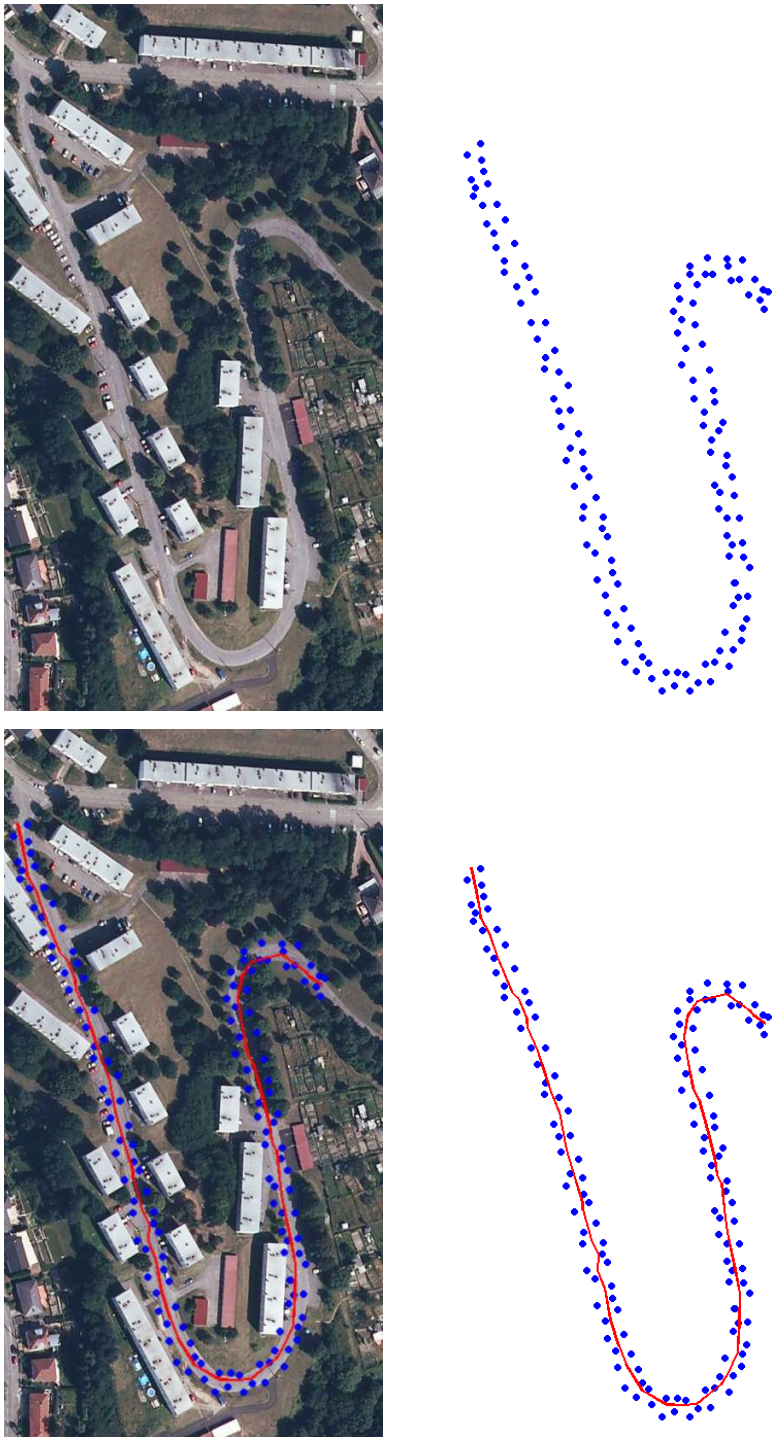


Figure 6: Principal curve fitted with the LS algorithm for the hairspin data.

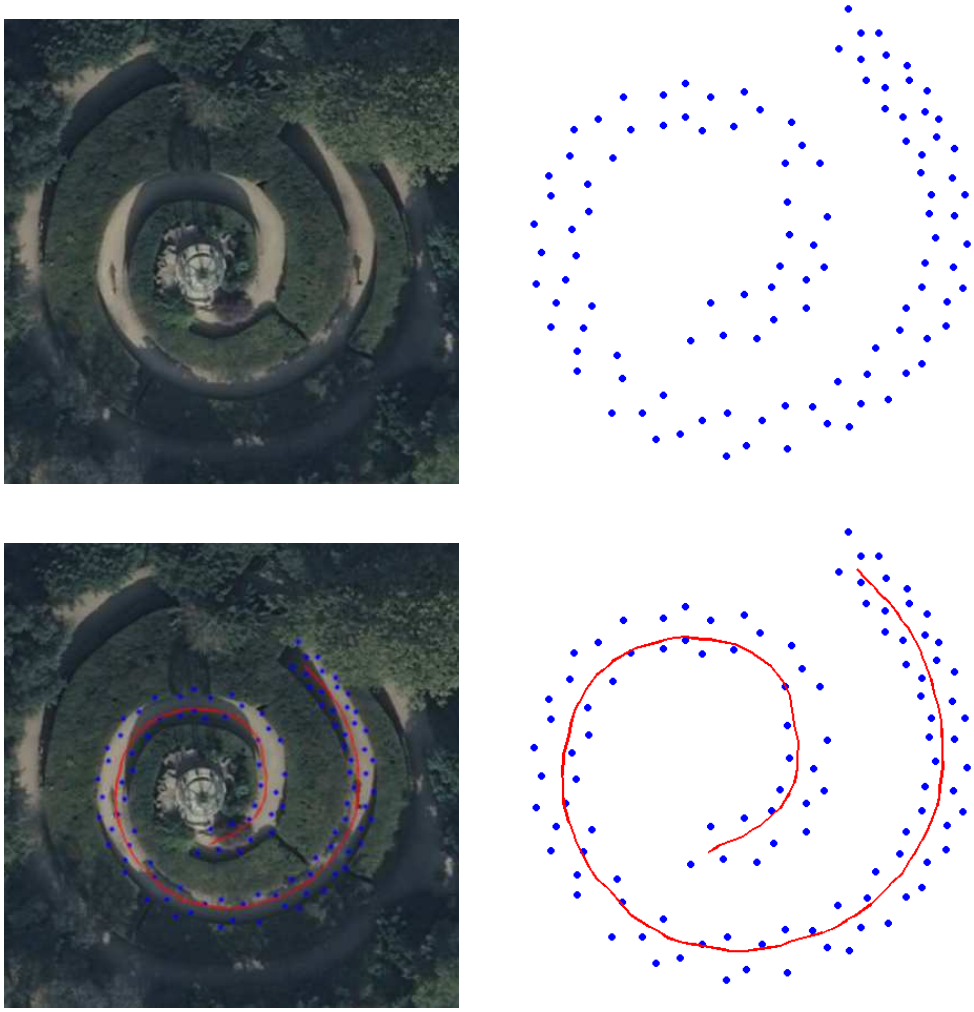
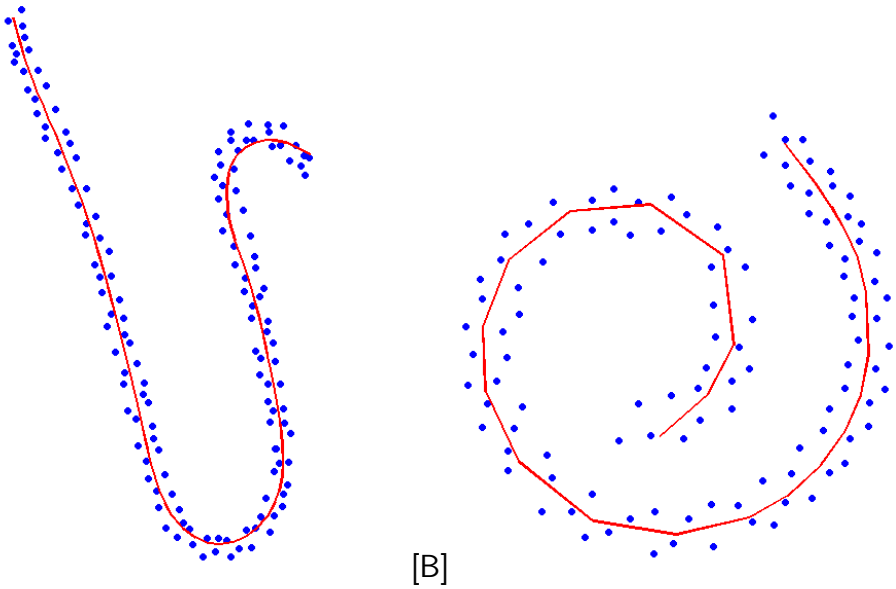


Figure 7: Principal curve fitted with the LS algorithm for the Labyrinth data.



[A] [B]

Figure 8: Principal curves fitted with the PL Algorithm. [A] Hairpin street data. [B] Labyrinth data.

On the whole, the considered streets are correctly recovered. For the hairpin road, we observe that the PL output is smoother, but the right-hand end of the curve looks better on the LS result. Note that the direction taken by this part of the PL curve is wrong. Regarding the Labyrinth, LS perform quite well and yields this time the smoothest curve. The PL principal curve is not at all as smooth as before: this can be explained by the fact that we used for both experiments the default parameters of this algorithm, which may be more or less suitable depending on the different characteristics of the data set, such as curvature or sampling density.

Noting that for the first example, there is a part of the street somewhat hidden by trees, a particularly interesting application could be to use principal curves to draw a map of paths in a forest.

4 Proofs

4.1 Proof of Lemma 2.1

Let $c = FG/2$. Note that $\ell > 2c$. In a well-chosen orthonormal coordinate system of $\mathbb{R}^d$, F has coordinates $(-c, 0, \ldots, 0)$ and G $(c, 0, \ldots, 0)$. A curve with length ℓ and endpoints F and G is included in the set delimited by the points $M(x_1, \ldots, x_d)$ such that

$$MF + MG = \ell.$$

Let us show that this equation defines an ellipsoid with first principal axis ℓ, the other axes having length λ. Let $M(x_1, \ldots, x_d)$ be such that $MF + MG = \ell$. Then,

$$MF^2 = (x_1 + c)^2 + \sum_{j=2}^{d} x_j^2$$

and

$$MG^2 = (x_1 - c)^2 + \sum_{j=2}^{d} x_j^2.$$

Therefore,

$$MF - MG = \frac{MF^2 - MG^2}{MF + MG} = \frac{(x_1 + c)^2 - (x_1 - c)^2}{\ell} = \frac{4x_1 c}{\ell}.$$

As a result, on the one hand,

$$(MF + MG)^2 + (MF - MG)^2 = \ell^2 + \frac{16x_1^2 c^2}{\ell^2},$$

and on the other hand,

$$(MF + MG)^2 + (MF - MG)^2 = 2(MF^2 + MG^2) = 4\sum_{j=1}^{d} x_j^2 + 4c^2.$$

Hence,

$$\ell^2 + \frac{16x_1^2 c^2}{\ell^2} = 4\sum_{j=1}^{d} x_j^2 + 4c^2,$$

which may be rewritten

$$x_1^2 \left(1 - \frac{4c^2}{\ell^2}\right) + \sum_{j=2}^{d} x_j^2 = \frac{\ell^2}{4} - c^2, \tag{7}$$

or, equivalently,

$$\frac{x_1^2}{\ell^2/4} + \sum_{j=2}^{d} \frac{x_j^2}{\ell^2/4 - c^2} = 1, \tag{8}$$

where $\ell^2/4 - c^2 > 0$ since $\ell > 2c$. In other words, the point M belongs to an ellipsoid with one axis of length ℓ and $d-1$ axes of length $2\sqrt{\ell^2/4 - c^2} = \sqrt{\ell^2 - FG^2} = \lambda$.

Reciprocally, if $M(x_1, \ldots, x_d)$ satisfies equation (7), with $\dfrac{\ell^2}{4} - c^2 > 0$, then

$$\begin{aligned} MF^2 &= (x_1 + c)^2 + \sum_{j=2}^{d} x_j^2 \\ &= (x_1 + c)^2 + \frac{\ell^2}{4} - c^2 - x_1^2 + \frac{4x_1^2 c^2}{\ell^2} \\ &= 2x_1 c + \frac{\ell^2}{4} + \frac{4x_1^2 c^2}{\ell^2} \\ &= \left(\frac{\ell}{2} + \frac{2x_1 c}{\ell}\right)^2. \end{aligned}$$

Hence, $MF = \left|\frac{\ell}{2} + \frac{2x_1 c}{\ell}\right|$. Similarly, $MG = \left|\frac{\ell}{2} - \frac{2x_1 c}{\ell}\right|$. Now, $|x_1| \leq \frac{\ell^2}{4c}$, since otherwise $\frac{x_1^2}{\ell^2/4} > \frac{\ell^2}{4c^2} > 1$, contradicting equation (8). Finally, $MF + MG = \frac{\ell}{2} + \frac{2x_1 c}{\ell} + \frac{\ell}{2} - \frac{2x_1 c}{\ell} = \ell$.

4.2 Proof of Lemma 2.2

First, we shall compute the covering number of an ellipsoid $\mathcal{E}_\ell$ of $\mathbb{R}^d$. The next lemma is a particular case of Proposition 5 in von Luxburg, Bousquet and Schölkopf [37].

Lemma 4.1. *Assume that $\ell \geq \lambda \geq \varepsilon$. The number of balls of radius ε needed to cover $\mathcal{E}_\ell$, ellipsoid in dimension d with principal axes $\ell, \lambda, \ldots, \lambda$, satisfies*

$$\mathcal{N}(\mathcal{E}_\ell, \|\cdot\|, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^d \ell \lambda^{d-1}.$$

Proof of Lemma 4.1. The number of balls of radius ε needed to cover $\mathcal{E}_\ell$ satisfies

$$\mathcal{N}(\mathcal{E}_\ell, \|\cdot\|, \varepsilon) \leq \left(\left\lfloor \frac{\ell}{\varepsilon} \right\rfloor + 1\right) \left(\left\lfloor \frac{\lambda}{\varepsilon} \right\rfloor + 1\right)^{d-1},$$

where $\lfloor y \rfloor$ denotes the floor of y, i.e., the largest integer less than or equal to y. Indeed, the ellipsoid $\mathcal{E}_\ell$ is inscribed in a parallelepiped with sides of lengths $\ell, \lambda, \ldots, \lambda$, and the number of balls of radius ε needed to cover a parallelepiped with sides of length $c_1, \ldots, c_d$ is

$$\prod_{j=1}^{d} \left(\left\lfloor \frac{c_j}{\varepsilon} \right\rfloor + 1\right).$$

By assumption, $\ell \geq \lambda \geq \varepsilon$, so that $\left\lfloor \frac{\ell}{\varepsilon} \right\rfloor + 1 \leq \frac{2\ell}{\varepsilon}$ and $\left\lfloor \frac{\lambda}{\varepsilon} \right\rfloor + 1 \leq \frac{2\lambda}{\varepsilon}$. Hence,

$$\mathcal{N}(\mathcal{E}_\ell, \|\cdot\|, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^d \ell \lambda^{d-1}.$$

□

Thus, according to Lemma 4.1,

$$\mathcal{N}(\mathcal{E}_\ell, \|\cdot\|, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^d \ell \lambda^{d-1}.$$

Let $\mathcal{U}$ be a collection of at most $\left(\frac{2}{\varepsilon}\right)^d \ell \lambda^{d-1}$ centers of balls in $(\mathbb{R}^d, \|\cdot\|)$, corresponding to an ε-covering of $\mathcal{E}_\ell$. For each vector $\overrightarrow{\mathbf{u}} = {}^t({}^t\mathbf{u}_1, \ldots, {}^t\mathbf{u}_n) \in \mathbb{R}^{nd}$, where the $\mathbf{u}_i$'s are elements of $\mathcal{U}$, we have $\prod_{i=1}^n B(\mathbf{u}_i, \varepsilon) \subset B(\overrightarrow{\mathbf{u}}, \varepsilon)$ (balls for the normalized norm of $\mathbb{R}^d$ and $\mathbb{R}^{nd}$ respectively). Consequently,

$$\mathcal{N}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^{nd} (\ell \lambda^{d-1})^n.$$

Remark 4.1. It is not possible to get rid of the exponent n in this covering number computation. Indeed, if we consider only one curve $\mathbf{f}$ of length ℓ, covered by N balls of $\mathbb{R}^d$, N^n balls of $\mathbb{R}^{nd}$ are needed to cover $(\Gamma_{\mathbf{f}})^n$, since every point among the n considered points along the curve can be in any one of the N balls. However, our upper bound for the number of balls in $\mathbb{R}^{nd}$ needed to recover $\mathcal{C}_\ell$ is probably too large. Indeed, since the n points are constrained to be located on the same curve of length ℓ, we would not need to use all balls of $\mathbb{R}^{nd}$ obtained by combining the centers of the balls of $\mathbb{R}^d$ covering $\mathcal{E}_\ell$. We could undoubtedly get a better upper bound by solving the combinatorial problem consisting in counting all acceptable combinations of balls in $\mathbb{R}^d$ for a class $\mathcal{C}_\ell$.

4.3 Proof of Lemma 2.3

We begin by a technical lemma which will be useful for bounding integrals.

Lemma 4.2. *For $x \in]0,1]$,*

$$\int_0^x \sqrt{\ln \frac{1}{t}}\mathrm{d}t \leq x\left(\sqrt{\ln \frac{1}{x}} + \sqrt{\pi}\right).$$

Proof of Lemma 4.2. We have

$$\begin{aligned}\int_0^x \sqrt{\ln \frac{1}{t}}\mathrm{d}t &= \left[t\sqrt{\ln 1/t}\right]_0^x + \int_0^x \frac{1}{2\sqrt{\ln 1/t}}\mathrm{d}t \\ &= x\sqrt{\ln 1/x} + \frac{1}{\sqrt{2}}\int_{\sqrt{2\ln 1/x}}^{+\infty} e^{-u^2/2}\mathrm{d}u \\ &\leq x(\sqrt{\ln 1/x} + \sqrt{\pi}).\end{aligned}$$

This inequality is due to the fact that, for $a \geq 0$,

$$\frac{1}{\sqrt{2\pi}}\int_a^{+\infty} e^{-u^2/2}\mathrm{d}u \leq e^{-a^2/2}. \tag{9}$$

Indeed, if g denotes the function defined by $g(a) = e^{-a^2/2} - \frac{1}{\sqrt{2\pi}}\int_a^{+\infty} e^{-u^2/2}\mathrm{d}u$, then $g'(a) = e^{-a^2/2}(\frac{1}{\sqrt{2\pi}} - a)$. This function g is increasing on $[0, 1/\sqrt{2\pi}]$ and decreasing on $[1/\sqrt{2\pi}, +\infty]$. Since $g(0) = 1/2$ and $\lim_{+\infty} g = 0$, we obtain $g(a) \geq 0$ for all $a \geq 0$. Hence, inequality (9) is proved. □

Back to the proof of Lemma 2.3, note that according to Lemma 2.2, the metric entropy $\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)$ satisfies

$$\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon) \leq nd\ln\left(\frac{2\ell^{1/d}\lambda^{1-1/d}}{\varepsilon}\right).$$

If $r \leq \lambda$,

$$\begin{aligned}\phi_\ell(r) &= \kappa \int_0^r \sqrt{\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)}\mathrm{d}\varepsilon \\ &\leq \kappa\sqrt{nd}\int_0^r \sqrt{\ln\left(\frac{2\ell^{1/d}\lambda^{1-1/d}}{\varepsilon}\right)}\mathrm{d}\varepsilon.\end{aligned}$$

The change of variables $t = \dfrac{\varepsilon}{2\ell^{1/d}\lambda^{1-1/d}}$ gives

$$\phi_\ell(r) \leq 2\kappa\sqrt{nd}\ell^{1/d}\lambda^{1-1/d}\int_0^{\frac{r}{2\ell^{1/d}\lambda^{1-1/d}}} \sqrt{\ln\frac{1}{t}}\mathrm{d}t.$$

Now, Lemma 4.2 indicates that for $x \in]0,1]$,

$$\int_0^x \sqrt{\ln\frac{1}{t}}\mathrm{d}t \leq x\left(\sqrt{\ln\frac{1}{x}} + \sqrt{\pi}\right).$$

Thus,

$$\phi_\ell(r) \leq \kappa r\sqrt{nd}\left(\sqrt{\ln\left(\frac{2\ell^{1/d}\lambda^{1-1/d}}{r}\right)} + \sqrt{\pi}\right).$$

For $r \leq \lambda$, let

$$\varphi_\ell(r) = \kappa r\sqrt{nd}\left(\sqrt{\ln\left(\frac{2\ell^{1/d}\lambda^{1-1/d}}{r}\right)} + \sqrt{\pi}\right).$$

If $r \geq \lambda$,

$$\begin{aligned}\phi_\ell(r) &= \kappa\int_0^r \sqrt{\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)}\mathrm{d}\varepsilon \\ &= \kappa\int_0^\lambda \sqrt{\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)}\mathrm{d}\varepsilon + \kappa\int_\lambda^r \sqrt{\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \varepsilon)}\mathrm{d}\varepsilon \\ &\leq \phi_\ell(\lambda) + (r-\lambda)H(\lambda) \\ &\leq \varphi_\ell(\lambda) + (r-\lambda)\varphi'_\ell(\lambda),\end{aligned}$$

where $H(\lambda) = \kappa\sqrt{\mathcal{H}(\mathcal{C}_\ell, \|\cdot\|_{nd}, \lambda)}$. Indeed, $\varphi_\ell(\lambda) \geq \phi_\ell(\lambda)$ by definition of φ_ℓ, and on the other hand

$$\begin{aligned}\frac{\varphi'_\ell(\lambda) - H(\lambda)}{\kappa\sqrt{nd}} &\geq \sqrt{\ln\left[2\left(\frac{\ell}{\lambda}\right)^{1/d}\right]} + \sqrt{\pi} - \frac{1}{2}\left(\ln\left[2\left(\frac{\ell}{\lambda}\right)^{1/d}\right]\right)^{-1/2} - \sqrt{\ln\left[2\left(\frac{\ell}{\lambda}\right)^{1/d}\right]} \\ &\geq \sqrt{\pi} - \frac{1}{2}\left(\ln\left(2\left(\frac{\ell}{\lambda}\right)^{1/d}\right)\right)^{-1/2} \\ &\geq \sqrt{\pi} - \frac{1}{2\sqrt{\ln 2}} \\ &\geq 0,\end{aligned}$$

which shows that $\varphi'_\ell(\lambda) \geq H(\lambda)$.

Then, let $\varphi_\ell(r) = \varphi_\ell(\lambda) + (r-\lambda)\varphi'_\ell(\lambda)$ for $r \geq \lambda$, so that, finally, $\phi_\ell(r) \leq \varphi_\ell(r)$ for all r.

4.4 Proof of Lemma 2.4

Notice first that φ_ℓ is concave. Indeed, the second derivative of the restriction $\varphi_\ell|_{]0,\lambda]}$ of φ_ℓ to $]0,\lambda]$ is equal to

$$-\frac{\kappa\sqrt{nd}}{2r}\left[\frac{1}{2}\left(\ln\left(\frac{2\ell^{1/d}\lambda^{1-1/d}}{r}\right)\right)^{-3/2}+\ln\left(\frac{2\ell^{1/d}\lambda^{1-1/d}}{r}\right)^{-1/2}\right]\leq 0,$$

which implies that $\varphi_\ell|_{]0,\lambda]}$ is concave. As φ_ℓ is obtained by extending $\varphi_\ell|_{]0,\lambda]}$ using the tangent to this function in λ, φ_ℓ is also concave.

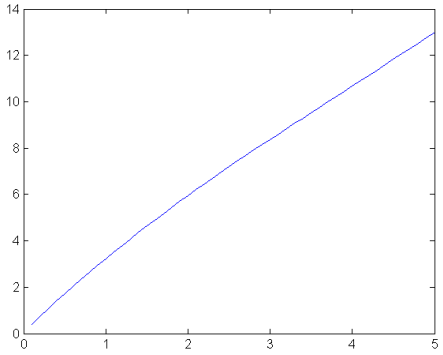


Figure 9: Function $\dfrac{\varphi_\ell}{\kappa\sqrt{nd}}$ for $d=2$, $\ell=6$ and $\lambda=3$.

Back to equation

$$\varphi_\ell\left(\frac{2\sigma\sqrt{d_\ell}}{\sqrt{nd}}\right)=\frac{\sigma d_\ell}{\sqrt{nd}},$$

observe that looking for a solution d_ℓ amounts to solving, for $r>0$,

$$\varphi_\ell(r)=\frac{\sqrt{nd}}{4\sigma}r^2.$$

This equation admits a unique solution $r_\ell=2\sigma\sqrt{\frac{d_\ell}{nd}}$, since φ_ℓ is concave and $r\mapsto r^2$ convex. Moreover, the solution r_ℓ satisfies $r_\ell\leq\lambda$ if, and only if,

$$\varphi_\ell(\lambda)\leq\frac{\sqrt{nd}}{4\sigma}\lambda^2,$$

that is

$$\kappa\lambda\sqrt{nd}\left(\sqrt{\ln 2+\frac{1}{d}\ln\left(\frac{\ell}{\lambda}\right)}+\sqrt{\pi}\right)\leq\frac{\sqrt{nd}}{4\sigma}\lambda^2,$$

which means that

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

If this condition is satisfied, the equation becomes

$$\kappa r_\ell \sqrt{nd} \left(\sqrt{\ln \left(\frac{2\ell^{1/d} \lambda^{1-1/d}}{r_\ell} \right)} + \sqrt{\pi} \right) = \frac{\sqrt{nd}}{4\sigma} r_\ell^2,$$

which is equivalent to

$$4\sigma\kappa \left(\sqrt{\ln \left(\frac{2\ell^{1/d} \lambda^{1-1/d}}{r_\ell} \right)} + \sqrt{\pi} \right) = r_\ell.$$

So, $4\sigma\kappa\sqrt{\pi} \leq r_\ell$, and then,

$$r_\ell \leq 4\sigma\kappa \left(\sqrt{\ln \left(\frac{\ell^{1/d} \lambda^{1-1/d}}{2\sigma\kappa\sqrt{\pi}} \right)} + \sqrt{\pi} \right).$$

Since $r_\ell = 2\sigma\sqrt{\frac{d_\ell}{nd}}$, we obtain

$$d_\ell \leq 8\kappa^2 nd \left(\ln \left(\frac{\ell^{1/d} \lambda^{1-1/d}}{2\sigma\kappa\sqrt{\pi}} \right) + \pi \right).$$

References

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

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

[3] 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.

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

[5] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 2011. In press. Available at `http://hal.archives-ouvertes.fr/docs/00/46/16/39/PDF/RR-7223.pdf`.

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

[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. Gaussian model selection. *Journal of the European Mathematical Society*, 3:203–268, 2001.

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

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

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

[12] C. Caillerie and B. Michel. Model selection for simplicial approximation. Technical Report 6981, INRIA, 2009. Available at `http://hal.inria.fr/docs/00/41/76/60/PDF/RR-6981.pdf`.

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

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

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

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

[17] 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.

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

[19] 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.

[20] 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`.

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

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

[23] 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.

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

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

[26] C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

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

[28] U. Ozertem and D. Erdogmus. Locally defined principal curves and surfaces. *Journal of Machine Learning Research*, 12:1249–1286, 2011.

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

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

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

[32] 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.

[33] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:33–73, 1978.

[34] 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.

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

[36] 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.

[37] U. von Luxburg, O. Bousquet, and B. Schölkopf. A compression approach to support vector model selection. *Journal of Machine Learning Research*, 5:293–323, 2004.

[38] 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.