Université Catholique de Louvain (LLN)

Thèse de Hugo Brunet

Estimation non paramétrique de données fonctionnelles avec erreurs fonctionnelles dans les covariables

---

WP1 : Estimation non paramétrique de données fonctionnelles avec erreurs de mesure via la déconvolution de la densité des scores des covariables

---

rédigé par
Hugo Brunet
sous la direction de
Eugen Pircalabelu
Germain Van Bever

14 Mar 2024

Abstract

Lorem ipsum dolor sit amet. Ut expedita sunt est delectus quia ad nostrum delectus eum magni dolor. Eos nemo minima sit deleniti porro et necessitatibus minima ab quia necessitatibus in beatae autem et voluptas labore.
Lorem ipsum dolor sit amet. Ut expedita sunt est delectus quia ad nostrum delectus eum magni dolor. Eos nemo minima sit deleniti porro et necessitatibus minima ab quia necessitatibus in beatae autem et voluptas labore.

contribution

si jamais vous apercevez des fautes dans le polycopié, merci de rédiger une issue sur Github à l'adresse:

correctif

fda-score_density_deconvolution/issues

contact

mail DEV: hugo.brunet@uclouvain.be

Notation	Signification
Données fonctionnelles	
X	Variable Aléatoire Fonctionnelle : $X : (\Omega, \mathscr{F}) \to \left(\mathscr{C}^0(I, \mathbb{R}), \mathscr{C}\right)$
Covariance de X	
C_X	Opérateur de Covariance : $C_X : \begin{array}{ccc} \mathbb{L}^2 & \longrightarrow & \mathbb{L}^2 \\ f & \longmapsto & \int c_X(u, \bullet) f(\bullet) du \end{array}$
c_X	kernel of the Covariance integral operator : $\displaystyle c_X : \begin{array}{ccc} \mathscr{T} \times \mathscr{T} & \longrightarrow & \mathbb{R} \\ (s,t) & \longmapsto & \sum_{k=1}^{r[X]} \lambda_k^{[X]} \phi_k^{[X]}(s) \phi_k^{[X]}(t) \end{array}$
Σ_X	Empirical covariance matrix of the functional random variable X : $\displaystyle \Sigma_X = \left[\operatorname{cov}\left[X(t_{\ell_1}), X(t_{\ell_2})\right]\right]_{1 \le \ell_1, \ell_2 \le L_X}$
Γ_X	Covariance Matrix Estimator: $\displaystyle \Gamma_X = \underset{\substack{\mathbf{S} \in S_L^{++}(\mathbb{R}) \\ \operatorname{rg} \mathbf{S} = \hat{r}_{L_*}[X]}}{\operatorname{argmin}} \left\| P_{L_X}(\delta) \odot \left(\widehat{\Sigma}_W - \mathbf{S}\right) \right\|_F^2$
$r[X]$	rank of C_X
Category B	

Contents

List of Figures

List of Algorithms

1 Introduction

1.1 Notation

The proposed model includes a variety of mathematical objects with distinct nature, but still related to each other. In this section we will introduce all the notations used throughout this paper :

X is a functional random variable, whose covariance operator $C_X : \mathbb{L}^2 \to \mathbb{L}^2 : f \mapsto \int c_X(u, \cdot) f(\cdot) du$ is an integral operator where c_X is the kernel of that integral operator. Sampled data from the real world is discrete though, that's why we consider the empirical covariance matrix $\Sigma_X = \left[\operatorname{cov}\left(X(t_{\ell_1}), X(t_{\ell_2}) \right) \right]_{1 \leq \ell_1, \ell_2 \leq L_X}$ where L_X is the number of points observed on the curve X. We will therefore name L_i the number of points observed on the curve X_i. In our problem, we do not have direct access to the covariates $\left(X_i \right)_{1,N}$ but a contaminated version of them. Hence, we do not have direct access to Σ_X, we must rely on an estimation of Σ_X, it is called Γ_X. Finally we call the rank of the covariance operator for the functional random variable X : $r[X] = \operatorname{rg} C_X$.

note : This is a general theme in the notations used that we specify the functional random variable we are referring to inside brackets : $[X]$ for instance. This becomes especially useful when considering projections of other functional random variables on the PCA basis of **another functional random variable**.

This paper focuses on a score smoothing based approach. In order to have more lightweight notations we will use lowercase letters for the PCA scores : $x_j = \langle\, X - \mathbb{E}X \mid \phi_j^{[X]} \,\rangle_{\mathbb{L}^2}$. Because X is finite rank we can define $x = [\, x_j \,]_{1 \leq j \leq r[X]}$, the vector of $\mathbb{R}^{r[X]}$ of the PCA scores of X. When looking at the projections on the PCA basis of another random variable, we use the following notation for the components of W on the PCA basis of X : $w_j^{[X]} = \langle\, W - \mathbb{E}W \mid \phi_j^{[X]} \,\rangle_{\mathbb{L}^2}$.

In order to keep some consistency in indexes, i (ranging from 1 to N) represents a functional data curve index, ℓ (ranging from 1 to L_i) represents a time observed index, k (ranging from 1 to r_N) represents a PCA component index, and j (ranging from 1 to $r[X]$) represents the additive term index in the additive model used.

K is the base kernel used for the deconvolution problem (epanechnikov for instance), which is then used to build the normalized deconvolution kernel $\widetilde{K}^{\star}$. Each function $f_{k,j}$ is therefore approximated by the SBF-solution using the deconvolution kernel, and thus named $\hat{f}^{\star}_{k,j}$ using the bandwidth $h_{k,j}$.

We observe a total of N covariate curves $\left(X_i \right)_{1,N}$. On each curve X_i we observe L_i, and L_Y points on the response curve Y. The critical number of points needed to be able to perform the retrieval of the eigenfunctions of C_X is called L_* as in [1, Panaretos, 2018]. The support of these curves is called $\mathcal{T}$, and a point on that support will be therefore called t. The interval on which we randomly observe the times $(T_i[\,\ell\,] : 1 \leq i \leq N, 1 \leq \ell \leq L_i)$, $[T_{i,(1)}, T_{i,(L_i)}]$ is noted I_{obs}.

The fourier transform of a real valued function f will be written as $\mathcal{F}[\,f\,]$. The characteristic function of a random variable variable $\mathbf{x}$ will be written $\varphi_{\mathbf{x}}$.

The density of the scores will be writtent as p_x for the full dimensional density of x on $\mathbb{R}^{r[X]}$ w.r.t Lebesgue measure $\lambda_{r[X]}$. The 1-dimensional density of the score x_j on the real line will be written as p_{x_j} and the joint density between the j_1^{th} and j_2^{th} component will be written $p_{x_{j_1}, x_{j_2}}$. The caracteristic function of a random variable x will be written as φ_x.

To make things clearer, dummy variables and fixed values will be set using cyrillic letters such as $ж$ or $ш$.

2 Methodology

2.1 The model

2.2 Estimation of the model

3 Theoretical Properties

3.1

$\mathbf{W}(\cdot) = \mathbf{X}(\cdot) + \mathbf{U}(\cdot)$

3.2 Optimal smoothing kernel bandwidth for the functional error deconvolution problem

3.3 Special case : $\| \mathbf{U} \|_{\mathbb{L}^2} \xrightarrow[\mathrm{N}\to\infty]{} \mathbf{0}$

4 Simulations & numerical study

A

B

References

[1] Anirvan Chakraborty and Victor M Panaretos. Regression with genuinely functional errors-in-covariates. arXiv preprint arXiv:1712.04290, 2017.

[2] Marie-Hélène Descary and Victor M Panaretos. Functional data analysis by matrix completion. 2019.

[3] Kyunghee Han and Byeong U Park. Smooth backfitting for errors-in-variables additive models. The Annals of Statistics, 46(5):2216–2250, 2018.

[4] Tailen Hsing and Randall Eubank. Theoretical foundations of functional data analysis, with an introduction to linear operators, volume 997. John Wiley & Sons, 2015.

[5] Sunny Wang Guang Wei, Valentin Patilea, and Nicolas Klutchnikoff. Adaptive functional principal components analysis. arXiv preprint arXiv:2306.16091, 2023.