Courbes Principales et Sélection de Modèle

Aurélie Fischer

Courbes principales et sélection de modèle

Aurélie Fischer

Université Pierre et Marie Curie - Paris VI
Laboratoire de Statistique Théorique et Appliquée (LSTA)
175, rue du Chevaleret, boîte 158
75013 PARIS

Résumé

Les courbes principales sont une généralisation non linéaire de la notion de première composante principale. Intuitivement, une courbe principale est une courbe de $\mathbb{R}^d$ passant au "milieu" d'une distribution de probabilité en dimension d ou d'un nuage de données de $\mathbb{R}^d$. Plusieurs définitions ont été proposées, dont l'une repose sur la minimisation d'un critère des moindres carrés. Nous nous intéressons au choix de la classe sur laquelle minimiser ce critère pour obtenir une courbe principale qui résume au mieux la forme des données sans pour autant conduire à de l'interpolation, et adoptons le point de vue de la sélection de modèle par pénalisation.

Mots-clés : courbes principales, sélection de modèle

Abstract

Principal curves are a non linear generalization of the notion of first principal component. Intuitively, a principal curve is a curve in $\mathbb{R}^d$ which pass through the "middle" of a d-dimensional probability distribution or a data cloud in $\mathbb{R}^d$. Several definitions have been proposed, one of which is based on the minimization of a least squares criterion. We are interested in choosing a suitable class over which this criterion should be minimized in order to obtain a principal curve which recovers the shape of the data without interpolating, and we consider this problem from the point of view of model selection via penalization.

Keywords : principal curves, model selection

1 Définition des courbes principales

La statistique utilise différents moyens de résumer de l'information, en représentant des données par certaines grandeurs "simplifiées". L'une de ces méthodes, fréquemment utilisée, est l'analyse en composantes principales, dont le but est de déterminer les axes de variabilité maximale d'un nuage de données. Dans certaines situations, plutôt que de les représenter à partir de droites, il peut être intéressant de résumer les données de manière non linéaire. Ceci conduit à la notion de courbe principale, qui est une généralisation de la première composante principale. De manière intuitive, il s'agit de rechercher une courbe régulière passant "au milieu" des données (Figure 1). Il existe plusieurs moyens de donner un sens mathématique à cette idée, et ainsi de définir les courbes principales. La définition dépend par exemple de la propriété des composantes principales que l'on choisit de généraliser.

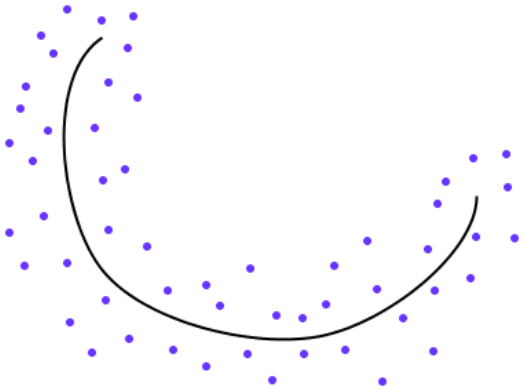

Figure 1: Un exemple de courbe principale

Soit X un vecteur aléatoire de $\mathbb{R}^d$ tel que $\mathbb{E}\|X\|^2 < +\infty$ et $\mathbb{E}X = 0$. Soit $f = (f_1, \ldots, f_d)$ une courbe de $\mathbb{R}^d$, continue, paramétrée par l'arc sur un intervalle $I = [a, b]$. Soit $t_f : \mathbb{R}^d \to \mathbb{R}$ l'indice de projection défini par

$$t_f(x) = \sup\{t, \|x - f(t)\| = \inf_{t'} \|x - f(t')\|\},$$

où $\|\cdot\|$ désigne la norme euclidienne de $\mathbb{R}^d$. La première définition, proposée par Hastie et Stuetzle (1989), repose sur la propriété d'auto-consistance. Une courbe f de classe C^∞ est une courbe principale pour X si elle est sans point double (f est injective), de longueur finie à l'intérieur de toute boule de $\mathbb{R}^d$, et auto-consistante, c'est-à-dire, pour presque tout t,

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

Hastie et Stuetzle (1989) donnent également un algorithme itératif qui permet d'approcher une courbe principale pour X en alternant une étape de projection et une étape de calcul d'espérance conditionnelle. Ces auteurs proposent également une version de l'algorithme pour trouver une courbe principale lorsque la loi de X est inconnue, mais que l'on dispose d'un échantillon $X_1, \ldots, X_n$.

Avec cette définition basée sur l'auto-consistance, l'existence de courbes principales n'a été prouvée que dans certains cas particuliers. Kégl, Krzyżak, Linder et Zeger (2000) donnent une définition un peu différente, qui leur permet de démontrer l'existence de courbes principales dès que $\mathbb{E}\|X\|^2 < +\infty$. Cette définition n'est pas implicite, mais repose sur la minimisation d'un critère des moindres carrés. Les courbes considérées sont continues, mais pas nécessairement différentiables, de sorte qu'une ligne polygonale peut être une courbe principale. Une courbe f^* est une courbe principale de longueur L pour X si f^* minimise

$$\Delta(f) = \mathbb{E} \inf_t \|X - f(t)\|^2 = \mathbb{E}\|X - f(t_f(X))\|^2$$

parmi toutes les courbes de longueur inférieure ou égale à L. La même définition est valable dans le cas d'un échantillon $X_1, \ldots, X_n$, en remplaçant $\Delta(f)$ par le critère empirique

$$\Delta_n(f) = \frac{1}{n} \sum_{i=1}^{n} \|X_i - f(t_f(X_i))\|^2.$$

Kégl, Krzyżak, Linder et Zeger (2000) proposent un algorithme construisant une courbe principale à partir de lignes polygonales. On peut observer que l'analyse en composantes principales classique ne s'inscrit pas tout à fait dans ce cadre, qui impose qu'une courbe principale soit de longueur bornée. Partant de ce constat, Sandilya et Kulkarni (2002) proposent une autre définition, basée sur les mêmes critères, mais dans laquelle la contrainte ne porte pas sur la longueur des courbes, mais sur une notion de courbure.

D'autres manières de définir les courbes principales ont été envisagées. Ainsi, Tibshirani (1992) présente une définition en termes de modèle de mélange, associée à un algorithme reposant sur l'algorithme EM. La caractérisation des courbes principales de Delicado (2001) est une extension au cas non linéaire d'une propriété de la loi normale multivariée, qui exprime que la projection sur l'hyperplan orthogonal à la première composante principale minimise, parmi toutes les projections sur des hyperplans, la trace de la matrice de covariance.

2 Un résultat de sélection de modèle

Une question qui se pose est celle du choix d'une classe appropriée de courbes dans laquelle chercher une courbe principale. Il s'agit de trouver un compromis entre un modèle trop petit pour donner une bonne approximation d'une courbe principale et un modèle contenant trop de courbes, qui pourrait conduire par exemple à une courbe principale passant par tous les points des données. Pour cela, nous adoptons le point de vue de la sélection de modèle par pénalisation développée par Birgé et Massart (voir par exemple Birgé et Massart, 2001).

Dorénavant, $\|\cdot\|$ désigne la norme de $\mathbb{R}^m$ définie par $\langle x, y\rangle = \frac{1}{m}\sum_{i=1}^m x_i y_i$. On suppose que $X_1, \ldots, X_n$ sont des vecteurs aléatoires à valeurs dans $\mathbb{R}^d$ tels que

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

où les x_i sont inconnus et supposés appartenir au support d'une courbe f^*, les ξ_i sont des vecteurs gaussiens standards de $\mathbb{R}^d$ indépendants, et σ désigne l'intensité du bruit. Soit $\mathbf{X} = {}^t({}^tX_1, \ldots, {}^tX_n)$ le vecteur constitué de tous les X_i. Les vecteurs $\mathbf{x}$ et $\boldsymbol{\xi}$ étant définis de même, (1) se récrit sous la forme

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

On a alors $\mathbf{x} \in (\text{supp } f^*)^n$. On considère le critère de Kégl, Krzyżak, Linder et Zeger (2000)

$$\Delta_n(f) = \frac{1}{n}\sum_{i=1}^n \|X_i - f(t_f(X_i))\|^2,$$

et on se donne une collection dénombrable $\{\mathcal{F}_\ell\}_{\ell\in L}$, où chaque $\mathcal{F}_\ell$ est une classe de courbes continues de $\mathbb{R}^d$ de longueur ℓ, ayant pour extrémités deux points bien choisis à partir des données. Le but est de sélectionner la longueur ℓ. On a

$$\Delta_n(f) = \frac{1}{n}\sum_{i=1}^n \inf_t \|X_i - f(t)\|^2 = \frac{1}{n}\sum_{i=1}^n \inf_{x_i \in \text{ supp}f} \|X_i - x_i\|^2.$$

Supposons que pour tout $\ell \in L$, $\hat{\mathbf{x}}_\ell$ minimise $\|\mathbf{X} - \mathbf{x}\|^2$ sur $\mathcal{C}_\ell = \bigcup_{f\in\mathcal{F}_\ell}(\text{supp}f)^n$. On cherche à minimiser en ℓ un critère du type

$$\text{crit}(\ell) = \|\mathbf{X} - \hat{\mathbf{x}}_\ell\|^2 + \text{pen}(\ell),$$

où pen : $L \to \mathbb{R}^+$ est une fonction de pénalité convenable.

On constate que les modèles $\mathcal{C}_\ell$ ne sont pas des sous-espaces vectoriels. Pour mesurer la complexité d'un modèle non linéaire, on utilise l'entropie métrique, qui est définie de la manière suivante. Pour S un sous-ensemble de $\mathbb{R}^{nd}$, un ε-réseau S_ε est un sous-ensemble fini S_ε de cardinal maximal tel que $\|x-y\| > \varepsilon$ pour tous $x, y \in S_\varepsilon$. Soit $N(S, \|\cdot\|, \varepsilon)$ le cardinal maximal d'un ε-réseau de S. L'entropie métrique de S est donnée par

$$H(S, \|\cdot\|, \varepsilon) = \ln N(S, \|\cdot\|, \varepsilon).$$

Souvent, il est plus facile de calculer $N'(S, \|\cdot\|, \varepsilon)$, le nombre minimal de boules de rayon ε nécessaires pour recouvrir S. Les deux quantités sont reliées par l'inégalité

$$N'(S, \|\cdot\|, \varepsilon) \leq N(S, \|\cdot\|, \varepsilon) \leq N'(S, \|\cdot\|, \varepsilon/2).$$

Nous obtenons un résultat de sélection de modèle de la forme du Théorème 4.18 de Massart (2007). Soit

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

où κ est une constante à calibrer. On suppose qu'il existe des poids $(w_\ell)_{\ell \in L}$ vérifiant

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

Si $\eta > 1$, σ n'est pas trop grand, et

$$\text{pen}(\ell) \geq \eta \frac{\sigma^2}{nd} (\sqrt{d_\ell} + \sqrt{2w_\ell})^2,$$

où d_ℓ dépend de Φ_ℓ, n, d et σ, alors, presque sûrement, il existe un minimiseur $\hat{\ell}$ du critère pénalisé

$$\text{crit}(\ell) = \|\mathbf{X} - \hat{\mathbf{x}}_\ell\|^2 + \text{pen}(\ell).$$

En notant $\tilde{\mathbf{x}} = \hat{\mathbf{x}}_{\hat{\ell}}$, on a, pour tout $\mathbf{x} \in \mathbb{R}^{nd}$,

$$\mathbb{E}_{\mathbf{x}} \|\tilde{\mathbf{x}} - \mathbf{x}\|^2 \leq c_\eta \Big[\inf_{\ell \in L} \{ (\inf_{\mathbf{y} \in \mathcal{C}_\ell} \|\mathbf{x} - \mathbf{y}\|)^2 + \text{pen}(\ell) \} + \frac{\sigma^2}{nd} (\Sigma + 1) \Big].$$

Bibliographie

[1] Birgé, L., et Massart, P. (2001) Gaussian model selection. *Journal of European Mathematical Society*, 3, 203–268.

[2] Hastie, T. et Stuetzle, W. (1989) Principal Curves. *Journal of the American Statistical Association*, 84(406), 502–516.

[3] Kégl, B., Krzyzak, A., Linder, T. et Zeger, K. (2000) Learning and Design of Principal Curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(3).

[4] Massart, P. (2007) Concentration Inequalities and Model Selection. *École d'eté de probabilités de Saint-Flour XXXIII - 2003, Lecture notes in mathematics, Springer.*

[5] Sandilya, S. et Kulkarni, S. R. (2002) Principal Curves With Bounded Turn. *IEEE Transactions on Information Theory*, 48(10), 2789–2793.

[6] Tibshirani, R. (1992) Principal curves revisited. *Statistics and Computing*, 2, 183–190.