Contribute 1
Multivariate functional data depth measure based on variance-covariance operators

Rachele Biasi, Francesca Ieva, Anna Maria Paganoni, Nicholas Tarabelloni

Abstract We introduce a generalization of the simplicial depth measure to multivariate functional data, exploiting the role of the variance-covariance operators in weighting the components that define the depth. We propose the use of this nonparametric method for supervised classification purpose.

Introduction

In recent times, more and more data coming from various real-life applications can be studied with the paradigm of functional data. This framework requires the development and use of ad-hoc statistical instruments (for an introduction, see [1]). Here we consider multivariate functional data, that means data where each observation is a set of possibly correlated functions. These functions can be viewed as trajectories of stochastic processes defined on a given infinite dimensional functional space, and characterized by a variance covariance operator (see the monograph [3] for more theoretical details). We consider the notion of depth measure for functional data introduced in [7, 8] for univariate functional data, and extended to multivariate functional framework in [6]. Depth measure belongs to the field of non-parametric functional data analysis (see, for instance, [2]) and seems to be promising in classifying curves belonging to different generating processes. According to definition in [6], it is necessary to make a choice of the weights averaging the contribution of each component of the multivariate

Rachele Biasi
Politecnico di Milano, Italy, e-mail: rachele.biasi@mail.polimi.it

Francesca Ieva
Università degli Studi di Milano, Italy, e-mail: francesca.ieva@unimi.it

Anna Maria Paganoni
Politecnico di Milano, Italy, e-mail: anna.paganoni@polimi.it

Nicholas Tarabelloni
Politecnico di Milano, Italy, e-mail: nicholas.tarabelloni@polimi.it

signal to the depth itself. This choice is usually problem-driven, and in general no gold rules have been given so far. We develop a general method for defining such weights. In particular, we propose to choose them taking into account the distance between the estimated covariance operators of the two groups. In fact, the covariance structure of the multivariate functional signals contains information about the reciprocal role of the signal components with respect one to each other. This should be taken into account in measuring the depth of a signal, and in general when comparing signal features with reference traces. In the following, we consider many different distances between covariance operators in the infinite dimensional setting, as discussed in [9], and by means of a simulation study we discuss the behavior and the robustness of the weight choice with respect to different distances and as long as the correlation between components changes. In [4] a wider study of this problem as well as an application to a case study for disease probability prediction from electrocardiographic signals are presented. All the analyses are carried out using `R` statistical software [10] and the ad-hoc C++/MPI parallel library for computational statistics `HPCS` (for further details see the website: https://github.com/ntarabelloni/HPCS, Code is available upon request) [5].

1.1 Multivariate depth measure with weights based on variance covariance operators

Let **X** be stochastic process taking values in the space $\mathcal{C}(I;\mathbb{R}^h)$ of continuous functions f $= (f_1, ..., f_h) : I \to \mathbb{R}^h$, where I is a compact interval of $\mathbb{R}$. The multivariate depth measure is defined as

$$MBD_n^J(\mathbf{f}) = \sum_{k=1}^{h} p_k MBD_{n,k}^J(f_k), \quad p_k > 0 \; \forall \; k = 1, ..., h, \quad \sum_{k=1}^{h} p_k = 1 \tag{1.1}$$

where for each function $f_k \in F \subset \mathcal{C}(I;\mathbb{R})$, $k = 1, ..., h$, the $MBD_{n,k}^J(f_k)$ measures the proportion of time interval I where the graph of f_k belongs to the envelopes of the j-tuples $(f_{i_1;k}, ..., f_{i_j;k})$, $j = 1, \ldots, J$, extracted from F. In other words, measuring that the curve f_k is in the band determined by the j curves $(f_{i_1;k}, ..., f_{i_j;k})$, means computing

$$MBD_{n,k}^J(f_k) = \sum_{j=2}^{J} \binom{n}{j}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_j \le n} \tilde{\lambda}\{E(f_k; f_{i_1;k}, ..., f_{i_j;k})\},$$

where $E(f_k; f_{i_1;k}, ..., f_{i_j;k}) = \{t \in I, \min_{r=i_1,...,i_j} f_{r;k}(t) \le f_k(t) \le \max_{r=i_1,...,i_j} f_{r;k}(t)\}$ and $\tilde{\lambda}(A) = \lambda(A)/\lambda(I)$, $\forall A \subseteq I$, with λ the Lebesgue measure on I. Statistical properties of the depth measure defined in (1.1) as well as inferential tools based on this concept are detailed in [6]. For subsequent analyses we fixed $J = 2$ in (1.1) once and for all.

Let us consider two different multivariate stochastic processes $\mathbf{X}$ and $\mathbf{Y}$. Without loss of generality we assume that their means are equal to zero. We indicate their covariance operators with $\mathcal{V}_{\mathbf{X}}$ and $\mathcal{V}_{\mathbf{Y}}$. The kernel of the covariance operator of $\mathbf{X}$, $V_{\mathbf{X}}(s,t)$, is defined by

$$V_{\mathbf{X}}(s,t) = \mathbb{E}[\mathbf{X}(s) \otimes \mathbf{X}(t)], \quad s,t \in I$$

where $\otimes$ is an outer product in $\mathbb{R}^h$. So $V_{\mathbf{X}}(s,t)$ is a $h \times h$ matrix, whose elements will be denoted as $V_{\mathbf{X}}^{kq}(s,t)$, for $k,q = 1,...,h$. Analogously for $V_{\mathbf{Y}}(s,t)$.
We consider the distances introduced in [9], generalizing them to the case of non necessarily positive definite operators, in fact we are interested in quantifying the distance also between the inter-component blocks, i.e. $V_{\mathbf{X}}^{kq}(s,t)$ and $V_{\mathbf{Y}}^{kq}(s,t)$ when $k \neq q$. Let $d(V,W)$ denote a distance between two operators. We compute for each $k = 1,\dots,h$ the quantity $d_k = \sum_{q=1}^{h} d(V_{\mathbf{X}}^{kq}(s,t), V_{\mathbf{Y}}^{kq}(s,t))$, considering the following distances:

- – L^2 **distance**

$$d_L(V,W) = \sqrt{\int_I \int_I (v(s,t) - w(s,t))^2 ds dt}, \tag{1.2}$$

where $v(s,t)$ and $w(s,t)$ are the kernels of the operators V and W respectively.

- – **Spectral distance**

$$d_S(V,W) = |\lambda_1|, \tag{1.3}$$

where $|\lambda_1|$ is the maximum eigenvalue of the difference operator $V - W$.

- – **Square root pseudo distance**

$$d_R(V,W) = \| \, |V|^{\frac{1}{2}} - |W|^{\frac{1}{2}} \, \|_{HS}, \tag{1.4}$$

where the Hibert-Schmidt norm of an Hilbert-Schmidt compact operator T is $\|T\|_{HS} = \sqrt{\text{trace} T^* T}$, T^* is the adjoint operator of T, $|T|^{\frac{1}{2}}$ is such that $|T|^{\frac{1}{2}} v_k = |\lambda_k|^{\frac{1}{2}} v_k$, $\{v_k\}_k$ is the orthonormal basis of L^2 of the eigenfunctions of T and $\{\lambda_k\}_k$ is the sequence of the related eigenvalues.

- – **Frobenius distance**

$$d_F(V,W) = \|V - W\|_{HS} = \sqrt{\text{trace}(V-W)^*(V-W)}. \tag{1.5}$$

- – **Procrustes pseudo distance**

$$d_P(V,W) = d_P(|V|,|W|) = \inf_{R \in O(L^2(I))} \|L_1 - L_2 R\|_{HS}, \tag{1.6}$$

where $O(L^2(I))$ is the space of all unitary operators on $L^2(I)$ and L_1 and L_2 are such that $V = L_1 L_1^*$ and $W = L_2 L_2^*$.

Let us note that in the case of square root and Procrustes we deal with pseudo distances since $d(V,W) = 0$ if and only if $|V| = |W|$. Based on the previous definitions, we then propose the following choice for the weights in the multivariate functional depth defined in (1.1):

$$p_k = \frac{d_k}{\sum_{l=1}^{h} d_l}, \qquad for\ k = 1, ..., h. \tag{1.7}$$

1.2 Simulation study: effect on weights of different distances and correlations between components

Without loss of generality we consider the case of bivariate functional data, i.e., $h = 2$. The time interval I is sampled over an evenly spaced grid of 50 points. The data are generated according to the following stochastic processes

$$\mathbf{X}\left(\mathbf{t}\right) \sim N\left(\mathbf{0}, S_1\right) \qquad \mathbf{Y}\left(\mathbf{t}\right) \sim N\left(\mathbf{0}, S_2\right)$$

for the first and the second population respectively. The structure of S_i, $i = 1, 2$, is the following:

$$S_i = \begin{pmatrix} A_i & C_i \\ C_i^T & B_i \end{pmatrix} \tag{1.8}$$

being

$$A_1 = \begin{pmatrix} 10^6 & 10^2 & & & \\ 10^2 & 10^6 & 10^2 & & \\ & \ddots & \ddots & \ddots & \\ & & 10^2 & 10^6 & 10^2 \\ & & & 10^2 & 10^6 \end{pmatrix}, B_1 = \begin{pmatrix} 10^5 & 1 & & & \\ 1 & 10^5 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 10^5 & 1 \\ & & & 1 & 10^5 \end{pmatrix}, \tag{1.9}$$

$$A_2 = \begin{pmatrix} 10^4 & 1 & & & \\ 1 & 10^4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ ferraty_{v}ieu & & 1 & 10^4 & 1 \\ & & & 1 & 10^4 \end{pmatrix}, B_2 = \begin{pmatrix} 10^3 & 10^2 & & & \\ 10^2 & 10^3 & 10^2 & & \\ & \ddots & \ddots & \ddots & \\ & & 10^2 & 10^3 & 10^2 \\ & & & 10^2 & 10^3 \end{pmatrix}, \tag{1.10}$$

$C_i = \rho(A_i B_i)^{1/2}$, and $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$.

It is worth noting that, for all the distances, the higher is the correlation ρ between components, the more balanced are the weights of the single components: if two components are strongly correlated, then their weights tend to be more balanced. This property is not only reasonable, but also shows a desirable flexibility in the behavior of such weights, which seem to be sensitive to the particular structure of the variance-covariance operators of the models at hand. For this reason taking into account not only the distances between intra-component variability, but also the inter-component ones is relevant for the weights choice, whichever distance we consider.

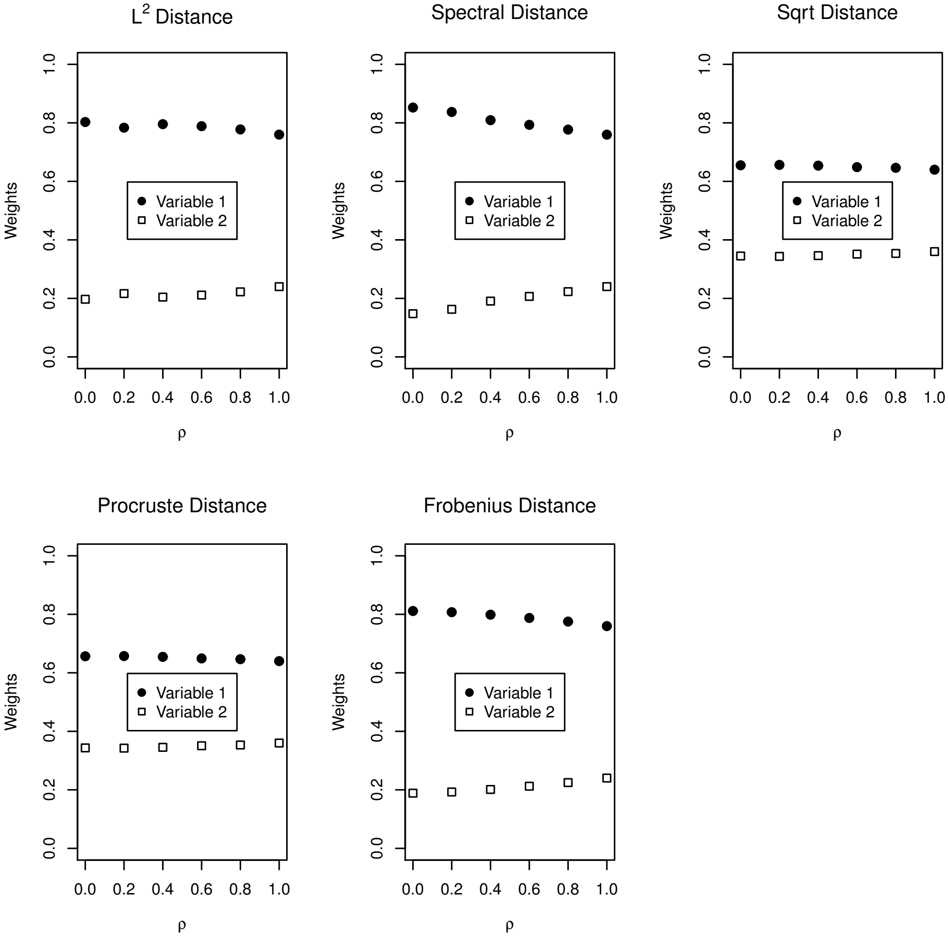

Figure 1.1: Weights of MBD in the simulation study, determined through the distances between variance-covariance operators previously defined.

Bibliography

[1] Silverman, B.W., Ramsay, J.O. (2005), Functional Data Analysis, *Springer*.

[2] Ferraty, F., Vieu, P. (2006), Nonparametric functional data analysis: theory and practice, *Springer*.

[3] Horváth, L., Kokoszka, P. (2012), Inference for functional data with applications, *Springer*.

[4] Biasi, R., Ieva, F., Paganoni, A.M., Tarabelloni, N. (2013), Use of depth measure for multivariate functional data in disease prediction: an application to electrocardiographic signals *Tech. Rep. MOX, Math. Dept.*, Politecnico di Milano. [Online] http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/54-2013.pdf

[5] Tarabelloni, N. (2013) Tools for computational statistics coded in C++, [online] https://github.com/ntarabelloni/HPCS

[6] Ieva, F., Paganoni, A.M. (2013a), Depth Measures for Multivariate Functional Data, *Communication in Statistics - Theory and Methods*, **42**, 7, 1265–1276.

[7] Lopez-Pintado, S., Romo, J. (2007), Depth-based inference for functional data, *Computational Statistics & Data Analysis*, **51**, 10, 4957–4968.

[8] Lopez-Pintado, S., Romo, J. (2009), On the Concept of Depth for Functional Data, *Journal of the American Statistical Association*, **104**, 486, 718–734.

[9] Pigoli, D., Aston, J.A.D., Dryden, I.L., Secchi, P. (2012), Distances and Inference for Covariance Functions *Tech. Rep. MOX, Math. Dept.*, Politecnico di Milano. [Online] http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/35-2012.pdf

[10] R Development Core Team (2009), R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. [online] http://www.R-project.org