Courbes principales dans un contexte d'apprentissage en ligne

Le Li [1], Benjamin Guedj [2] & Sébastien Loustau [3]

[1] *Université d'Angers & iAdvize, le@iadvize.com*
[2] *Équipe-projet MODAL, Inria, benjamin.guedj@inria.fr*
[3] *Artfact, artfact64@gmail.fr*

Résumé. Les courbes principales sont la généralisation non-linéaire de l'analyse en composantes principales. En général, une courbe principale passe continûment à travers le «milieu» des données avec une hypothèse probabiliste et en donne un résumé de manière non linéaire. Nous nous intéressons dans ce travail à la construction et à la mise en oeuvre d'une implémentation d'un programme qui utilise les *courbes principales* en ligne. Nous proposons un nouvel algorithme pour la construction de *courbes principales* qui résument séquentiellement des données en ligne, reposant sur l'approche quasi-bayésienne. En particulier, les sommets nécessaires pour former la *courbe principale* sont estimés dynamiquement (c'est-à-dire qu'ils peuvent changer au cours du temps). Nous démontrons de borne de regret de notre algorithme et nous donnons une implémentation via MCMC.

Mots-clés. Borne de regret, Courbes principales en ligne, MCMC, Estimateur quasi-bayésien.

Abstract. The principal curves are nonlinear generalisation of principal component analysis. By passing through the "middle" of a data set, a principal curve provides a summary of data set where they are often assumed to be observations from some probabilistic model. In this work, we will consider the principal curves in an online setting and we introduce a new algorithm to form principal curves for online data. Our procedure relies on quasi-Bayesian approach, allowing for a dynamic (*i.e.*, time-dependent) estimation of vertices that are necessary for the construction of principal curves. Its theoretical merits are supported by a regret bound and an MCMC-flavored implementation.

Keywords. Online principal curves, Quasi-Bayesian estimator, MCMC, Regret bound.

1 Introduction

Given a data set consisting of T observations of two variables x and y, we are interested in plotting them by a scatter plot and in trying to summarize the pattern exhibited by the observations. To do this, one often uses Principal Component Analysis (PCA) to orthogonally transform the observations into a set of values of linearly uncorrelated

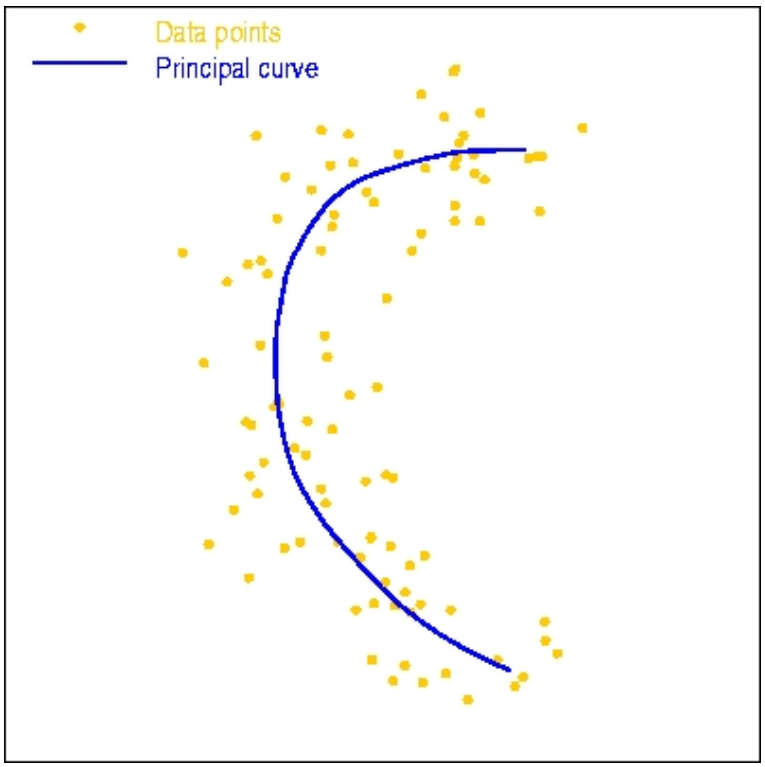

Figure 1: An example of principal curve

variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance of data set and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. However, in some circumstances, it is preferred to represent and summarize the data by a nonlinear pattern rather than by straight lines. One may resort to nonlinear regression where we treat one of the variables as an explanatory variable (eg, x) and another one as response variable (eg, y). We seek therefore to find a model to predict the response using the value of explanatory variable. More precisely, the expectation of y is modeled as a function f of x where f belongs to a family of functions with certain regularity. But sometimes choosing a suitable variable as response and others as explanatory for data in high dimension can be annoying. In addition, even in low dimension, we sometimes wish to treat them symmetrically rather than labelling one as a response variable.

The non-linearity of PCA and non-symmetry of regression leads to the notion of principal curves which pass through the middle of observations in a smooth way, as illustrated in Figure 1. The principal curves have been applied in many applications including physics, character recognition etc. The original definition of principal curve

dates back to Hastie and Stuetzle [6], where a smooth curve $\mathbf{f}(s) = (f_1(s), \ldots, f_d(s))$ is a principal curve if:

1. $\mathbf{f}$ does not intersect itself.
2. $\mathbf{f}$ has finite length inside any bounded subset of $\mathbb{R}^d$.
3. $\mathbf{f}$ is self-consistent, *i.e.*,

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

where $X \in \mathbb{R}^d$ is a random vector with a certain distribution, and $s_{\mathbf{f}}(x)$ is the largest s such that $\mathbf{f}(s)$ is closest to x by Euclidean distance $|\cdot|_2$. More precisely,

$$s_{\mathbf{f}}(x) = \sup \left\{ s : |x - \mathbf{f}(s)|_2 = \inf_{\tau} |x - \mathbf{f}(\tau)|_2 \right\}.$$

This self-consistency means that each point s of $\mathbf{f}$ is the average (under the distribution of X) of all points projected on s. However, an unfortunate property of principal curves defined by Hastie and Stuetzle [6] is that in general, it is not known if they exist for a given source of distribution, let along online data for which no probabilistic assumption is assumed. In our work, we will adopt the definition of principal curve by Kégl, Krzyżak, Linder and Zeger [7] where a constraint of the length of curve is imposed. The advantage of this definition is the existence of principal curve (in the new sense) for large class of source distribution and it avoids the computing of condition expectation in (1) which is infeasible in online setting.

Learning in online setting has been extensively studied last decades in game theory and statistics, see [1]. Research efforts in online learning began in the framework of prediction with experts advices, then was extended to regression and clustering framework [2]. The ambition of our work is to propose a principal curve algorithm suited to online setting. For this task, our strategy consists in two steps:

1. The Gibbs quasi-posterior which depends on the choice of a sparsity-inducing prior. This approach is motivated by the quasi-Bayesian theory which has been extensively studied by numerous researchers in batch setting and developed by Audibert [3], Loustau [4] and Gerchinovitz [5] in online setting, among others.
2. The MCMC algorithm which allows effective simulations from the Gibbs quasi-posterior of the complex-structured space.

2 Notation

A parameterized curve in $\mathbb{R}^d$ is a continuous function $\mathbf{f} : I \longrightarrow \mathbb{R}^d$ where $I = [a, b]$ is a closed interval of the real line. The length of $\mathbf{f}$ is defined by

$$\mathcal{L}(\mathbf{f}) = \sup \sum_{i=1}^{M} |\mathbf{f}(s_i) - \mathbf{f}(s_{i-1})|_2,$$

where the supremum is taken over all subdivisions $a = s_0 < s_1 < \cdots < s_M = b, M > 1$.

Let $x_1, x_2, \ldots, x_T \in \mathbf{B}_d(R) \subset \mathbb{R}^d$ be an online sequence, where $\mathbf{B}_d(R)$ is an ℓ_2-ball centered in $0 \in \mathbb{R}^d$ with radius $R > 0$. Let $\mathcal{C}$ be a grid over $\mathbf{B}_d(R)$, that is $\mathcal{C} = \mathbf{B}_d(R) \cap \Gamma$, where Γ is the lattice of distance Δ of $\mathbb{R}^d$. Denote by $\mathcal{V} = \{v_1, v_2, \ldots, v_p\}$ all the vertices in $\mathcal{C}$. Let $L > 0$ and define for each $k \in \{1, \ldots, p-1\}$ a collection:

$$\mathcal{F}_{k,L} = \{\mathbf{f} : \text{polygonal line with } k \text{ segments whose vertices are in } \mathcal{V} \text{ and } \mathcal{L}(\mathbf{f}) \leq L\}.$$

Denote by $\mathcal{F} = \cup_{k=1}^{p-1} \mathcal{F}_{k,L}$ all polygonal lines whose vertices are in $\mathcal{V}$ and length at most L. It is clear that the smaller the distance Δ, the bigger the number of vertices p in $\mathcal{C}$. However, to ensure that the lengths of all polygonal lines do not exceed L, it suffices that $\Delta \geq (2R)^{(d+2)/d}/L^{1/d}$, hence we have $p \leq L/(4R^2 d^{d/2})$.

Our goal is to learn a time-dependent polygonal line which passes through the middle and gives a summary of all available observations $(x_s)_{1:(t-1)}$ right before time t. To this aim, the output of our algorithm at time t is a polygonal line $\hat{\mathbf{f}}_t \in \mathcal{F}$, depending on past information $(x_s)_{1:(t-1)}$ and past predictions $(\hat{\mathbf{f}}_s)_{1:(t-1)}$. When x_t is revealed, the instantaneous loss of $\hat{\mathbf{f}}_t$ and x_t is computed as

$$\ell\left(\hat{\mathbf{f}}_t, x_t\right) = \inf_{s \in I} |\hat{\mathbf{f}}_t(s) - x_t|_2^2.$$

In the sequel, denote by π a prior on $\mathcal{F}$ and let $\lambda > 0$ be some (inverse temperature) parameter. The output $\hat{\mathbf{f}}_{t+1}$ is constructed as follows: at each time t, we observe x_t and a polygonal line $\hat{\mathbf{f}}_{t+1} \in \mathcal{F}$ is sampled from the Gibbs quasi-posterior $\hat{\rho}_{t+1}$ in Algorithm 1 below.

Note that the last term in step 5 is a consequence of the non-convexity of the loss ℓ (see [3]) and that the curve $\hat{\mathbf{f}}_{t+1}$ is a realization of $\hat{\rho}_{t+1}$.

3 Regret bound of the algorithm

We present here our main theoretical result.

Algorithm 1 A quasi-Bayesian algorithm for online principal curves

1: **Input parameters**: $p > 0, \pi$, $\lambda > 0$ and $S_0 \equiv 0$
2: **Initialization**: Draw $\hat{\mathbf{f}}_1 \sim \pi$
3: **For** $t = 1, \ldots, T-1$
4: Get the data x_t
5: Draw $\hat{\mathbf{f}}_{t+1} \sim \hat{\rho}_{t+1}(\mathbf{f})$ where $\mathrm{d}\hat{\rho}_{t+1}(\mathbf{f}) \propto \exp\big(-\lambda S_t(\mathbf{f})\big)\mathrm{d}\pi(\mathbf{f})$, and

$$S_t(\mathbf{f}) = S_{t-1}(\mathbf{f}) + \ell\left(\mathbf{f}, x_t\right) + \frac{\lambda}{2}\left(\ell\left(\mathbf{f}, x_t\right) - \ell\left(\hat{\mathbf{f}}_t, x_t\right)\right)^2.$$

6: **End for**

Theorem 1. *For any sequence $(x_t)_{1:T} \in \mathbb{R}^{dT}$, and $L > 0$, there exists a prior π and a $\lambda > 0$ such that the procedure described in Algorithm 1 satisfies*

$$\sum_{t=1}^{T} \mathbb{E}_{(\hat{\rho}_1, \hat{\rho}_2, \ldots, \hat{\rho}_t)} \ell(\hat{\mathbf{f}}_t, x_t) \;\leq\; \inf_{k \in \{1,\ldots,p-1\}} \left\{ \inf_{\mathbf{f} \in \mathcal{F}(k,L)} \sum_{t=1}^{T} \ell(\mathbf{f}, x_t) + C_{R,d,L}\sqrt{kT} \right\} + C_d\sqrt{LT}.$$

where $R \geq \max_{t=1,\ldots,T} |x_t|_2$ and $C_{R,d,L}$ is a constant depending on R, d, L while C_d is a constant only depending on d.

This theorem indicates that the expected cumulative loss of polygonal lines $\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_T$ is close to the minimal cumulative loss over all $k \in \{1, \ldots, p-1\}$, up to a term that is sublinear in T. In addition, we see that this term is proportional to $\sqrt{kT}$ and penalizes the number k of segments which is a measurement of the complexity of polygonal lines when L is fixed. This result is consistent in order of k and number T of observations with that of Biau and Fischer [8] where parameter selection for principal curves via penalization is considered in a statistical setting. The implementation required to sample at each t from the Gibbs quasi-posterior $\hat{\rho}_{t+1}$. Since $\hat{\rho}_t$ is defined on the massive and complex-structured space $\mathcal{F}$ (let us recall that $\mathcal{F}$ is a union of heterogeneous spaces), direct sampling from $\hat{\rho}_t$ is not an option and is much rather an algorithmic challenge. Our approach consists in approximating $\hat{\rho}_t$ by MCMC.

Bibliographie

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

[2] L. Li, B. Guedj and S. Loustau (2016), *PAC-Bayesian Online Clustering*, HAL preprint https://hal.inria.fr/hal-01264233.

[3] J. Y. Audibert (2009), *Fast learning rates in statistical inference through aggregation*, The Annals of Statistics, 37(4): 1591–1646.

[4] S. Loustau (2014), *Online clustering of individual sequence*, HAL preprint https://hal.archives-ouvertes.fr/hal-00943384.
[5] S. Gerchinovitz (2011), *Prédiction de suites individuelles et cadre statistique classique : étude de quelques liens autour de la régression parcimonieuse et des techniques d'agrégation*, PhD thesis, Université Paris-Sud.
[6] T. Hastie and W. Stuetzle (1989), *Principal curves*, Journal of the American Statistical Association, 84:502–516.
[7] B. Kégl, A. Krzyżak, T. Linder and K. Zeger (2000), *Learning and design of principal curves*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22:281–297.
[8] G. Biau and A. Fischer (2012), *Parameter selection for principal curves*, IEEE Transactions on Information Theory 58(3).