MOX–Report No. 54/2013

Use of depth measure for multivariate functional data in disease prediction: an application to electrocardiographic signals

Biasi, R.; Ieva, F.; Paganoni, A.M.; Tarabelloni, N.

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

Use of depth measure for multivariate functional data in disease prediction: an application to electrocardiographic signals

Rachele Biasi[♯], Francesca Ieva[♯], Anna Maria Paganoni[♯] and Nicholas Tarabelloni[♯]

November 13, 2013

[♯] MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy

rachele.biasi@mail.polimi.it, francesca.ieva@polimi.it,
anna.paganoni@polimi.it, nicholas.tarabelloni@mail.polimi.it

Keywords: Depth measures; multivariate functional data; covariance operators; ECG signals, generalized linear models.

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. 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.

1 Introduction

In many real-life applications of statistics nowadays, data are becoming more and more complex. This is particulary true in the biomedical and healthcare 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 [5] 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 data describes the same biological event, i.e., the representative heartbeat of a patient (see [7] for a deeper explanation of this). 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 multivariate functional data as predictors, with both classification and prediction purposes.
In [6] a similar problem has been faced, mainly performing a data dimensionality reduction by a Multivariate Functional Principal Component Analysis, see [14, 1]. It consists of summarizing the information carried out by the 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. On the other hand, in this work we summarize some relevant feature of data through non parametric statistical objects, i.e., the depth measures. In [5] the depth measure introduced in [11, 12] for univariate functional data was extended to the multivariate functional setting. However, to compute the depth measure proposed in that paper, 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 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 (and possibly of the derivatives) contains information about the reciprocal role of the signal (derivative) components with respect one to each other. 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, we consider many different distances between covariance operators in the infinite dimensional setting,

as discussed in Pigoli et al. (2012).
A different definition of depth for multivariate functional data could be find in [3]. More specifically they construct their definition of depth measure starting from the Tuckey's halfspace depth [15] as building block. The definition proposed in [3] averages a multivariate depth function over the time points, and includes a weight function which accounts for information provided by the warping functions used to register functional data. Since our motivating problem deals with ECG data that are characterized by strongly localized features (peaks, oscillations,...) and the information catched by warping functions of the registration procedure based on landmarks is poor for classification purposes, see [7], we prefer to go on with the definition of depth proposed in [5].

The paper is structured as follows: in Section 2, the definition of multivariate functional depth measure is presented and the choice of the weights is discussed. In Section 3, a simulation study to support the robustness of the proposed method is detailed. Section 4 concerns the analysis of ECG data arising from PROMETEO dataset. Finally, in Section 5, conclusions are drawn and further developments are proposed. All the analyses are carried out using `R` statistical software [13] and the ad-hoc C++/MPI parallel library for computational statistics `HPCS`[1] [4].

2 Multivariate depth measures with covariance driven weights

Let us start by recalling the definition of band depth for multivariate functional data introduced in [5]. This definition has been introduced to generalize to the multivariate framework the concept of band depth for functional data introduced in [11, 12].
Let $\mathbf{X}$ be stochastic process taking values in the space $\mathcal{C}(I;\mathbb{R}^h)$ 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

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

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

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

curves $(f_{i_1;k}, ..., f_{i_j;k})$, means computing

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

where $E(f_k) := E(f_k; f_{i_1;k}, ..., f_{i_j;k}) = \{t \in I, \min_{r=i_1,...,i_j} f_{r;k}(t) \le f_k(t) \le \max_{r=i_1,...,i_j} f_{r;k}(t)\}$ and $\tilde{\lambda}(f_k) = \lambda(E(f_k))/\lambda(I)$ with λ the Lebesgue measure on I. Statistical properties of the depth measure defined in (1) as well as inferential tools based on this concept are detailed in Ieva et al. (2013a).
An open problem in (1) is how to choose the weights $p_1, ..., p_h$. In general this choice is problem driven. Whenever the principal aim of the analysis is the comparison between two different populations it would be useful to take advantage of the study of the differences between the variance-covariance structures of the two samples. In fact, it is reasonable to expect that they carry out information about possibly differences between the two populations.

In the following we consider the depth measure (1) with $J = 2$. This choice is motivated by a robustness study conducted in [2], where the stability of the order induced by different choices of J is verified, taking advantage of an efficient implementation within the parallel library environment `HPCS` [4].

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 $1 \le l \le h$, 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 $\mathcal{V}_{\mathbf{X}}$ 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:

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

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

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

where $\otimes$ is 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 $k, q = 1, ..., h$.

In what follows, we deal with two different stochastic processes $\mathbf{X}$ and $\mathbf{Y}$ with covariance operators $\mathcal{V}_{\mathbf{X}}$ and $\mathcal{V}_{\mathbf{Y}}$, respectively, possibly different in the cross covariance structure among their components. As mentioned before, several

distances can be used to measure the differences between the two covariance operators. We consider the distances introduced in [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^2 **distance**

$$d_L(V,W) = \sqrt{\int_I \int_I (v(s,t) - w(s,t))^2 ds dt}, \tag{3}$$

where $v(s,t)$ and $w(s,t)$ are the kernels of the operators V and W respectively, see (2).

•$-$ **Spectral distance**

$$d_S(V,W) = |\lambda_1|, \tag{4}$$

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

•$-$ **Square root pseudo distance**

$$d_R(V,W) = \| \, |V|^{\frac{1}{2}} - |W|^{\frac{1}{2}} \, \|_{HS}, \tag{5}$$

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**

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

•$-$ **Procrustes pseudo distance**

$$d_P(V,W) = d_P(|V|,|W|) = \inf_{R \in O(L^2(I))} \|L_1 - L_2 R\|_{HS}, \tag{7}$$

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):

$$p_k = \frac{d_k}{\sum_{k=1}^{h} d_k}, \qquad for\ k = 1, ..., h. \tag{8}$$

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.

We also want to generalize to this framework a non parametric rank test where two samples of multivariate functions can be compared, based on depth proposed in (1), with the weights choice stated in (8). We deal with a sample characterized by more than one center since data come from a mixture of distributions (physiological and pathological subjects in the application of interest). It is well known in literature about depth measures that the right way of extending the Wilcoxon rank sum test to the multivariate case is the following: consider a sample $\mathbf{f}_1, ..., \mathbf{f}_n$ generated according to a distribution $P_{\mathbf{X}}$ and another sample $\mathbf{g}_1, ..., \mathbf{g}_m$ generated according to a distribution $P_{\mathbf{Y}}$. We assume that there is a third reference sample, say $\mathbf{h}_1, ..., \mathbf{h}_N$, from one of the two populations, say $P_{\mathbf{X}}$ without loss of generality. We then compute the MBD of each $\mathbf{f}_i, i = 1, ..., n$ and each $\mathbf{g}_j, j = 1, ..., m$ with respect to the reference sample $\mathbf{h}_1, ..., \mathbf{h}_N$. In so doing, it is possible to rank the functions $\mathbf{f}_1, ..., \mathbf{f}_n, \mathbf{g}_1, ..., \mathbf{g}_m$. Let $R(P_N, \mathbf{f}_i)$ be the proportion of $\mathbf{h}_k$'s, $k = 1, ..., N$, with MBD less than or equal to the MBD of $\mathbf{f}_i$, where the MBD is computed with respect to the reference sample $\mathbf{h}_1, ..., \mathbf{h}_N$. An analogous definition is assumed for $R(P_N, \mathbf{g}_i)$. Then we order these values, $R(P_N, \mathbf{f}_i)$ and $R(P_N, \mathbf{g}_i)$, from the smallest to the highest giving them a rank from 1 to $n + m$. This induces a rank on the functions $\mathbf{f}_1, ..., \mathbf{f}_n, \mathbf{g}_1, ..., \mathbf{g}_m$. According to [10], 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 $R(P_N, \mathbf{g}_1), ..., R(P_N, \mathbf{g}_m)$. According to the null hypothesis (H_0) there are no differences between the distributions generating the data. Hence $R(P_N, \mathbf{g}_1), ..., R(P_N, \mathbf{g}_m)$ can be viewed as a random sample of size m drawn without replacement from the set $(1, ..., n + m)$, 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 [9]. The presence of ties is treated as explained in [10] and [12].

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. So, once the MBDs of data are computed according to the explained procedure, we consider a logistic regression model, where the response variable is $Y_i \sim Be(p_i)$ for $i \in 1, \ldots, n$ and $\theta_i = \log(p_i/(1-p_i))$. Y_i is a general binary outcome that depends on the multivariate functional data. θ_i can then be modelled as a linear transformation of the covariates related to $i-$th statistical unit, that is,

$$\theta_i = \beta_0 + \beta_1 MBD_i + \sum_{h=1}^{w} d_{ih}\gamma_h. \tag{9}$$

The vector $\mathbf{d}_i = (d_{i1}, ..., d_{iw})^T$, $\mathbf{d}_i \in \mathbb{R}^w$, $i = 1, \ldots, n$, contains the traditional covariates that are possibly 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. For the multivariate functional covariate, this can be done adding to a classical regression model an effective summary of the multivariate functional data information content.

3 Simulation results

In this Section we present two simulation studies, in order to support the methodology presented in the previous section. The first one is aimed at exploring the weights behavior with respect to different (pseudo) distance notions and different correlations between the components of the multivariate functional processes generating the data. The second one provides a benchmark for classification via logistic regression with depth measures used as predictors into the model.

3.1 Case A: weights behaviour with respect to the choice of the distance and of the correlation 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

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

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

$$S_i = \begin{pmatrix} A_i & C_i \\ C_i^T & B_i \end{pmatrix} \tag{10}$$

being

$$A_1 = \begin{pmatrix} 10^6 & 10^2 & & & & & \\ 10^2 & 10^6 & 10^2 & & & & \\ & \cdot & \cdot & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & 10^2 & 10^6 & 10^2 & \\ & & & & 10^2 & 10^6 & \end{pmatrix}, B_1 = \begin{pmatrix} 10^5 & 1 & & & & & \\ 1 & 10^5 & 1 & & & & \\ & \cdot & \cdot & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & 1 & 10^5 & 1 & \\ & & & & 1 & 10^5 & \end{pmatrix}, \tag{11}$$

$$A_2 = \begin{pmatrix} 10^4 & 1 & & & & & \\ 1 & 10^4 & 1 & & & & \\ & \cdot & \cdot & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & 1 & 10^4 & 1 & \\ & & & & 1 & 10^4 & \end{pmatrix}, B_2 = \begin{pmatrix} 10^3 & 10^2 & & & & & \\ 10^2 & 10^3 & 10^2 & & & & \\ & \cdot & \cdot & \cdot & & & \\ & & \cdot & \cdot & \cdot & & \\ & & & 10^2 & 10^3 & 10^2 & \\ & & & & 10^2 & 10^3 & \end{pmatrix}, \tag{12}$$

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

We generate 100 statistical units for each population and we repeat the procedure 100 times. Simulated curves are represented in Figure 1.

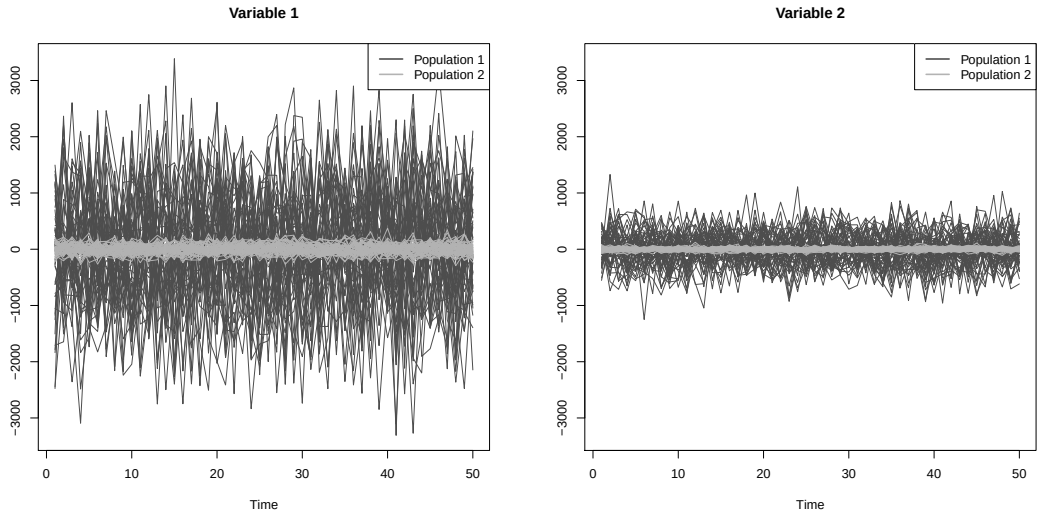

Figure 1: Sample of 100 curves generated according to models detailed in (11) and (12), with $\rho = 0.4$.

Table 1 shows the mean and standard deviation of weights of each component computed according to each choice of the distance between covariance operators.

(Pseudo) distance	ρ	**Mean weight Variable 1**	**Mean weight Variable 2**	**St. Dev. weights**
L^2	0	0.8062	0.1938	0.0129
	0.2	0.8037	0.1963	0.0137
	0.4	0.7949	0.2051	0.0152
	0.6	0.7842	0.2158	0.0122
	0.8	0.7718	0.2282	0.0099
	1	0.7599	0.2401	$2.3*10^{-5}$
Spectral	0	0.8539	0.1461	0.0071
	0.2	0.8368	0.1632	0.0080
	0.4	0.8194	0.1806	0.0086
	0.6	0.7985	0.2015	0.0081
	0.8	0.7785	0.2215	0.0054
	1	0.7597	0.2403	$1.1*10^{-5}$
Square root	0	0.6525	0.3475	0.0063
	0.2	0.6522	0.3478	0.0039
	0.4	0.6501	0.3499	0.0030
	0.6	0.6472	0.3528	0.0022
	0.8	0.6441	0.3559	0.0011
	1	0.6400	0.3600	$1.2*10^{-5}$
Frobenius	0	0.8064	0.1936	0.0032
	0.2	0.8035	0.1965	0.0029
	0.4	0.7958	0.2042	0.0027
	0.6	0.7725	0.2275	0.0027
	0.8	0.7725	0.2275	0.0018
	1	0.7597	0.2403	$2.7*10^{-6}$
Procrustes	0	0.6537	0.3463	0.0063
	0.2	0.6533	0.3466	0.0039
	0.4	0.6511	0.3489	0.0030
	0.6	0.6479	0.3521	0.0022
	0.8	0.6444	0.3556	0.0011
	1	0.6400	0.3600	$1.2*10^{-5}$

Table 1: Mean and standard deviation of the weights of the two components of the simulated bivariate data for different values of ρ and different (pseudo) distances.

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. 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.

3.2 Case B: classification via logistic regression with depth measures used as predictor

In this case we use the depth measures computed according to the procedure detailed in Section 2 as predictor in a logistic regression model. The time interval I is sampled over an evenly spaced grid of 50 points as in the first case. The data are generated according the following stochastic processes

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

for the first and the second population respectively. The covariances S_i, for $i = 1, 2$ are as in (10) but with the following choices: $A_1 = 10 * \mathbf{1}_{50}$, $B_1 = 8 * \mathbf{1}_{50}$, $A_2 = 5 * \mathbf{1}_{50}$, $B_2 = 4 * \mathbf{1}_{50}$, where $\mathbf{1}_n$ is the identity matrix $(n \times n)$ and $\rho = 0.4$. We consider a reference sample of 125 units from the first population, and another one of dimension 25 from the second population. For 20 different runs of the simulation and different distance choices we randomly extract 25 units from the first sample and we compute the depth measures of the 50 data (25 from the first sample and 25 from the second one), with respect to the 100 remaining of the reference sample. The p-value of the Wilcoxon tests carried out to compare the distributions over all the 20 cases is always less or equal to $3.39 * 10^{-5}$.

The results are very robust with respect to the pseudo-distance choice. Anyway among all the Procrustes pseudo-distance provides a slightly better discrimination power. For this reason we choose to report only the results concerning the Procrustes pseudo-distance.

Table 2 presents the mean confusion matrix obtained comparing the true and the estimated labels given by a logistic regression model like in (9), with MBDs of the ECG signals as unique covariates. We set the threshold for the classification carried out by the logistic model as equal to 0.5.

	Units from sample 1	Units from sample 2
Classified as units from sample 1	20.05	4.2
Classified as units from sample 2	4.95	20.80

Table 2: Confusion matrix. 50 units (25 from the first sample and 25 from the second one) are classified via logistic regression.

The following quantities have been computing from Table 2: sensitivity, specificity, the correct classification rate and the leave-one-out cross validation error of the generalized linear model. The sensitivity is equal to 83.20 ($\pm$2.46) %, the specificity is equal to 80.20 ($\pm$4.94) %, the correct classification rate is equal to 81.70 ($\pm$3.20) % and the leave-one-out cross validation error is equal to 15.02 ($\pm$1.96) %.

This second simulation study leads to a satisfactory correct classification rate and acceptable values for sensitivity and specificity. It is worth noting that in

general the choice of the distance is problem driven. Each kind of data requires the distance which better distinguishes the differences between populations.

4 Application to ECG signals

In this Section we apply the methods presented in Section 2 to the ECGs data. In this case, the basic statistical unit is the 8-variate function (the ECG) which describes the heart dynamics of each patient 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 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 (10000 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 1200 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% to 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 2 shows denoised and registered data we consider for our analysis (see [7] for further details on wavelet denoising and landmarks registration adopted for preprocessing data). The black solid lines represent the mean functions.
To compute the MBDs according to the procedure described in Section 2, we

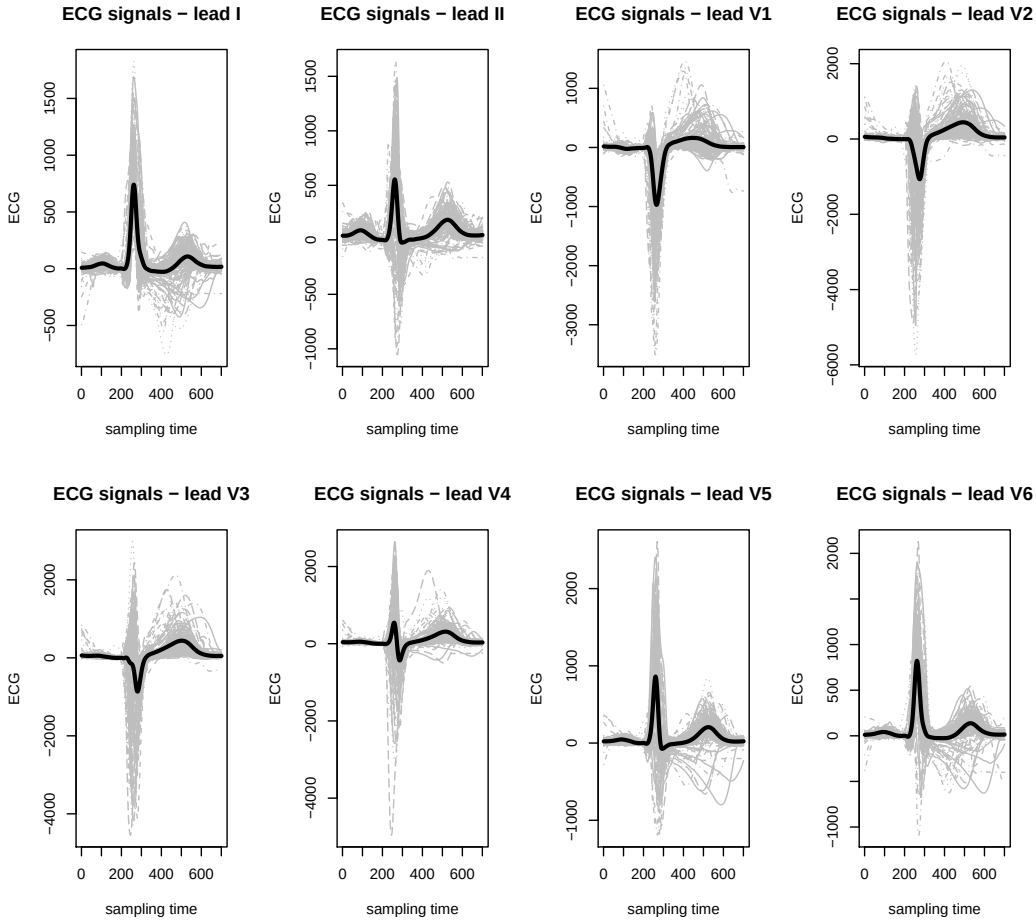

Figure 2: Denoised and registered data (8 leads) for the 149 patients with superimposed the mean functions (black solid lines).

randomly chose 50 ECGs from the physiological traces for being the reference group, and 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 very robust with respect to the distances choice, so we will present in the following the results obtained with the Procrustes pseudo-distances we did in the simulation study.

The weights to plug in the formula (1) are the following (leads are ordered according to decreasing weight):
Figure 3 shows the multivariate functional boxplots (only the lead V2, the most relevant according to the weights reported in Table 3) for the 101 physiological (left panel) and the 48 LBBB (right panel) signals.

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 3: Weights induced by the Procrustes pseudo-distance, to be inserted in the MDB formula given in (1).

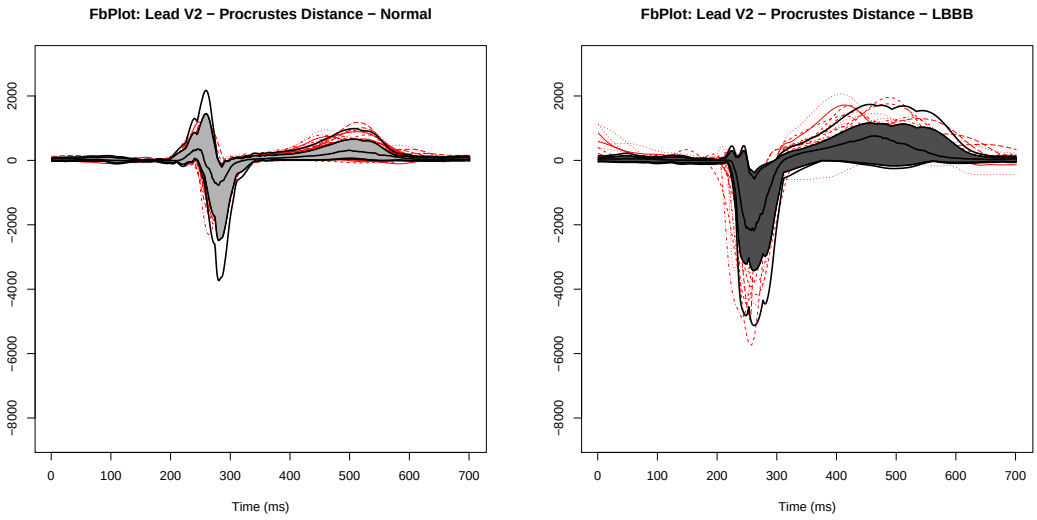

Figure 3: Functional boxplots (only lead V2 is depicted) of 101 physiological traces (left panel) and of the 48 LBBB signals (right panel). The central bands (grey area), the fences (solid lines) and outliers (dotted lines) are computed according to the ranking induced by $MBD_n^J(\mathbf{f})$ defined in (1), and all the leads are weighted using Table 3.

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 4 shows the distributions of the MBDs, stratified by the presence/absence of Left Bundle Branch Block in one of representative case over the 20 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 (9) for $i = 1, \ldots, n$. Considering both signals and derivatives, only MDBs of the signals come out 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 (9) is missing because we do not have additional covariates in our dataset.
The model (9) reduces then to

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

The estimates of parameters in model (13) are reported in Table 4.

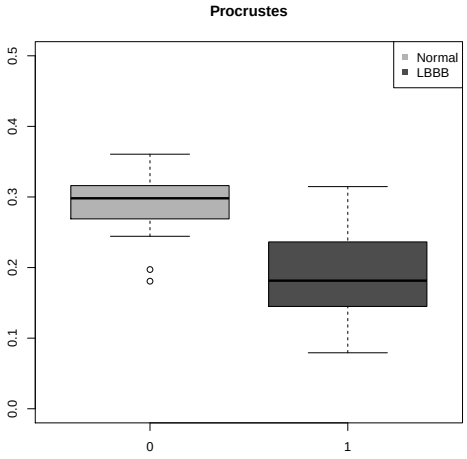

Figure 4: Distributions of MBDs, stratified by the presence of disease, for the data.

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 4: Estimates, standard errors and p-values for the parameters of the logistic model.

The confusion matrix obtained comparing the true and the estimated labels of the patients is reported in Table 5. We set the threshold for the classification carried out by the logistic model in (13) equal to 0.5.

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

Table 5: Confusion matrix.

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 best performing one in term of correct classification rate.

As in the simulation study (case B), the results are satisfactory. We obtain a good correct classification rate and high values of sensitivity and specificity. 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 through 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 new way for handling complexity of multivariate functional data. 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] Berrendero, J.R., Justel A., Svarc M. (2011), Principal components for multivariate functional data. *Computational Statistics and Data Analysis*, **55**, 9, 2619–2634.

[2] Biasi, R., Ieva, F., Paganoni, A.M., Tarabelloni, N. (2013), An efficient framework for computational functional data analysis. In progress.

[3] Claeskens, G., Hubert, M., Slaets, L., Vakili, K. (2013), Multivariate functional halfspace depth, *Journal of the American Statistical Association*, in press.

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

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

[6] Ieva, F., Paganoni, A.M. (2013b), Risk Prediction for Myocardial Infarction via Generalized Functional Regression Models, *Statistical Methods in Medical Research*. Forthcoming. DOI: 10.1177/0962280213495988

[7] Ieva, F., Paganoni, A.M., Pigoli, D., Vitelli, V. (2013), Multivariate Functional Clustering for the Morphological Analysis of ECG Curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **62** (3), 401-418.

[8] 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

[9] Li, J., Liu, R. (2004), New Nonparametric Tests of Multivariate Locations and Scales using Data Depth, *Statistical Science*, **19**, 686-696.

[10] Liu, R., Singh, K. (1993), A Quality Index Based on Data Depth and Multivariate Rank Tests, *Journal of the American Statistical Association*, **88**, (421), 252–260.

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

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

[13] 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

[14] Ramsay, J.O. Silverman, B.W. *Functional Data Analysis* (2nd ed.), Springer, New York, 2005.

[15] Tukey, J. (1975). Mathematics and the picturing of data. In Proceedings of the International Congress of Mathematicians, volume 2, pages 523531, Vancouver.

MOX Technical Reports, last issues

Dipartimento di Matematica "F. Brioschi",
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

54/2013 Biasi, R.; Ieva, F.; Paganoni, A.M.; Tarabelloni, N.
Use of depth measure for multivariate functional data in disease prediction: an application to electrocardiographic signals

53/2013 Micheletti, S.
A continuum variational approach based on optimal control to adaptive moving mesh methods

52/2013 Chen, P.; Quarteroni, A.; Rozza, G.
Multilevel and weighted reduced basis method for stochastic optimal control problems constrained by Stokes equations

51/2013 Chen, P.; Quarteroni, A.
Weighted reduced basis method for stochastic optimal control problems with elliptic PDE constraint

47/2013 Chkifa, A.; Cohen, A.; Migliorati, G.; Nobile, F.; Tempone, R.
Discrete least squares polynomial approximation with random evaluations - application to parametric and stochastic elliptic PDEs

50/2013 Antonietti, P.F.; Verani, M.; Zikatanov, L.
A two-level method for Mimetic Finite Difference discretizations of elliptic problems

49/2013 Micheletti, S.
Fast simulations in Matlab for Scientific Computing

48/2013 Simone Palamara, Christian Vergara, Elena Faggiano, Fabio Nobile
An effective algorithm for the generation of patient-specific Purkinje networks in computational electrocardiology

45/2013 Sangalli, L.M.; Secchi, P.; Vantini, S.
Analysis of AneuRisk65 data: K-mean Alignment

46/2013 Marron, J.S.; Ramsay, J.O.; Sangalli, L.M.; Srivastava, A.
Statistics of Time Warpings and Phase Variations