Depth measures for multivariate functional data with data-driven weights

Misure di profondità per dati funzionali multivariati con pesi indotti dai dati

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

Abstract The notion of statistical depth have recently been extended to the case of multivariate functional data. Its definition involves the choice of proper weights, averaging the univariate functional depths of each component. The choice of weights is crucial and must be carefully done according to the problem at hand. We describe a procedure that, starting from data, allows to compute a set of weights which are suitable for classification based on depths. These weights incorporate information on distances between covariance operators of the sub-populations. We show the validity of our strategy through a case study in which we perform supervised classification on ECG traces referring to both physiological and pathological subjects.

Abstract *Recentemente la nozione di misura di profondità statistica è stata estesa al caso di dati funzionali multivariati. La definizione richiede la scelta di un insieme di pesi con cui mediare le profondità univariate delle componenti. Nel seguito descriviamo una procedura che, a partire dai dati, permette di calcolare un insieme di pesi opportuni per una classificazione basata sulle profondità. Questi pesi incorporano informazioni sulla distanza degli operatori di covarianza delle sottopopolazioni del dataset. L'approccio viene validato attraverso un'applicazione alla classificazione supervisionata di segnali ECG, sia riferiti ad uno stato fisiologico che patologico.*

Rachele Biasi
Politecnico di Milano, P.za Leonardo da Vinci 32, 20133 Milano (Italy)
e-mail: rachele.biasi@mail.polimi.it

Francesca Ieva
Università degli Studi di Milano, via Saldini 50, 20133 Milano (Italy)
e-mail: francesca.ieva@unimi.it

Anna Maria Paganoni
Politecnico di Milano, P.za Leonardo da Vinci 32, 20133 Milano (Italy)
e-mail: anna.paganoni@polimi.it

Nicholas Tarabelloni
Politecnico di Milano, P.za Leonardo da Vinci 32, 20133 Milano (Italy)
e-mail: nicholas.tarabelloni@polimi.it

Key words: Functional data analysis, multivariate functional depths, covariance operator, operators distance.

1 Introduction

In recent years the model of functional data has been increasingly applied to the description of processes occurring within a broad set of fields, such as medicine, biology and engineering (see, e.g., the monograph [7]). Due to the difficulty, in practice, of assessing distributional hypotheses regarding functional data, a leading role is played by nonparametric techniques. In this paper we focus on depth measures, a classical tool of nonparametric multivariate analysis, expressing the centrality versus outlyingness of an element with respect to the entire sample, thus inducing a global ordering on the dataset. Starting from original definitions for multivariate data, in [4, 5] authors have built an analogous definition for the univariate functional case. A possible extension to multivariate random functions has been proposed in [2], where the overall depth of the multivariate functional variable is given by a weighted mean of the univariate depths in each dimension. Indeed, weights must be chosen accordingly to some prior knowledge on the problem at hand, and no golden rule is available so far. In this paper we address the problem of determining a set of weights from which to construct depth measures that are suitable for classification purposes. In particular, we will show an expression for the weights which directly involves the distances between covariance operators of the two populations, thus is completely data-driven.
We will apply this procedure to the supervised classification, through logistic regression on multivariate functional depths, of a dataset composed of ECG traces both physiological and pathological, and show that considerable results can be obtained in terms of classification power.

In Section 2 we recall the definition of multivariate functional depth. In Section 3 we describe the construction of weights starting from several possible distances between covariance operators. In Section 4 we show the application to ECG signals.

2 Multivariate depth measures

We recall the definition of depth for multivariate functional data as it is introduced in [2]. Let $\mathbf{X}$ be a stochastic process taking values in the space $\mathscr{C}(I;\mathbb{R}^h)$ of continuous functions $\mathbf{f} = (f_1,...,f_h) : I \to \mathbb{R}^h$, where I is a compact interval of $\mathbb{R}$. The multivariate depth measure is defined as:

$$\mathrm{MBD}_n^J(\mathbf{f}) = \sum_{k=1}^{h} p_k \, \mathrm{MBD}_{n,k}^J(f_k), \quad p_k > 0 \quad \forall\, k = 1,...,h, \quad \sum_{k=1}^{h} p_k = 1 \qquad (1)$$

where for each function $f_k \in F \subset \mathscr{C}(I;\mathbb{R})$, $k = 1,...,h$, and $\mathrm{MBD}^J_{n,k}(f_k)$, the (modified) univariate functional depth as it is defined in [4, 5], measures the proportion of 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, i.e:

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

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 $\tilde{\lambda}$ is the Lebesgue measure normalized with respect to $\lambda(I)$. Statistical properties of the depth measure defined in (1) as well as inferential tools based on this concept are detailed in [2]. In the following we fix $J = 2$ in (1), to have a lighter computational burden. This is motivated by the robustness of ranks induced by depths, which is assessed in a study conducted in [1], thanks to an efficient implementation within a parallel environment.

3 Weighting strategy for multivariate functional depths

The set of weights $\{p_k\}$ defining the depths in (1) must be chosen carefully depending on the application at hand. We recall that our purpose is to perform supervised classification of signals belonging to two populations with different covariance operators. We focus on one population, and think it is generated according to $\mathbf{X}$, a stochastic process with law $P_{\mathbf{X}}$ taking values on the space $L^2(I;\mathbb{R}^h)$ of square integrable functions. Let $\mu_l(t) = \mathbb{E}[X_l(t)]$, for each $t \in I$, denote the mean function of the $l-$component $X_l(t)$, for $1 \le l \le h$, then $\mu_{\mathbf{X}}(t) := \mathbb{E}[\mathbf{X}(t)]$ is the mean function of $\mathbf{X}$. The covariance operator $\mathscr{V}_{\mathbf{X}}$ of $\mathbf{X}$ is a linear, compact operator from $L^2(I;\mathbb{R}^h)$ to $L^2(I;\mathbb{R}^h)$ acting on a function $\mathbf{g}$ as follows:

$$(\mathscr{V}_{\mathbf{X}}\mathbf{g})(s) = \int_I V_{\mathbf{X}}(s,t)\mathbf{g}(t)dt, \tag{2}$$

where the kernel $V_{\mathbf{X}}(s,t)$ is defined by

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

with $\otimes$ an outer product in $\mathbb{R}^h$. $V_{\mathbf{X}}(s,t)$ is a $h \times h$ matrix, whose elements will be denoted as $V^{kq}_{\mathbf{X}}(s,t)$, for $k,q = 1,...,h$.
Populations will correspond to two different stochastic processes, say $\mathbf{X}$ and $\mathbf{Y}$, with covariance operators $\mathscr{V}_{\mathbf{X}}$ and $\mathscr{V}_{\mathbf{Y}}$ respectively. In order to enhance classification, we aim at building weights incorporating information on the difference in covariance operators, i.e. distances between blocks of $\mathscr{V}_{\mathbf{X}}$ and $\mathscr{V}_{\mathbf{Y}}$. Thus, let us denote by $d(V,W)$ a (pseudo)-distance between two operators V and W. We define for each $k = 1,\ldots,h$ the quantity $d_k = \sum_{q=1}^{h} d(V^{kq}_{\mathbf{X}}(s,t), V^{kq}_{\mathbf{Y}}(s,t))$.

Then, our proposal for the weights is

$$p_k = \frac{d_k}{\sum_{j=1}^{h} d_j}. \tag{3}$$

We explicitly remark that, through this definition of weights, we are taking into account not only the distances between intra-component covariance operators, but also the distance between inter-component ones.
We chose the distance d among those proposed in [6], after proper extensions to semi-definite (i.e. not only positive-definite) operators, as blocks $V_{\mathbf{X}}^{k,q}$ with $k \neq q$ are:

- L^2 distance

$$d_L(V,W) = \sqrt{\int_I \int_I (v(s,t) - w(s,t))^2 ds dt}, \tag{4}$$

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

$$d_S(V,W) = |\lambda_1|, \tag{5}$$

  where $|\lambda_1|$ is the first eigenvalue of the difference operator $V - W$.
- Frobenius distance

$$d_F(V,W) = \|V - W\|_F = \sqrt{\text{trace}(V-W)^*(V-W)}. \tag{6}$$

  where T^* indicates the adjoint of T.
- Square root pseudo distance

$$d_R(V,W) = \| \, |V|^{\frac{1}{2}} - |W|^{\frac{1}{2}} \, \|_F, \tag{7}$$

  $|T|^{\frac{1}{2}}$ is such that $|T|^{\frac{1}{2}} v_k = |\lambda_k|^{\frac{1}{2}} v_k$; $\{v_k\}_k$ and $\{\lambda_k\}_k$ are the sequences of eigenfunctions and eigenvalues of T.
- Procrustes pseudo distance

$$d_P(V,W) = d_P(|V|,|W|) = \inf_{R \in O(L^2(I))} \|L_1 - L_2 R\|_F, \tag{8}$$

  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 $W = L_2 L_2^*$.

4 Application to ECG signals

In this section we apply the procedure previously described to the computation of MBDs (with distance-driven weights) of a dataset of ECG signals, and perform a

supervised classification via logistic regression, in order to distinguish the healthy from the pathological traces (we considered the pathology of Left Bundle Branch Block, LBBB). As basic statistical unit, we consider the 8-variate ECG record defined by leads I, II, V1, V2, V3, V4, V5 and V6. We model the label as a Bernoulli random variable Y_i, taking value 1 if LBBB is diagnosed, and 0 otherwise.
We analyse ECG traces from PROMETEO (PROgetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall'Extra Ospedaliero) database, containing also an automatic diagnosis, established by the commercial Mortara-Rangoni VERITAS^{TM} algorithm, which we want to reproduce.
The dataset is constituted of ECG traces of $n = 149$ subjects, among which 101 are physiological and 48 are affected by LBBB. Before the analysis, data have been denoised and registered through landmarks, in order to separate amplitude variability from phase's one (for further details on this procedure, see [3]).
To compute the MBDs, we randomly chose 50 ECGs from the physiological traces to be used as a reference group, then we compute the ranks of the remaining 51 physiological and 48 LBBB traces with respect to them. The procedure has been repeated 20 times to avoid bias selection in the choice of the reference group.
We performed the analyses on our case study considering all the (pseudo) distances introduced in Section 3. The results are quite robust with respect to the choice of distance, and we will present the results obtained with the Procrustes pseudo-distance, which is the best performing (we report in Tab. 1 the weights computed with this choice of distance).

Lead	V2	V3	V1	V4	V5	V6	I	II
Weights	0.1722	0.1607	0.1385	0.1357	0.1132	0.1104	0.0872	0.0821

Table 1 Weights induced by the Procrustes pseudo-distance, to be inserted in (1).

We carry out a Wilcoxon rank sum test on MBDs, to assess their ability to express the difference between the two populations of samples. The p-value over all the 20 cases is always less or equal to $3.02 * 10^{-12}$, thus supporting the belief that evidence for the difference among the two population exists and is significant.
We use the MBDs within a logistic regression model for the prediction of the label variable $Y_i \sim \text{Be}(p_i)$:

$$\theta_i = \beta_0 + \beta_1 MBD_i, \quad \theta_i = \text{logit}(p_i), \quad \forall i = 1, \dots, n. \tag{9}$$

The predictors have great statistical significance, since their p-values are both less than 10^{-6}. The confusion matrix obtained comparing the true and the estimated labels of the patients is reported in Table 3. We set the threshold for the classification carried out by the logistic model equal to 0.5.
At the end of the 20 runs of our analysis we obtain a sensitivity (mean $\pm$ standard deviation) of $(84.48 \pm 2.29)\,\%$, specificity of $(89.80 \pm 1.87)\,\%$ and a correct

Parameter	Estimate	Std. Error	p-value
β^0 (Intercept)	11.484	2.483	$3.75*10^{-06}$
β^1 (MBD)	-46.268	9.619	$1.51*10^{-06}$

Table 2 Estimates, standard errors and p-values for the parameters of the logistic model 9.

	Normal	LBBB
Classified as Normal	47	8
Classified as LBBB	4	40

Table 3 Confusion matrix for the classification via logistic model based on MBDs.

classification rate of $(87.22 \pm 1.58)\,\%$. We conclude that considering the distances between covariance operators could be a good method to produce multivariate functional depths that are able to emphasize, within classification procedures, differences in populations characterized by different covariance operators.

5 Acknowledgements

This work is part of PROMETEO (PROgetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall'Extra Ospedaliero). Data are provided by Mortara Rangoni Europe s.r.l.. The authors wish to thank 118 Dispatch Centre of Milano.

References

1. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer (2005)
2. Ieva, F., Paganoni, A.M.: Depth Measures for Multivariate Functional Data. Communications in Statistics. **42**, 7, 1265–1276 (2013)
3. López-Pintado, S., Romo, J.: Depth-based inference for functional data. Computational Statistics and Data Analysis. **51**, 10, 4957–4968 (2006)
4. López-Pintado, S., Romo, J.: On the Concept of Depth for Functional Data. Journal of the American Statistical Association. **104**, 486, 718–734 (2009)
5. Pigoli, D., Aston, J.A.D., Dryden, I.L., Secchi, P.: Distances and Inference for Covariance Functions. Biometrika (To appear)
6. Ieva, F., Paganoni, A.M., Pigoli, D., Vitelli, V.: Multivariate functional clustering for the morphological analysis of electrocardiograph curves. Journal of the Royal Statistical Society: Series C (Applied Statistics). **62**, 3, 401–418 (2013)
7. Biasi, R., Ieva, F., Paganoni, A.M., Tarabelloni, N.: An efficient framework for computational functional data analysis. MOX - Math. Dept. Politecnico di Milano. (Work in progress)