ACP pour données fonctionnelles discrétisées, estimation minimax et contraintes spectrales

Nassim Bourarach [1] & Franck Picard [2]& Vincent Rivoirard [1] & Angelina Roche [1]

[1] *CEREMADE, Université Paris-Dauphine PSL , nassim.bourarach@dauphine.psl.eu, Vincent.Rivoirard@dauphine.fr, roche@ceremade.dauphine.fr*
[2] *LBMC, ENS de Lyon, franck.picard@ens-lyon.fr*

Résumé. Dans cette présentation, nous explorerons le problème de l'estimation des fonctions propres et valeurs propres de l'opérateur de covariance d'un échantillon de données fonctionnelles discrétisées et bruitées. L'analyse de l'impact de la discrétisation et du bruit sur l'inférence se fera grâce à une double asymptotique : en n, le nombre de courbes observées, et p, la taille de la grille de discrétisation.
Nous aborderons le problème pour une classe de processus très large (trajectoires m-Höldériennes avec $m \in \mathbb{R}^*$) et expliciterons le rôle de la régularité dans les bornes. Ces nouvelles bornes inférieures minimax seront accompagnées d'un 'nouvel' estimateur non-paramétrique, fondé sur des ondelettes, qui est optimal sans nécessiter de régularisation. Après avoir présenté le cadre et nos résultats, nous discuterons plus en détail des contraintes spectrales nécessaires à travers des illustrations et des résultats d'inconsistance.

Mots-clés. données fonctionnelles, ACP, borne minimax, estimation non-paramétrique

Abstract. In this presentation, we will explore the problem of estimating the eigenfunctions and eigenvalues of the covariance operator of a sample of discretised and noisy functional data. The impact of discretisation and noise on inference will be analysed through a double asymptotic: in n, the number of observed curves, and p, the size of the discretisation grid. We will address the problem for a very broad class of processes (m-Hölderian trajectories with $m \in \mathbb{R}_+^*$) and specify the role of regularity in the bounds. These new minimax lower bounds will be accompanied by a 'new' non-parametric estimator, based on wavelets, that is optimal without the need for regularisation. After presenting the framework and the main results, we will discuss the necessary spectral constraints in more details by means of inconsistency results and illustrations.

Keywords. functional data, PCA, minimax bound, non-parametric estimation

1 Introduction

1.1 Les données fonctionnelles

Pour faire face à l'afflux de données de plus en plus massives, beaucoup d'approches statistiques différentes ont vu le jour. Dans de nombreuses situations, les données dont on dispose se présentent comme des vecteurs d'observations en grande dimension qui cachent une dépendance importante entre les différentes entrées du vecteur. Il peut s'agir d'une dépendance spatiale ou temporelle par exemple. Ces dépendances peuvent d'une certaine façon être encapsulées dans une modélisation fonctionnelle des observations, à travers des contraintes de régularité. Ainsi, plutôt que de considérer qu'on observe des vecteurs aléatoires en très grande dimension, on considère qu'on observe des réalisations (potentiellement bruitées) d'une variable aléatoire $X_i \overset{i.i.d.}{\sim} X$ à valeur dans un espace de fonctions $\mathcal{H} \subset \mathbb{L}_2([0,1])$ l'espace des fonctions de carré intégrable sur $[0,1]$. C'est notamment le point de vue adopté dans de nombreux ouvrages dédiés aux données fonctionnelles (Ferraty et Vieu (2006), Ramsay et Silverman (2010), Hsing et Eubank (2015)).

Ce premier saut conceptuel nous éloigne de la réalité des données (on n'observe jamais plus que des vecteurs) mais permet l'accès à la multitude de résultats relatifs aux espaces de fonctions. Nos travaux s'inscrivent alors dans une seconde perspective où l'on essaie de réconcilier l'approche fonctionnelle avec les données observées, par la prise en compte de l'aspect discret des données. Les nouveaux résultats que nous obtenons prennent place dans le prolongement de ceux de Belhakem et al. (2022).

De ce fait, en ayant p, n deux entiers naturels strictement positifs, nos données consistent en p évaluations bruitées de n réalisations de fonctions aléatoires sur une grille $(t_j)_{j=1}^p \in [0,1]$:

$$Y_i(t_j) = X_i(t_j) + \varepsilon_{i,j}, \qquad (i,j) \in [\![1,n]\!] \times [\![1,p]\!]. \tag{1}$$

où les $\varepsilon_{i,j}$ sont i.i.d. gaussiennes, centrées, de variance $\sigma^2 > 0$ fixée, indépendantes des X_i qui sont aussi i.i.d. .

1.2 Le problème

L'Analyse en Composantes Principales fonctionelles (ACPf), tout comme son équivalent vectoriel, joue un rôle important, à la fois comme outil de réduction de la dimension ou d'analyse exploratoire des données fonctionnelles. En pratique c'est ce qui est observé dans les papiers de Viviani et al. (2005) sur des données fMRI ou plus récemment par Warmenhoven et al. (2021) pour des données issues de la biomécanique.

Soit D une dimension fixée, l'ACPf consiste à trouver un système de D fonctions $(\psi_1^*, \ldots, \psi_D^*)$ minimisant l'écart quadratique entre une donnée fonctionnelle X et sa projection orthogonale $\Pi_{S_D} X$ sur l'espace $S_D^* = \text{Vect}\{\psi_1^*, \ldots, \psi_D^*\}$ autrement dit

$$S_D^* \in \arg \min_{\substack{S \text{ s.e.v. de } \mathbb{L}_2([0,1],\\ \dim(S)=D}} E\left[\|X - \Pi_S X\|^2\right], \tag{2}$$

où $\Pi_S X$ est la projection orthogonale sur l'espace $S \subset \mathbb{L}_2([0,1])$ avec le produit scalaire usuel défini par $\langle f, g\rangle = \int_0^1 f(t)g(t)\,dt$, $f, g \in \mathbb{L}_2$ et $\|\cdot\|$ la norme associée. Dans l'expression précédente et le reste du document, E désigne l'espérance sous la loi de X.

Supposons $E\left[\|X\|^2\right] < +\infty$, pour que (2) ait un sens et définissons

$$\begin{aligned}\Gamma : \mathbb{L}_2 &\to \mathbb{L}_2\\ f &\mapsto E[\langle X - E[X], f\rangle(X - E[X])],\end{aligned}$$

l'opérateur de covariance associé aux données. Comme Γ est un opérateur compact et auto-adjoint, il existe une base hilbertienne de $\mathbb{L}_2$ composée de fonctions propres de Γ. D'autre part, il est possible de montrer que la solution de (2) est l'espace engendré par les fonctions propres $\psi_1^*, \ldots, \psi_D^*$ de l'opérateur Γ associées aux D plus grandes valeurs propres $\lambda_1^*, \ldots, \lambda_D^*$ (comptées avec multiplicité) et que cette solution est unique si ces valeurs propres sont toutes distinctes.

On va chercher à étudier l'erreur quadratique d'estimation des éléments propres en question à partir de données qui se présentent sous la forme discrétisée bruitée (1).

2 Nouveaux résultats pour la théorie minimax du problème

2.1 Le choix de modélisation de l'aléa

Nous adoptons une modélisation des variables aléatoires $X_i \overset{i.i.d.}{\sim} P_X$ avec $P_X \in \mathcal{P}$ une classe non-paramétrique de mesure de probabilité. En revanche, pour construire des estimateurs consistants, des contraintes sur la richesse de la classe $\mathcal{P}$ sont nécessaires. Nous définissons cette classe en deux temps :
On va s'intéresser à l'estimation des ℓ-ièmes éléments propres de l'opérateur de covariance associé aux données, avec $\ell \in [\![1, p]\!]$.
On définit dans un premier temps une classe de noyaux dotés d'une certaine régularité pour tout $(m, L) \in {\mathbb{R}_+^*}^2$

$$\mathcal{K}(m,L) := \left\{ K : [0,1]^2 \mapsto \mathbb{R} \ \ \lfloor m \rfloor\text{-differentiable} \middle| \ \forall u, u', t \in]0,1[^3, K(u,t) = K(t,u), \right.$$
$$\left. \left| \frac{\partial^{\lfloor m \rfloor} K}{(\partial x_1)^{\lfloor m \rfloor}}(u,t) - \frac{\partial^{\lfloor m \rfloor} K}{(\partial x_1)^{\lfloor m \rfloor}}(u',t) \right| \leq L \left|u - u'\right|^{1+(m-\lfloor m \rfloor - 1)\mathbb{I}_{m \notin \mathbb{N}\setminus\{0\}}} \right\}.$$

Puis on définit, sur l'espace des opérateurs compacts définis positifs de $\mathbb{L}_2([0,1])$ (qu'on notera $\mathcal{L}$), la forme suivante

$$\begin{aligned}\mathrm{r}_\ell : \mathcal{L} &\to \mathbb{R}\\ G &\mapsto r_\ell(G) := \frac{\lambda_\ell^*(G)\lambda_{\ell+1}^*(G)}{\left|\lambda_\ell^*(G) - \lambda_{\ell+1}^*(G)\right|^2} + \mathbb{I}_{\ell \neq 1}\frac{\lambda_\ell^*(G)\lambda_{\ell-1}^*(G)}{\left|\lambda_\ell^*(G) - \lambda_{\ell-1}^*(G)\right|^2},\end{aligned}$$

où $\lambda_j^*(G)$ correspond à la j-ième valeur propre de G.
Notons que des quantités analogues sont aussi apparues dans d'autres travaux liés à l'ACPf (voir la notion de "rang relatif" de Jirak et Wahl (2018) ou encore les travaux de Mas et Ruymgaart (2015)).

Enfin, on spécifie, pour toutes suites $(c_n)_{n\in\mathbb{N}\setminus\{0\}}, (d_p)_{p\in\mathbb{N}\setminus\{0\}} \in \mathbb{R}_+^{\mathbb{N}}$ telles que $c_n = o_n(n)$ et $d_p = o_p\left(p^m\right)$, la classe de mesures de probabilité modélisant la loi des X_i,

$$\mathcal{P}(\ell, c_n, d_p, m) := \Bigg\{ P, \text{ mesure de probabilité associée à un processus à trajectoires continues centrés et dont l'opérateur de covariance } \Gamma \text{ est tel que } \mathrm{r}_\ell(\Gamma) \leq c_n, \max\left(\frac{\lambda_1^*(\Gamma)}{\lambda_\ell^*(\Gamma)}, \lambda_\ell^*(\Gamma)^{-1}\right) \leq d_p, \left(\lambda_{\ell-1}^*(\Gamma)\lambda_\ell^*(\Gamma)\right)^2 = o_n(n) \text{ et que } \forall (s,t) \in [0,1]^2, K(s,t) := \int_{\mathcal{C}^0} z(t)z(s)\, dP(z) \in \mathcal{K}\left(m, 8(2\pi)^m \sum_{j=1}^{\ell-1} j^m \lambda_j^*(\Gamma)\right). \Bigg\},$$

on rappelle que $o_n(\cdot)$ et $o_p(\cdot)$ désignent des quantités négligeables face à la quantité $\cdot$ lorsque, respectivement, $n \to \infty$ et $p \to \infty$.

Nous justifierons ce choix de modèle de façon intuitive, puis nous verrons qu'il est soutenu théoriquement par des théorèmes d'inconsistance et empiriquement par des simulations numériques.

2.2 Les bornes inférieures minimax

Nous présenterons ce qui est à notre connaissance la première borne inférieure minimax d'estimation des fonctions propres qui prend en compte la discrétisation et l'impact des contraintes mises sur les spectres des opérateurs de covariance des processus en question. On peut aussi y ajouter une borne inférieure minimax d'estimation des valeurs propres pour obtenir le résultat suivant

Théorème. *(Borne inférieure minimax pour les ℓ-ièmes éléments propres)*
Soit $\ell \in [\![1, p]\!], m \in \mathbb{R}_+^*$, $(c_n)_{n\in\mathbb{N}\setminus\{0\}}, (d_p)_{p\in\mathbb{N}\setminus\{0\}} \in \mathbb{R}_+^{\mathbb{N}}$ telles que $c_n = o_n(n)$ et $d_p = o_p\left(p^m\right)$.
Si on dispose de données sous la forme (1) avec $X_i \overset{i.i.d.}{\sim} P_X$, alors il existe $C, C' > 0$ qui ne dépendent pas des autres paramètres du modèle telles que

$$\inf_{\widehat{\psi}_\ell} \sup_{P_X \in \mathcal{P}(\ell, c_n, d_p, m)} \mathbb{E}\left[\left\|\widehat{\psi}_\ell - \psi_\ell^*\right\|^2\right] \geq C\left(\frac{c_n}{n} + \frac{d_p^2}{p^{2m}}\right),$$
$$\inf_{\widehat{\lambda}_\ell} \sup_{P_X \in \mathcal{P}(\ell, c_n, d_p, m)} \mathbb{E}\left[\left(\widehat{\lambda}_\ell - \lambda_\ell^*\right)^2\right] \geq C'\left(\frac{1}{n} + \frac{d_p^2}{p^{4m}}\right).$$

Dans les cas usuellement traités dans la littérature, on suppose simplement (en plus de la régularité du noyau de covariance) que nos processus sont tels qu'on a $a, C_0 \in \mathbb{R}_+^*$,

$$\begin{aligned} &\lambda_j(\Gamma) = C_0 j^{-(a+1)}, \ \forall j, n, p \in \mathbb{N} \setminus \{0\}, \\ \text{ou } &\lambda_j(\Gamma) = C_0 \exp(-ja), \ \forall j, n, p \in \mathbb{N} \setminus \{0\}. \end{aligned} \tag{3}$$

Il s'agit d'un cas particulier de notre modèle. Le cas échéant donne à partir du Théorème le Corollaire suivant :

Corollaire. *(Borne inférieure minimax pour les ℓ-ièmes éléments propres)*
Soit $\ell \in [\![1, p]\!], m \in \mathbb{R}_+^*$.
Sous le schéma d'observation (1) avec $X_i \overset{i.i.d.}{\sim} P_X$, en supposant qu'on a (3), il existe $C_\ell, C'_\ell > 0$, dont les seules dépendances en les paramètres du modèle sont en ℓ, telles que

$$\inf_{\widehat{\psi}_\ell} \sup_{P_X \in \mathcal{P}(\ell, C_\ell, C_\ell, m)} \mathbb{E}\left[\left\|\widehat{\psi}_\ell - \psi_\ell^*\right\|^2\right] \geq C'_\ell \left(\frac{1}{n} + \frac{1}{p^{2m}}\right),$$
$$\inf_{\widehat{\lambda}_\ell} \sup_{P_X \in \mathcal{P}(\ell, C_\ell, C_\ell, m)} \mathbb{E}\left[\left(\widehat{\lambda}_\ell - \lambda_\ell^*\right)^2\right] \geq C'_\ell \left(\frac{1}{n} + \frac{1}{p^{4m}}\right).$$

2.3 Estimateurs minimax-optimaux

Dans un dernier temps, nous présenterons des estimateurs simples, basés sur une projection sur une base d'ondelettes, qui sont optimaux au sens minimax sous quelques hypothèses supplémentaires. Nous expliciterons le cheminement des idées ayant abouti à de tels estimateurs. Si on se place dans le cadre du corollaire on a alors le théorème suivant

Théorème. *(Borne supérieure pour les ℓ-ièmes éléments propres)*
Soit $(\widehat{\lambda}_\ell, \widehat{\psi}_\ell)$ nos estimateurs.
On se place sous le schéma d'observation (1) avec $X_i \overset{i.i.d.}{\sim} P_X$, en supposant qu'on a (3), que les trajectoires des X_i sont p.s. m-Hölder et qu'il existe $M > 0$ tel que $\|X\| \leq M$ p.s..
Alors il existe $C_{\ell,\sigma,M} > 0$ qui ne dépend que de ℓ, σ et M telle que

$$\mathbb{E}\left[\left\|\widehat{\psi}_\ell - \psi_\ell^*\right\|^2\right] \leq C_{\ell,\sigma,M}\left(n^{-1} + p^{-2m}\right),$$
$$\mathbb{E}\left[\left(\widehat{\lambda}_\ell - \lambda_\ell^*\right)^2\right] \leq C_{\ell,\sigma,M}\left(n^{-1} + p^{-2m}\right).$$

Remarque. Les conditions du théorème précédent impliquent qu'il existe $C_\ell > 0$ telle que $P_X \in \mathcal{P}(\ell, C_\ell, C_\ell, m)$, et on se retrouve donc bien dans le même cadre que le Corollaire de borne inférieure.

De façon assez surprenante, et ce même si le problème est abordé de façon non-paramétrique avec des hypothèses de régularités, il n'y a en fait pas besoin d'étape de régularisation/lissage des données. Cette observation n'est pas spécifique à notre approche et semble être intrinsèque à l'aspect fonctionnel des données. On retrouve notamment des phénomènes et des bornes similaires pour un autre problème d'Analyse de Données Fonctionnelles défini sur le schéma d'observation (1) dans l'article de Cai et Yuan (2011).

Bibliographie

Belhakem R., Picard F., Rivoirard, V. et Roche A. (2022) Minimax estimation of Functional Principal Components from noisy discretized functional data.

Cai, T.T. et Yuan, M. (2011), Optimal estimation of the mean function based on discretely sampled functional data : Phase transition, *The Annals of Statistics*, 39(5), pp. 2330–2355.

Ferraty F. et Vieu P. (2006), *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York. Theory and practice.

Hsing, T. et Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*.

Jirak M. et Wahl M. (2018), Relative perturbation bounds with applications to empirical covariance operators, *Advances in Mathematics*.

Mas A., Ruymgaart F. (2015), High Dimensional Principal Projections, *Complex Analysis and Operator Theory*, 9, 35-63.

Viviani R., Grön G. et Spitzer M. (2005) Functional principal component analysis of fmri data, *Human Brain Mapping*, 24.

Warmenhoven J., Bargary N., Liebl D., Harrison A., Robinson M.A., Gunning E., Hooker G. (2021), PCA of waveforms and functional PCA: A primer for biomechanics, *Journal of Biomechanics*, Volume 116.