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

Use of Depth Measure for Multivariate Functional Data in Disease Prediction: An Application to Electrocardiograph Signals

DOI 10.1515/ijb-2014-0041

Abstract: In this paper we develop statistical methods to compare two independent samples of multivariate functional data that differ in terms of covariance operators. In particular we generalize the concept of depth measure to this kind of data, exploiting the role of the covariance operators in weighting the components that define the depth. Two simulation studies are carried out to validate the robustness of the proposed methods and to test their effectiveness in some settings of interest. We present an application to Electrocardiographic (ECG) signals aimed at comparing physiological subjects and patients affected by Left Bundle Branch Block. The proposed depth measures computed on data are then used to perform a nonparametric comparison test among these two populations. They are also introduced into a generalized regression model aimed at classifying the ECG signals.

Keywords: depth measures, multivariate functional data, covariance operators, ECG signals, generalized linear models

1 Introduction

In many real-life applications of statistics nowadays, data are becoming more and more complex. This is particularly true in the biomedical context, where data produced by medical devices are signals, functions of vital parameters, images or even a combination of these. This drives statistical research towards the identification of suitable models and inferential techniques for handling the complexity of such data.

This paper is mainly focused on supervised learning from multivariate functional data. By multivariate functional data we mean 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, as it has been proposed in Ieva and Paganoni [1] and as it will be detailed in Section 2. In particular the motivating aim is the analysis of the 8-leads Electrocardiographic (ECG) traces of patients whose pre-hospital ECG has been sent to 118 Dispatch Center of Milan (the Italian free-toll number for emergencies) by life support personnel of the basic rescue units. ECG signals can be inherently considered as multivariate functional data with correlated components. In fact, each component describes the same biological event, i.e., the representative heartbeat of a patient (see Ieva et al. [2] for a deeper explanation of this).

In this work we aim at modeling the binary outcome representing the presence of cardiovascular ischaemic event in order to estimate the probability of each patient to be affected by Acute Myocardial Infarction. Beyond the application of interest, we develop a general framework to model a binary outcome by means of proper summaries of multivariate functional data addressing the problem of classification and prediction.

In Ieva and Paganoni [3] a similar problem has been faced, mainly performing a data dimensionality reduction by a Multivariate Functional Principal Component Analysis (see Ramsay and Silverman [4] and Berrendero et al. [5]). It consists of summarizing the information contained in covariance operators of the signals and their first derivatives by the corresponding scores. Scores are obtained projecting data and derivatives on the related Karhunen-Loève bases. In this work we summarize relevant features of data

***Corresponding author: Francesca Ieva,** Department of Mathematics, Università degli Studi di Milano, Milano, Italy, E-mail: francesca.ieva@unimi.it
Nicholas Tarabelloni, Rachele Biasi, Anna Maria Paganoni, Department of Mathematics, Politecnico di Milano, Milano, Italy

through some non parametric statistical objects, i.e., depth measures. Depth measures are a classic tool from multivariate statistics, and have recently been extended to functional data, originally in order to enable functional outlier detection. Depth measures allow to order and rank high-dimensional data, thus besides their use in robust statistics, they are a promising tool to develop new methods involving objects, like functional data, for which inferential methods are lacking.

We exploit the definition of depth measure introduced in Ieva and Paganoni [1], which is a generalization to the multivariate functional case of the univariate depths proposed in Lopez-Pintado and Romo [6, 7]. In order to compute the depth measure proposed in Ieva and Paganoni [1], it is necessary to make a choice of the weights averaging the contribution of each component of the multivariate signal to the depth itself. This choice is usually problem-driven, and in general no gold rules have been given so far. In this paper we develop a 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 identified by the binary outcome. In fact, the covariance structure of the multivariate functional signals (and possibly of the corresponding derivatives) contains information about the reciprocal role of the signal (derivative) components with respect one to one another. This should be taken into account in measuring the depth of a signal (derivative), and in general when comparing signal (derivative) features with reference traces. In fact, it may drive the weights definition giving emphasis to data components according to the way they are correlated one to each other. In the following, in order to measure the distance between variance-covariance operators we consider many different definitions of distances in the infinite dimensional setting, according to what discussed in Pigoli et al. [8]. By choosing the weights of multivariate depths, exploiting information about the processes generating data, we mean to enhance the depth-based inference regarding the binary outcome under study.

The paper is structured as follows: in Section 2, the definition of multivariate functional depth measure is presented, the choice of the weights is discussed and the use of depths to perform a functional Wilcoxon's sum rank test is explained. In Section 3, a simulation study to support a choice for the reduction of complexity in the computation of depths is detailed, together with a study concerning the behaviour of the proposed depth measure in presence of data with only location shift (equal covariance). Section 4 concerns the analysis of ECG data arising from PROMETEO dataset through a generalized regression model based on depths. Finally, in Section 5, conclusions are drawn and further developments are proposed. All the analyses are carried out using R statistical software (see R Core Team [9]) and the ad-hoc $C++$/MPI parallel library for computational statistics HPCS[1] (see Tarabelloni [10]).

2 Multivariate depth measures with covariance driven weights

Let us start by recalling the definition of band depth for multivariate functional data introduced in Ieva and Paganoni [1]. This definition has been introduced to generalize the concept of band depth for univariate functional data introduced in Lopez-Pintado and Romo [6, 7] to the multivariate framework.

Let $\mathbf{X}$ be a stochastic process taking values in the space $L^2(I;\mathbb{R}^h)$ of square integrable functions $\mathbf{f} = (f_1, ..., f_h) : I \to \mathbb{R}^h$, where I is a compact interval of $\mathbb{R}$. We imagine to have a dataset F_n constituted of $n \in \mathbb{N}$ sample observations of this process, namely $\mathbf{f}_1, \ldots, \mathbf{f}_n$. F_n^k is the n-tuple made by the k-th components of the functions included in the reference sample. According to Definition 2 in Lopez-Pintado and Romo [7], given a generic function $g \in L^2(I;\mathbb{R})$, we introduce its modified band depth with respect to F_n^k:

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

1 Code is available upon request, for further details see the website: https://github.com/ntarabelloni/HPCS.

where $E(g) := E(g; f_{i_1;k}, ..., f_{i_j;k}) = \{t \in I, \min_{r=i_1,...,i_j} f_{r;k}(t) \leq g(t) \leq \max_{r=i_1,...,i_j} f_{r;k}(t)\}$ and $\tilde{\lambda}(g) = \lambda(E(g))/\lambda(I)$ being $\lambda(\cdot)$ the Lebesgue measure on I. For each function $g \in L^2(I;\mathbb{R})$, the quantity $MBD^J_{n,k}(g)$ measures the proportion of I where the graph of g belongs to the envelopes of the j-tuples, $(f_{i_1;k}, ..., f_{i_j;k}), j = 2, \ldots, J$, sampled from F^k_n.

The multivariate depth measure of a vector of functions $\mathbf{f} = (f_1, ..., f_h)$ with respect to F_n is defined in Ieva and Paganoni [1] as

$$MBD^J_n(\mathbf{f}) = \sum_{k=1}^{h} p_k MBD^J_{n,k}(f_k), \quad p_k > 0 \ \forall\ k = 1, ..., h, \quad \sum_{k=1}^{h} p_k = 1. \tag{2}$$

Statistical properties of the depth measure defined in (2) as well as inferential tools based on this concept are detailed in Ieva and Paganoni [1].

An open problem in eq. (2) is how to choose the weights $p_1, ..., p_h$. In general this choice is problem driven. Nevertheless, whenever the principal aim of the analysis is the comparison between two different populations, it is possible to exploit the difference between the variance-covariance structures of the two groups. In fact, it is reasonable to expect that they point out information about differences between the two populations.

In the following we consider the depth measure defined in eq. (2) with $J = 2$. As it will be detailed in Section 3, this choice can be done without loss of generality, since we tested the stability of the ranking induced by different choices of J.

The reason why we choose definition (2), among the range of possible generalizations of depths, is the following: given that the depth of a multivariate function should depend on the behaviour of each component, and therefore its centrality/outlyingness is a global property of the function as a whole, a natural possibility to generalise depths to multivariate functions could consider multi-dimensional versions of function envelopes, i.e. simplexes in $L^2(I;\mathbb{R}^h)$, thus producing a functional version of the well-known finite dimensional simplicial depth (see e.g. Lopez-Pintadoet al. [11]). Yet, according to this strategy, for a dataset of n h-dimensional multivariate functions, in order to compute the depth of a target function $\mathbf{g} \in L^2(I;\mathbb{R}^h)$, one should consider n choose $(h+1)$ subsets of data and build all the corresponding simplexes, then repeat this task for each time instant of the discrete grid.

For common values of n, time grid size, and in case of ECG signals – which will constitute our application further on -, this is absolutely not feasible. As an example, by considering $n = 50$ reference signals, 700 time measurements and $h = 8$ leads, we should compute $\sim 1.75 * 10^{12}$ 8-dimensional polytopes and check for the inclusion of the corresponding value of $\mathbf{g}$ in each one of them, resulting in a forbidding computational complexity. It is then clear that the crucial drawback in this application is the combinatorial complexity with respect to h. On the contrary, the complexity of definition (2) is linear in the number of leads, which makes inference possible in our applicative case, and only a negligible extra effort is required to compute the weights based on covariance structures information.

2.1 On the choice of depths' weights

In order to define the p_ks, and to properly account for the information about the correlation among components in the dataset, we make use of distances defined on variance-covariance operators. We focus on a stochastic process $\mathbf{X}$ 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 $1 \leq l \leq h$, then

$$\boldsymbol{\mu}_{\mathbf{X}}(t) := (\mu_1(t), \ldots, \mu_h(t))^T = \mathbb{E}[\mathbf{X}(t)]$$

is the mean function of $\mathbf{X}$. The covariance operator $\mathcal{V}_X$ of $\mathbf{X}$ is a linear compact integral operator from $L^2(I;\mathbb{R}^h)$ to $L^2(I;\mathbb{R}^h)$ acting on a function $\mathbf{g}$ as follows:

$$(\mathcal{V}_{\mathbf{X}}\mathbf{g})(s) = \int_I V_{\mathbf{X}}(s,t)\mathbf{g}(t)dt, \tag{3}$$

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

$$V_{\mathbf{X}}(s,t) = \mathbb{E}[(\mathbf{X}(s) - \boldsymbol{\mu}_{\mathbf{X}}(s)) \otimes (\mathbf{X}(t) - \boldsymbol{\mu}_{\mathbf{X}}(t))], \quad s,t \in I \tag{4}$$

where $\otimes$ is an outer product in $\mathbb{R}^h$. For s,t fixed, $V_{\mathbf{X}}(s,t)$ is a $h \times h$ matrix, whose elements will be denoted as $V_{\mathbf{X}}^{kq}(s,t)$, for $k,q = 1,\ldots,h$. For k,q fixed, $V_{\mathbf{X}}^{kq}(s,t)$ is a covariance operator.

In what follows, we deal with two different stochastic processes $\mathbf{X}$ and $\mathbf{Y}$ with covariance operators $\mathcal{V}_{\mathbf{X}}$ and $\mathcal{V}_{\mathbf{Y}}$, possibly diff$k,q = 1,\ldots,h$erent in the cross covariance structure among their components. As mentioned before, several distances can be used to measure the differences between the two co$\mathbf{X}$ variance operators. We consider the distances introduced in Pigoli et al. [8], generalizing them to the case of non necessarily positive definite operators. In fact we are interested in quantifying also the distance between $V_{\mathbf{X}}^{kq}(s,t)$ and $V_{\mathbf{Y}}^{kq}(s,t)$ with $k \neq q$.

Let $d(V,W)$ denote a distance between two operators. We compute for each $k = 1,\ldots,h$ 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^2distance

$$d_L(V,W) = \sqrt{\int_I \int_I (v(s,t) - w(s,t))^2 ds dt}, \tag{5}$$

where $v(s,t)$ and $w(s,t)$ are the kernels of the operators V and W respectively, see eq. (4).

– Spectral distance

$$d_S(V,W) = |\lambda_1|, \tag{6}$$

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

– Square root pseudo distance

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

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

$$d_F(V,W) = \|V - W\|_{HS} = \sqrt{\text{trace}(V - W)^*(V - W)}. \tag{8}$$

– Procrustes pseudo distance

$$d_P(V,W) = d_P(|V|,|W|) = \inf_{R \in O(L^2(I))} \|L_1 - L_2 R\|_{HS}, \tag{9}$$

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 eq. (2):

$$p_k = \frac{d_k}{\sum_{s=1}^{h} d_s}, \qquad \text{for } k = 1,\ldots,h. \tag{10}$$

being $d_k = \sum_{q=1}^{h} d(V_{\mathbf{X}}^{kq}(s,t), V_{\mathbf{Y}}^{kq}(s,t))$ computed according the (pseudo) distances above. With this choice we should take into account not only the distances between intra-component variability, but also the inter-component ones. In fact, for each $k = 1,\ldots,h$ we compute the distance between the variance structures of the marginal components X_k and Y_k of the two stochastic processes, and we then sum up

the distances between the covariances with the remaining $h-1$ components. The higher is this distance, the higher is the weight of the corresponding component in calculating the depth measure of the multivariate functional data.

2.2 Inference on ranks induced by the depths

Let's now consider two groups of multivariate functional data generated by two distributions, say $P_{\mathbf{X}}$ and $P_{\mathbf{Y}}$. We want to generalize to this framework an inferential procedure to compare the two distributions $P_{\mathbf{X}}$ and $P_{\mathbf{Y}}$. Motivated also by the application at hand we aim at constructing a non parametric rank test where two samples of multivariate functions, generated according to $P_{\mathbf{X}}$ and $P_{\mathbf{Y}}$, can be compared (ECGs concerning physiological vs. pathological subjects). This test is based on depth proposed in eq. (2), with the weights choice stated in eq. (10). We deal with a sample characterized by more than one center since data come from a mixture of two distributions (physiological and pathological subjects in the application of interest). It is well known in liter ature about depth measures that the right way of extending the Wilcoxon rank sum test to the multivariate case is the following: consider a sample $\mathbf{f}_1, \ldots, \mathbf{f}_n$ generated according to a distribution $P_{\mathbf{X}}$ and another sample $\mathbf{g}_1, \ldots, \mathbf{g}_m$ generated according to a distribution $P_{\mathbf{Y}}$. We assume that there is a third reference sample, say $\mathbf{h}_1, \ldots, \mathbf{h}_N$, from one of the two populations, say $P_{\mathbf{X}}$ without loss of generality. We then compute the depths MBD of each $\mathbf{f}_i, i = 1, \ldots, n$ and each $\mathbf{g}_j, j = 1, \ldots, m$ with respect to the reference sample $\mathbf{h}_1, \ldots, \mathbf{h}_N$. In doing so, it is possible to rank the functions $\mathbf{f}_1, \ldots, \mathbf{f}_n, \mathbf{g}_1, \ldots, \mathbf{g}_m$. according to the related depths. Let us call these ranks $R_{\mathbf{h}_1,\ldots,\mathbf{h}_N}(\mathbf{f}_i), i = 1, \ldots, n$ and $R_{\mathbf{h}_1,\ldots,\mathbf{h}_N}(\mathbf{g}_j), j = 1, \ldots, m$, respectively. These ranks are transferred to the functions $\mathbf{f}_1, \ldots, \mathbf{f}_n, \mathbf{g}_1, \ldots, \mathbf{g}_m$. According to Liu and Singh [12], we can apply the Wilcoxon test to the induced ranks. In particular, the lower the depth the lower the rank. The proposed test statistic R is the sum of the ranks of the second sample $R_{\mathbf{h}_1,\ldots,\mathbf{h}_N}(\mathbf{g}_1), \ldots, R_{\mathbf{h}_1,\ldots,\mathbf{h}_N}(\mathbf{g}_m)$. According to the null hypothesis (H_0) there are no differences between the distributions generating the data. Hence $R_{\mathbf{h}_1,\ldots,\mathbf{h}_N}(\mathbf{g}_1), \ldots, R_{\mathbf{h}_1,\ldots,\mathbf{h}_N}(\mathbf{g}_m)$ can be viewed as a random sample of size m drawn without replacement from the set $(1, \ldots, n+m)$, and we reject H_0 for values of R too small. For large values of n and m it is possible to use a Normal approximation, see Li and Liu [13]. The presence of ties is treated as explained in Liu and Singh [12] and Lopez-Pintado and Romo [7].

2.3 Multivariate depth measures and prediction

Our aim is not only to test the differences between the two stochastic processes that generate the multivariate functional data, but also to predict the membership of a new statistical unit entering the study. This means that we aim both at comparing ECG traces (physiological and pathological subjects), and at quantifying the probability of being affected by the disease for new patients who enter the study. Once the MBDs of data are computed according to the explained procedure, a logistic regression model may be fitted as follows: given, $Y_i \sim Be(p_i)$ a response variable for $i \in 1, \ldots, n$ with $\theta_i = \log(p_i/(1-p_i))$, θ_i can be modelled as a linear transformation of the covariates (possibly functional) related to $i-$th statistical unit. Depths, in this case, may be considered as a dimensional reduction of the functional covariates and inserted as predictors in the following way:

$$\theta_i = \beta_0 + \beta_1 MBD_i + \sum_{h=1}^{w} d_{ih}\gamma_h. \tag{11}$$

The vector $\mathbf{d}_i = (d_{i1}, \ldots, d_{iw})^T$, $\mathbf{d}_i \in \mathbb{R}^w$, $i = 1, \ldots, n$, contains the traditional covariates that are available for the i-th statistical unit. The idea is to exploit all the available information concerning the statistical unit (the patient condition in the application) to predict her/his status (disease) at best.

3 Simulation studies

3.1 Robustness of ranks induced by multivariate functional depth measures

As previously mentioned, the definition of functional depth in eq. (1) is based on the definition of univariate depth by Lopez-Pintado and Romo [7]. The depth defined in eq. (1) involves the choice of the parameter $J \in \mathbb{N}$, which determines the number of units to be extracted from the dataset in order to build up the envelopes. Clearly, the entity of J strongly affects the computational effort needed for computing depths. In fact for $J > 2$ the computation of eq. (1) becomes quickly unaffordable. In Lopez-Pintado and Romo [7], authors state that the magnitude of J doesn't affect the ordering induced on (univariate) data by the values of depths; for this reason, it is possible to set $J = 2$ and greatly ease the computation of such quantities. In the following we will adopt the same value for J, supported by a simulation study to assess the validity of such a decision. Our simulation study takes advantage of an efficient implementation (see Tarabelloni [10]) of the algorithm for the computation of depths, which enables the use of values $J > 2$ in the computation of depths for datasets with standard dimensions. This implementation is done in $C++$ and exploits a MPI-parallel acceleration to subdivide the workload of computing the band depths among different processors. We have conducted a simulation study with $n = 50$ univariate functional data, $X_1, \ldots, X_N$, generated as follows:

$$X(t) = (1 + \alpha)\sin(2\pi t + \beta) + N(t). \tag{12}$$

In eq. (12) α and β are independent gaussian random variables with zero means and variances equal to 0.09 and 0.04, respectively. Moreover $\forall t \in (0, 2)$, $N(t)$ is assumed to be a zero mean gaussian random variable with variance equal to $2.5 * 10^{-4}$, independent of α and β. Such a choice for the family of simulated signals leads to a dataset of functional data (Figure 1) that shows variability in both amplitude and phase, as well as a moderate additive noise.

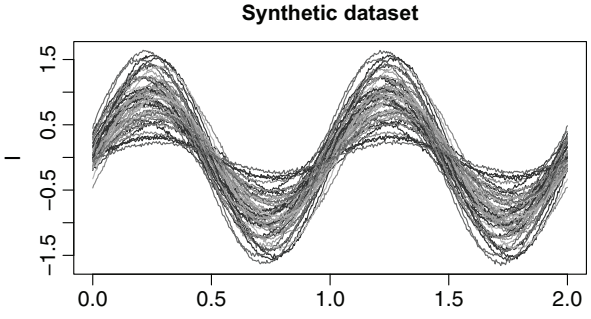

Figure 1: Plot of the the synthetic dataset simulated according to eq. (12), used to assess the robustness of ranks induced by depths defined in eq. (1) w.r.t the choice of parameter J.

We repeated the study of the stability of induced ranks also on the dataset of ECGs. In particular, we chose to consider the first lead (lead I) of the ECG and to conduct a robustness study on a subsample of $n = 50$ elements from the full dataset (see Section 4 for more details about the data). Just as before, we computed the depths of the 50 curves for $J = 2, 3, 4$, determining the corresponding ranks and studying their stability as J increases. Data are depicted in Figure 2, they are clearly more complex than simulated ones. We display a summary of the results of the study for both simulated and real data in Figure 3. On the horizontal axis are the ranks of the observations, and the corresponding depths, for each value of J considered, are on the vertical axis. Since the attribution of ranks to observations may change as J increases, we joined the points corresponding to the same observations with red lines. As a consequence it is possible to notice all the changes in the ordering of the dataset by looking for the crossings of red lines in the plot. The left panel of Figure 3 shows that a complete stability characterises the ranks of simulated data, and no swap occurs. In particular, the values of depths for a single observation increase with J (this is a consequence of definition (1)), but the reciprocal orders of the curves are the same, saying that the induced ranks are robust.

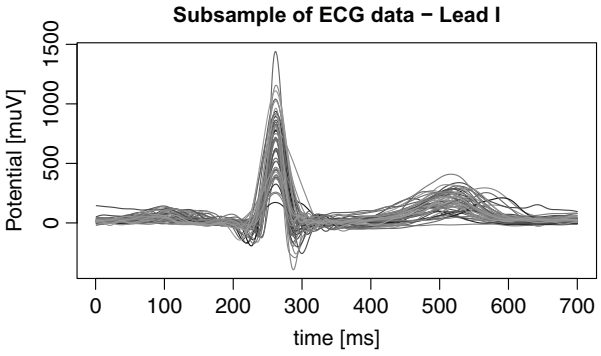

Figure 2: Plot of the subpopulation of $n = 50$ ECGs used to assess the robustness of ranks induced by depths defined in eq. (1) w.r.t the choice of parameter J.

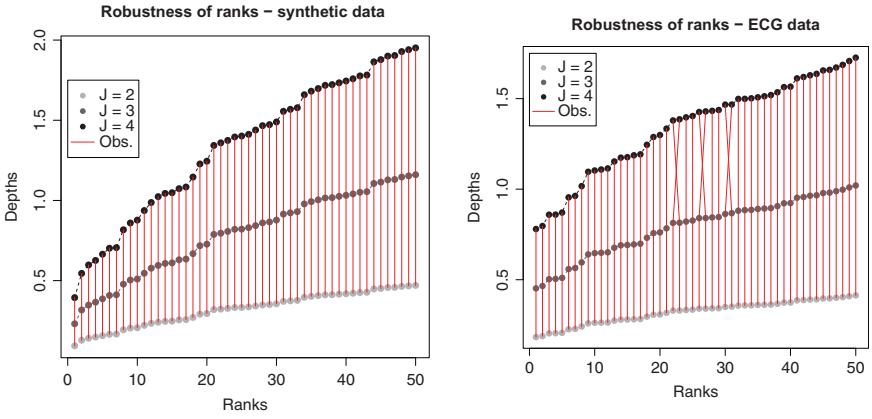

Figure 3: Representation of the of ranks induced by depths for increasing values of J, in case of simulated data (left) and a subsample of Lead I of healthy ECGs (right). The sets of points with the same colour indicate the same value of J for the computation of depths, while the red lines join the values of depths corresponding to the same observation in the dataset. This representation allows us to detect the changes of ranks by the crossings of the red lines, and to compare the values of depth of the observations that swapped their ranks.

The right panel of Figure 3 represents the results obtained using the same procedure in the case of ECG data. Here the shapes of curves defined by depths corresponding to equal J are still similar in the three cases, yet three swaps do happen between $J = 3$ and $J = 4$. Nonetheless, the observations whose ranks are exchanged have almost equal depth, thus ranks involved in the swapping are consecutive. As a consequence, the impact of the swaps on quantities depending on ranks or on depths is pretty negligible. The results we showed here are representative of the general performances one might expect from the ECG dataset. In fact Figure 3 shows the median outcome (in terms of swaps) obtained from sampling 100 times a subpopulation of $n = 50$ data from the ECG dataset, and repeating the stability analysis.

To conclude, the two simulation studies we proposed here, on both simulated and real data, give empirical evidence to assume the stability of ranks induced by functional depths. Therefore, in order to reduce computational efforts, hereafter we can set $J = 2$.

3.2 Populations with pure mean difference

As we have explained before, our suggestion for the choice of weights $\{p_k\}$ in eq. (2) directly takes into account the possible differences in the variance-covariance structure of multivariate functional data. In particular, our aim is to use data depth in order to carry out a supervised learning between two different populations of functions, $\mathbf{X} \sim P_{\mathbf{X}}$ and $\mathbf{Y} \sim P_{\mathbf{Y}}$. It is our belief, then, that the weights built upon the variance-covariance operators distances are able to capture information regarding the variability of the two generating processes.

In principle, (relative) depths are mainly concerned with (relative) amplitude of signals, despite they strongly depend on location shift of data. In other words, it's reasonable to expect that differences in the means of the two generating processes have direct impact on the magnitude of relative depths, which in general will be lower for data belonging to the population with location shift. In view of this, the choice of weights relying on variance-covariance structures of the processes has both the role of regaining information on variability and modulating the (marginal) centrality/outlyingness on each component by the corresponding amount of difference in variability.

This can be observed through an example carried out on simulated data. In particular, we want to show that, when we consider two multivariate processes $\mathbf{X}$ and $\mathbf{Y}$ with equal variance-covariance structures and different means, univariate functional depths are able to capture this difference, and definition (2) gives overall multivariate depths which directly reflect the contributions of the single components.

The simulation setting is as follows. We considered $n_1 = 10^3$ $n_2 = 10^3$, $N = 2 \cdot 10^3$ independent observations from gaussian, bivariate processes, respectively $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{R}$. We can assume $\mathbf{R}$ with the same law of $\mathbf{X}$ as a bigger population reference. The two processes share the same variance-covariance structure, i.e.,

$$\mathcal{C} = \begin{pmatrix} \mathcal{C}_1 & 0 \\ 0 & \mathcal{C}_2 \end{pmatrix} \qquad \mathcal{C}_1 \equiv \mathcal{C}_2 = \sum_{i=1}^{L} \lambda_i \psi_i \otimes \psi_i$$

We chose $L = 3$, $\lambda_i = 1$, $\forall i = 1, \ldots, L$ and $\{\psi_1, \ldots, \psi_L\}$ the truncated $L-$dimensional Fourier basis of $L^2([0,1])$. It can be noticed that the two processes have equal variability and that the two components are equally responsible for the overall variability of the corresponding process. Being the covariances of the three processes exactly the same, we cannot use directly (10), but we have to compute the two weights of each population by a perturbation argument: take $\varepsilon > 0$, and let $\mathcal{C}_\varepsilon = (1+\varepsilon)\,\mathcal{C}$ be the covariance of $\mathbf{X}_\varepsilon$ and $\mathbf{Y}_\varepsilon$; then $p_{k,\varepsilon} \equiv 1/2$, $k = 1, 2$ and $\forall \varepsilon > 0$, hence we can set $p_k = 1/2$, $k = 1, 2$ also for $\mathbf{X}$ and $\mathbf{Y}$. This means that equal weight is put on each component, due to the equal informative content brought by the two components in terms of variability. Of course, in a practical context, the weights will not be equal, due to the inaccuracy of both sampling of the observations and the estimation of covariances. This makes their computation possible also when covariances are equal.

For what concerns the means of the processes, we set $\mu_{\mathbf{X}} = \mathbf{0}$ and $\mu_{\mathbf{Y}} = (0, \alpha)$ with $\alpha = 4$. To sum up, the observations are generated according to:

$$\mathbf{X} = \begin{cases} \sum_{i=1}^{L} \sqrt{\lambda_i}\, \xi_{i,1}\, \psi_i \\ \sum_{i=1}^{L} \sqrt{\lambda_i}\, \xi_{i,2}\, \psi_i \end{cases} \qquad \mathbf{Y} = \begin{cases} \sum_{i=1}^{L} \sqrt{\lambda_i}\, \eta_{i,1}\, \psi_i \\ \alpha + \sum_{i=1}^{L} \sqrt{\lambda_i}\, \eta_{i,2}\, \psi_i \end{cases}.$$

where $\xi_{i,1}, \xi_{i,2}, \eta_{i,1}$ and $\eta_{i,2}$ are standard normal i.i.d random variables for $i = 1, 2$. The population $\mathbf{R}$ is generated according to the same process of $\mathbf{X}$.

After simulating the three datasets, we computed the empirical weights, which for $\mathbf{X}$ are $p_1 = 0.497, p_2 = 0.503$ and for $\mathbf{Y}$ are $p_1 = 0.504, p_2 = 0.496$, along with mariginal univariate and multivariate depths. We display the results in Figure 4. Looking at the boxplots of univariate depths, and considering the original data in Figure 5, it is clear that the offset in the means of the two families induces a quick decrease in the (relative) depth of the second one apart from those observations that are in the same range of the first family due to variability. The boxplot of the multivariate depth, computed with the weights specified before shows that the strong outlyingness of second components in the second family is, on average, mitigated by the first ones. Nonetheless, the two families are truly different in just one component, and yet a clear distinction between the two populations is still

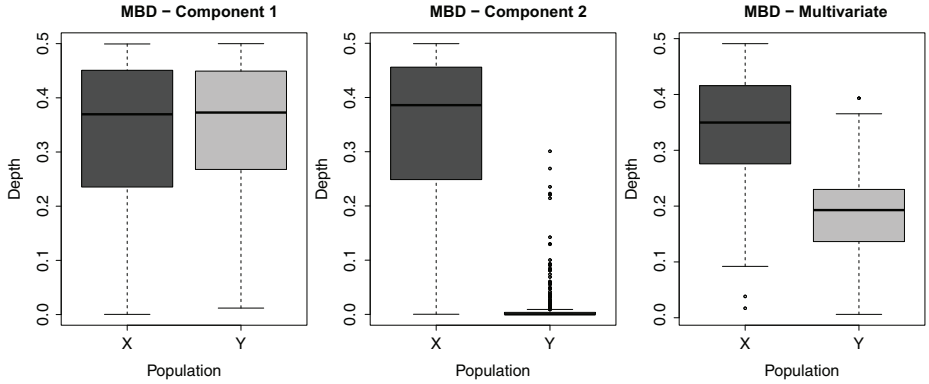

Figure 4: Boxplots of the univariate (left and center) and multivariate (right) depths of populations **X** (dark grey) and **Y** (light grey) in the simulation study.

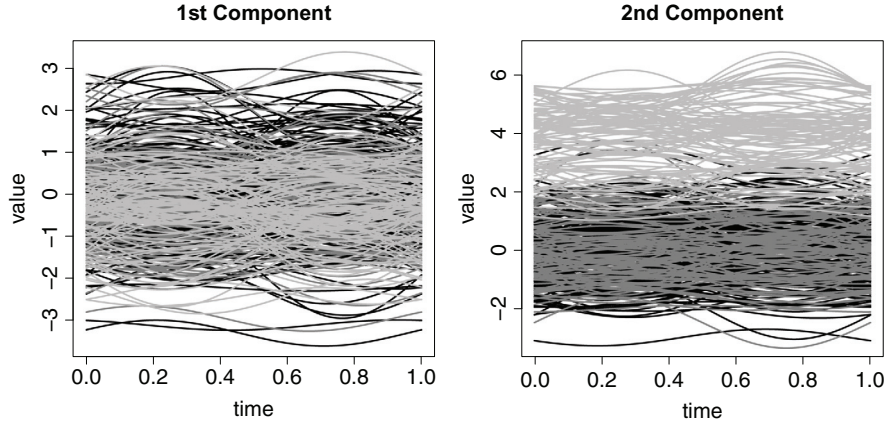

Figure 5: Example of populations simulated according to the laws of processes **X** (dark grey), **Y** (light grey) and **R** (black). In particular, this is a subsample of the three populations used to conduct the simulation study.

possible, meaning that pure differences in means observed only in some components are still detected by choosing weights based on covariance operators.

4 Application to ECG signals

In this section we apply the methods presented in Section 2 to ECGs. In this case, the basic statistical unit (patient) is characterized by a 8-variate function (ECG) which describes his/her heart dynamics on the eight leads I, II, V1, V2, V3, V4, V5 and V6, together with the corresponding derivatives. Here, the binary outcome we consider is the group label, indicating the presence/absence of the disease. It is modeled by a Bernoulli random variable Y_i, which takes value 1 if Left Bundle Branch Block is diagnosed, and 0 if the trace is physiological. We analyse ECG traces from PROMETEO (PROgetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall'Extra Ospedaliero) database. PROMETEO has been started with the aim of spreading the intensive use of ECGs as pre-hospital diagnostic tool. The project was also a way of constructing a new database of ECGs with features never recorded before in any other data collection on heart diseases. Indeed, ECG recorders with GSM transmission have been installed on all Basic Rescue Units of Milan urban area thanks to the partnerships of Azienda Regionale Emergenza Urgenza (AREU), Abbott Vascular and Mortara Rangoni Europe s.r.l.

Each file contained in PROMETEO can be associated to three sub-files. The first is called *Details* and consists of technical information, useful for signal processing and analysis. More precisely, it includes waves repolarisation and depolarisation times, landmarks indicating onset and offset times of the main ECG's subintervals and an automatic diagnosis, established by the commercial Mortara-Rangoni VERITAS^{TM} algorithm. We used these automatic diagnoses to label the ECG traces we analysed. The second sub-file is called *Rhythm* and contains the output of an ECG recorder. Specifically, it registers 10 seconds (10,000 sampled points) of the ECG signal. The third file is called *Median*. It is built from the *Rhythm* file, and depicts a *reference* beat lasting 1.2 seconds on a grid of 1,200 points. We carried out the analysis using the *Median* files, i.e., using 8 curves (one for each ECG lead) for each patient, representing patient's "Median" beat for that lead. This representative heartbeat is a trace of a single cardiac cycle (heartbeat), i.e., of a *P* wave, a *QRS* complex, a *T* wave, and a *U* wave, which are normally visible in 50%–75% of ECGs.

The sample we analyse consists of the ECG signals of $n = 149$ subjects, among which 101 are Normal and 48 are affected by Left Bundle Branch Block. Figure 6 shows denoised and registered data we consider for our analysis (see Ieva et al. [2] for further details on wavelet denoising and landmarks registration adopted for preprocessing data). The black solid lines represent the mean functions.

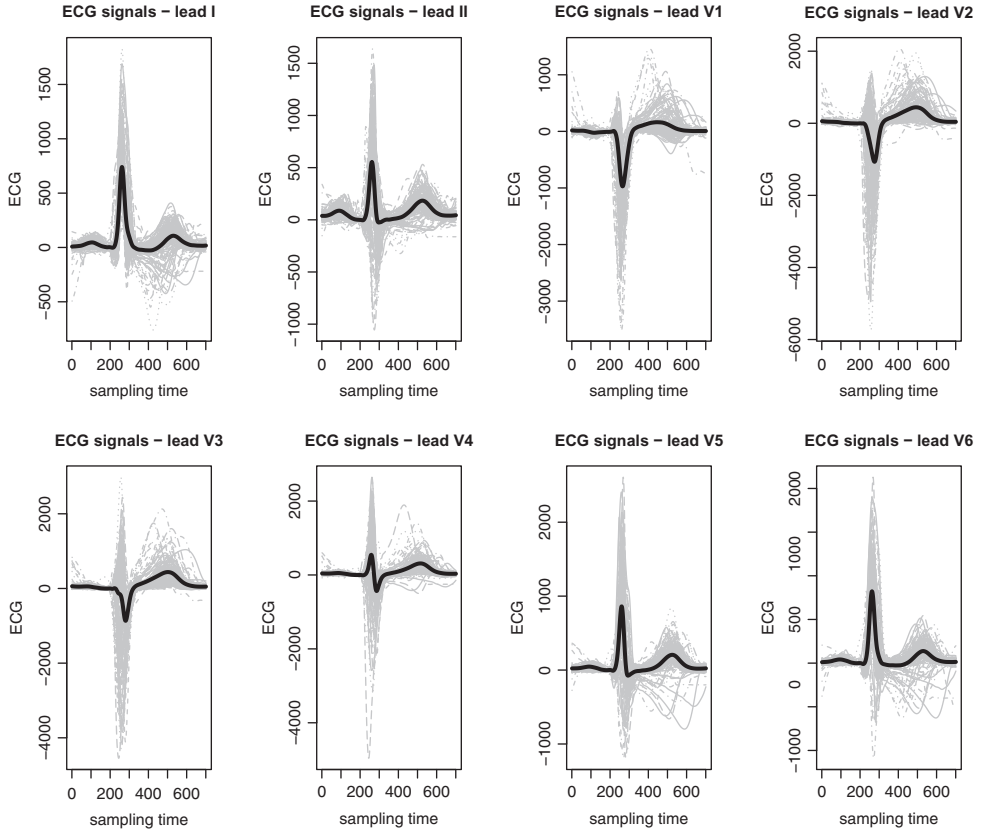

Figure 6: Denoised and registered data (8 leads) for the 149 patients with superimposed the mean functions (black solid lines).

To compute the MBDs according to the procedure described in Section 2, we randomly chose 50 ECGs from the physiological traces for being the reference group. Then we computed the ranks of the remaining 51 physiological and 48 LBBB traces with repect 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 2. The results are robust with respect to the distances choice, so we will present in the following the results obtained with the Procrustes pseudo-distance.

Table 1: Weights p_k to be inserted in eq. (2) arising from the choice of the Procrustes pseudo-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

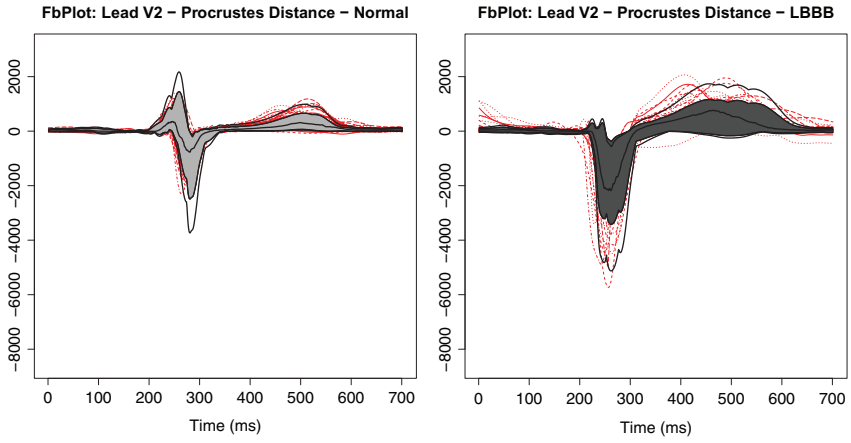

Figure 7: Functional boxplots (only lead V2 is depicted) of 101 physiological traces (left panel) and of the 48 LBBB signals (right panel). The central bands (grey area), the fences (solid lines) and outliers (dotted lines) are computed according to the ranking induced by $MBD_n^J(\mathbf{f})$ defined in eq. (2), and all the leads are weighted using Table 1.

The weights to plug in the formula (2) are shown in Table 1 (leads are ordered according to decreasing weight).

Figure 7 shows the multivariate functional boxplots (only the lead V2, the most relevant according to the weights reported in Table 1) for the 101 physiological (left panel) and the 48 LBBB (right panel) signals.

The p-value of the Wilcoxon tests carried out to compare the distributions of the depths is $5.352 * 10^{-14}$ and over all the 20 cases is always less or equal to $3.02 * 10^{-12}$. Figure 8 shows the distributions of the MBDs, stratified by the presence/absence of Left Bundle Branch Block in one representative case over the 20

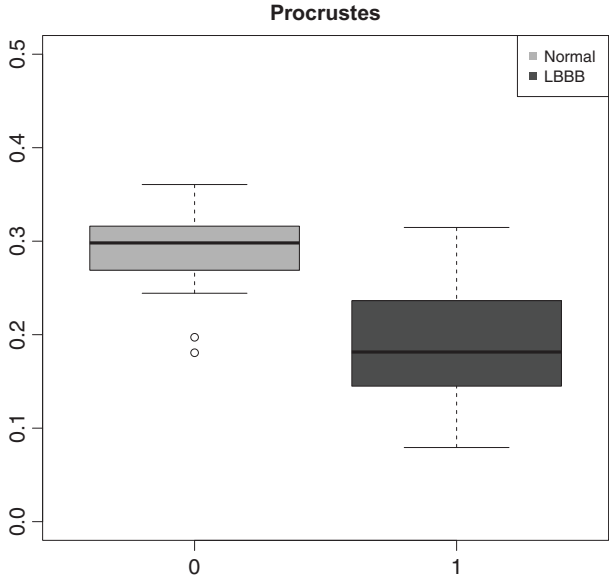

Figure 8: Distributions of data depth measures computed according to eq. (2) with the weights reported in Table 1. On the left data concerning Normal subjects, on the right data concerning patients affected by LBBB.

explored. This picture further supports belief that evidence for the difference among the two population exists and is significant.

Thus we fitted the logistic regression model in eq. (11) to these data, considering both signals and derivatives as functional covariates to be summarized through MDBs. Being Y_i, $i = 1, \ldots, n$ the label of healthy ($Y_i = 0$) and unhealthy ($Y_i = 1$) signals, the final model was:

$$\theta_i = \beta_0 + \beta_1 MBD_i \tag{13}$$

Only MBDs of signals result to be significant for the generalized linear regression model so we dropped the MBDs of the derivatives out. They are not significant, probably due to the high correlation with the corresponding MBDs of signals. The term $\sum_{h=1}^{w} d_{ih}\gamma_h$ in eq. (11) is missing because we do not have additional covariates in our dataset. The estimates of parameters in model (13) are reported in Table 2.

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 in eq. (13) equal to 0.5.

Table 2: Estimates, standard errors and p-values for the parameters of the logistic model (13).

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 3: Confusion matrix of the classification carried out by the logistic model (13) with threshold set equal to 0.5.

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

Considering all the 20 different cases adopted for carrying out the Wilcoxon test, we obtain the following summary results, in terms of mean ($\pm$ std. dev.): sensitivity equal to $84.48(\pm 2.29)\%$, specificity equal to $89.80(\pm 1.87)\%$; the correct classification rate equal to $87.22(\pm 1.58)\%$ and the leave-one-out cross validation error equal to $9.61(\pm 1.18)\%$. These results refer to Procrustes distance, which is the one performing best in term of correct classification rate. The results are satisfactory. We conclude that considering the distances between covariance operators is a good prognostic factor for identifying group membership of patients via logistic regression.

5 Conclusions

In this paper, we focus on supervised learning from multivariate functional data, that is 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. In particular the motivating aim is the analysis of the 8-leads Electrocardiographic traces of patients whose pre-hospital ECG has been sent to 118 Dispatch Center of Milan (the Italian free-toll number for emergencies) by life support personnel of the basic rescue units.

We focus on the binary outcome indicating the presence of cardiovascular ischaemic event, in order to estimate the probability of each patient to be affected by Acute Myocardial Infarction. This can be done

summarizing some relevant features of data (multivariate functional curves) by means of non parametric statistical objects, i.e., the depth measures. In fact, computing the depth measure of multivariate functional data and averaging the contribution of each component of the multivariate signal to the depth itself revealed to be an effective way for defining powerful predictors of the disease presence. We showed how to choose the weights taking into account the distance between the estimated covariance operators of the two groups. This has been carried out considering many different distances between covariance operators in the infinite dimensional setting, so that a robustness assessment and a comprehensive performances evaluation can be provided.

The methodological framework proposed in this paper represents a different way for handling complexity of multivariate functional data using robust statistics such depth measures. In fact it is often complex to summarize and quantify the information embedded in signals in order to make inference and predictions, especially when they are proxies of complex disease mechanisms. The results obtained in the paper show that we can rely on some robust procedures in order to accomplish all these goals.

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. Ieva F, Paganoni A. Depth measures for multivariate functional data. Commun Stat Theory Met 2013a;42:1265–76.
2. Ieva F, Paganoni A, Pigoli D, Vitelli V. Multivariate functional clustering for the morphological analysis of ECG curves. J R Stat Soc Ser C (Appl Stat) 2013;62:401–18.
3. Ieva F, Paganoni A. Risk prediction for myocardial infarction via generalized functional regression models. Stat Met Med Res [Internet]. 2013b Jul 18. Available from: http://smm.sagepub.com/content/early/2013/07/09/0962280213495988. DOI: 10.1177/0962280213495988.
4. Ramsay J, Silverman B. Functional data analysis, 2nd ed. New York: Springer, 2005.
5. Berrendero J, Justel A, Svarc M. Principal components for multivariate functional data. Comput Stat Data Anal 2011;55:2619–34.
6. Lopez-Pintado S, Romo J. Depth-based inference for functional data. Comput Stat Data Anal 2007;51:4957–68.
7. Lopez-Pintado S, Romo J. On the concept of depth for functional data. J Am Stat Assoc 2009;104:718–34.
8. Pigoli D, Aston J, Dryden I, Secchi P. Distances and inference for covariance functions. Biometrika 2014;101:409–22.
9. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2013 [cited 2015 Jun 15]. Accessed June 15, 2015. Available from: http://www.R-project.org/, ISBN 3-900051-07-0.
10. Tarabelloni N. Tools for computational statistics coded in C++, 2013 [cited 2015 Jun 15]. Available from: https://github.com/ntarabelloni/HPCS.
11. Lopez-Pintado S, Sun Y, Genton M. Simplicial band depth for multivariate functional data. Adv Data Anal Classif 2014;8:321–38.
12. Liu R, Singh K. A quality index based on data depth and multivariate rank tests. J Am Stat Assoc 1993;88:252–60.
13. Li J, Liu R. New nonparametric tests of multivariate locations and scales using data depth. Stat Sci 2004;19:686–96.