MOX–Report No. 56/2012

Risk Prediction for Myocardial Infarction via Generalized Functional Regression Models

Ieva, F.; Paganoni, A.M.

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

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

Risk Prediction for Myocardial Infarction via Generalized Functional Regression Models

Francesca Ieva[♯] and Anna Maria Paganoni[♯]

December 21, 2012

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

francesca.ieva@mail.polimi.it, anna.paganoni@polimi.it

Keywords: multivariate functional data; ECG signals; generalized linear models.

Abstract

In this paper, we propose a generalized functional linear regression model for a binary outcome indicating the presence/absence of a cardiac disease with a multivariate functional data among the relevant predictors. In particular the motivating problem is an analysis of Electrocardiographic (ECG) traces of patients whose prehospital 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. The statistical analysis starts with a preprocessing step of ECGs, treated as multivariate functional data. They are reconstructed from noisy observations, then the biological variability is removed by a nonlinear registration procedure based on landmarks. Thus, a Multivariate Functional Principal Component Analysis (MFPCA) is carried out on the variance-covariace matrix of the reconstructed and registered ECGs as well as of their first derivatives, in order to perform a data-driven dimensional reduction. The scores of the principal components that result to be significant are then used within a generalized functional regression model, together with other standard covariates of interest. Hence, a new semi-automatic diagnostic procedure is proposed to model the probability of disease (in the case of interest, the probability of being affected by Left Bundle Brunch Block) and to classify patients. Finally, the performance of this classification method is evaluated through cross validation and compared with other methods proposed in literature.

1 Introduction

The use of telemedicine systems in prehospital emergency rescues has allowed diagnoses for patients with cardiovascular ischaemic diseases to be performed more rapidly. The literature has shown that prehospital ECG reduces treatment times and in-hospital mortality (see Canto et al., 1997; Ting et al., 2008 and Diercks et al., 2009 among others) and it also suggests that prehospital ECG may either be transmitted for interpretation by hospital staff or can be interpreted locally by paramedics, who then communicate their diagnosis to the hospital (see Brown et al., 2008; Trivedi et al., 2009).

Starting from 2006, in the Milanese urban area a working group collecting 23 Cardiology Units and the 118 Dispatch Center (the Italian free-toll number for emergencies), performed monthly data collections twice a year on all patients admitted to any hospital in Milano with coronary artery disease (MOMI^2: MOnth MOnitoring Myocardial Infarction in MIlan survey). The statistical analysis of the collected data (see Ieva and Paganoni 2010, Grieco et al., 2012a, 2012b) confirmed the time of first ECG teletransmission as the most important factor to guarantee a quick access to an effective treatment for patients. Then, since 2008, a project named PROMETEO (PROgetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall'Extra Ospedaliero) has been started with the aim of spreading the intensive use of ECG as prehospital diagnostic tool and of constructing a new database of ECGs with features never recorded before in any other data collection on heart diseases. Thanks to the partnerships of Azienda Regionale Emergenza Urgenza (AREU), Abbott Vascular and Mortara Rangoni Europe s.r.l., ECG recorder with GSM transmission have been installed on all Basic Rescue Units (BRUs) of Milanese urban area.

The principal aim of this work is the development of a new semi-automatic diagnostic procedure for classification of the ECG signals generated by telemedicine equipment of the BRUs.

In Ieva et al. (2013), an identification, from a statistical perspective, of specific ECG patterns which could benefit by an early invasive approach has been performed, on a sample of data arising from PROMETEO database. In fact, the identification of statistical tools capable of classifying curves using their sole shape could support an early detection of coronary disease, not based on usual clinical criteria. In order to do this, in Ieva et al. (2013) ECG traces are considered as a noisy multivariate functional data. A real time procedure consisting of preliminary steps like reconstructing signals, wavelets denoising and removing biological variability in the signals through data registration has been tuned and tested. Then a multivariate functional k-means has been considered, thus simultaneously clustering all 8 leads of each patient. This classification procedure uses group centroids as reference signals. The technique proposed in Ieva et al. (2013) allowed diagnoses to be consistent with clinical practice, starting from purely statistical considerations. Anyway, despite the attractiveness of this

method, the estimation of the number of groups as well as their identification might not be straightforward.

In this work we approach the problem in a different way: we aim at constructing and validating a statistical procedure to model the binary outcome of interest (i.e., the presence of cardiovascular acute ischaemic event) by means of suitable covariates (i.e., patients characteristics, whenever available) and of multivariate functional predictors (i.e., the ECG signal available for each patient). In particular, we focus our attention on estimating the probability to belong to Left Bundle Branch Block (LBBB) group, using as predictor the 8-leads ECG trace of each patient and its first derivative, which are inserted in a suitable generalized functional regression model. Specifically, following for example James (2002), Ratcliffe et al. (2002), Escabias et al. (2004), Müller and Stadtmüller (2005) and Zhu and Cox (2009), we perform a dimensionality reduction by a Multivariate Functional Principal Component Analysis (MFPCA, see Ramsay and Silvermann, 2005), summarizing the information carried out by the covariance matrices of the signals and their first derivatives by the corresponding scores, obtained projecting data and derivatives on the corresponding Karhunen-Loève bases. Then we introduce the scores into the generalized regression model where the response is the Bernoulli variable indicating the presence of LBBB. A consequent classification of patients as well as a comparison with previous results are proposed and discussed.

The paper is then structured as follows: Section 2 contains the theoretical framework of the MFPCA we adopt for carrying out dimensional reduction of the multivariate functional data (§2.1) and the corresponding first derivatives (§2.2). As we said above, this step is performed in order to point out relevant components to be inserted in the generalized regression model for risk prediction of the presence of the disease (§2.3). In Section 3 the analysis of ECG data arising from PROMETEO dataset is presented, together with the cross validation analysis aimed at testing the robustness of the procedure. Finally, in Section 4 conclusions are drawn and further developments are discussed.
All the analyses are carried out using `R` statistical software (see R Development Core Team, 2009).

2 Models and Methods

A common strategy to deal with complex or high-dimensional data is to perform a dimensional reduction (see Ramsay and Silverman, 2005). In the motivating example we consider, the 8-leads ECG signal of each patient (a multivariate functional curve) is considered as a predictor of the presence of LBBB, then we deal with the dimensional reduction of such data and their first derivatives in order to input them in a generalized regression model for predicting the risk of LBBB.

2.1 Multivariate Functional Principal Component Analysis

Also in the functional setting, Principal Components Analysis (PCA) provides a way of looking at covariance structure of data that can be much more informative and can complement, or even replace altogether, a direct examination of the variance-covariance function, as detailed in Ramsay and Silverman (2005).

Let $\mathbf{X}$ a stochastic process with law P taking values on the space $L^2(I;\mathbb{R}^h)$ of square integrable functions $\mathbf{X}(t) = (X_1(t), \ldots, X_h(t))^T : I \to \mathbb{R}^h$, where I is a compact interval of $\mathbb{R}$. 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

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

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

$$(\mathcal{V}\mathbf{g})(s) = \int_I V(s,t)\mathbf{g}(t)dt,$$

The kernel $V(s,t)$ is defined by

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

where $\otimes$ is a outer product in $\mathbb{R}^h$. $V(s,t)$ is a $h \times h$ matrix, whose elements will be denoted as $V_{rq}(s,t)$, for $r,q = 1, ..., h$.
In what follows, the model formulation is already intended for the application of interest, where the ECG of each patient $j = 1, ..., n$ is a 8-variate functional data generated by the stochastic process $\mathbf{X}$ taking values on the Hilbert space $L^2(I;\mathbb{R}^8)$. The general case of $h \geq 2,\ h \neq 8$ follows straightforwardly. We consider, as data reduction strategy, the Multivariate Functional Principal Component Analysis (MFPCA) proposed in Ramsay and Silverman (2005).
So let $V_{rr}(s,t)$, $r = 1, \ldots, 8$, be the variance functions of the components of $\mathbf{X}$, as well as $V_{rq}(s,t)$, $r,q = 1, \ldots, 8,\ r \neq q$ the cross covariance functions. Thus, for any $(s,t) \in I \times I$, $V_{rq}(s,t) = \mathrm{Cov}(X_r(s), X_q(t))$, $r,q = 1, \ldots, 8$.
Consider the usual scalar product between two elements $\mathbf{U}$ and $\mathbf{W}$ in $L^2(I;\mathbb{R}^8)$

$$\langle \mathbf{U}, \mathbf{W} \rangle = \sum_{r=1}^{8} \int_I U_r(t) W_r(t) dt. \tag{1}$$

Call $\mathbf{e}^k(t) = (e_1^k(t), \ldots, e_8^k(t))^T$ the $k-$element of the Karhunen-Loève expansion, that is the solution of the eigenequation system

$$\begin{aligned}
\int_I V_{11}(s,t)e_1^k(t)dt + \cdots + \int_I V_{18}(s,t)e_8^k(t)dt &= \rho^k e_1^k(s), \\
&\vdots = \vdots \\
\int_I V_{81}(s,t)e_1^k(t)dt + \cdots + \int_I V_{88}(s,t)e_8^k(t)dt &= \rho^k e_8^k(s).
\end{aligned}$$

The eight-leads ECGs $\{(\mathbf{X})_i\}$, $(i = 1, ..., n)$, are a sample from $\mathbf{X}$. Eigenfunction-eigenvalue couples $\{(\mathbf{e}^k, \rho^k)\}_{k\in\mathbb{N}}$ completely explain modes of variation in the data, in the sense that eigenfunctions represent orthonormal directions of decreasing variability with respect to the explained variances expressed by the corresponding eigenvalues. Thanks to the basis expansion given by principal components, it is possible to represent data using just the first K elements of $\{\mathbf{e}^k\}_{k\in\mathbb{N}}$, the linear combination of which is, by construction, a good approximation for the original curves. The interpretation of eigenvalues as variances is useful also to determine a criterion to choose the most relevant modes. Since $\sum_{k=1}^{K} \rho_k$ represents variance captured by the first K components, we can choose K so that the proportion of variance described by these components is higher than a given threshold c, i.e.,

$$\frac{\sum_{k=1}^{K} \rho_k}{\sum_{k=1}^{m} \rho_k} \geq c, \tag{2}$$

where m is the number of abscissa values on which functional data are known, which is an upper bound to the number of components that can be estimated. In the analysis of data, as literature advises, we deal with centered and scaled data, that is:

$$\mathbf{Z}(t) = (Z_1(t), \ldots, Z_8(t))^T = \left(\frac{X_1(t) - \mu_1(t)}{\sqrt{V_{11}(t,t)}}, \ldots, \frac{X_8(t) - \mu_8(t)}{\sqrt{V_{88}(t,t)}} \right)^T.$$

2.2 Derivatives Refinements

The problem with discrete and noisy observations, is amplified when the interest focuses also on data derivatives. In our case, since the information on the presence of the disease is carried out not only by morphological changes we observe on the original signals, but also by changes that happen in their first derivatives, it is even more necessary to smooth data in a suitable way. In fact, the smoothing procedure is essential not only for an accurate reconstruction of data, but also for a proper estimate of their derivatives (see Ieva et al., 2013 for deeper discussion of such arguments and comparison of different derivatives' computations). Moreover, since the eight ECG leads of interest (I, II, V1, V2, V3, V4, V5 and V6) jointly describe the complex heart dynamics, the smoothing technique should take into account simultaneously all the components of the multivariate functional data (i.e., the leads).

Among possible smoothing methods, wavelet bases seem suitable for smoothing our data because every basis function is localized both in time and in frequency and is therefore able to capture strongly localized ECG features (peaks, oscillations,...). Details of the wavelets smoothing applied to ECG data can be found in Pigoli and Sangalli (2012). The procedure proposed there is able to take jointly into account the multi-dimensionality of the data, obtaining smoothed

estimates of the 8-dimensional curves of the ECGs. It has also the advantage of providing an estimate of the curves derivatives, which is straightforward when functional reconstruction is obtained via a basis expansion: each derivative can be obtained simply by a linear combination of the corresponding basis function derivatives.

Once we get the smoothed data and the corresponding first derivatives, we carry out the same dimensionality reduction also on the variance-covariance matrix of the first derivatives, in order to take into account also the variability of the first derivatives when modelling the risk prediction, according to the procedure described in §2.1. So doing we really take advantage of the functional nature of data.

2.3 Generalized regression with multivariate functional predictors

We consider now a logistic regression model, where the response variable is $Y_i \sim Be(p_i)$ for $i \in 1, \ldots, n$ and $\theta_i = \log(p_i/(1-p_i))$. We model θ_i as linear transformation of the covariates related to $i-$th statistical unit.

$$\theta_i = \int_I \boldsymbol{\delta}^T(t)\mathbf{Z}_i(t)dt + \int_I \boldsymbol{\delta}_d^T(t)\mathbf{Z}_i^{'}(t)dt + \sum_{h=1}^{q} d_{ih}\gamma_h \tag{3}$$

being $\mathbf{Z}_i(t)$ the centered and scaled multivariate functional data concerning the i-th statistical unit, and $\mathbf{Z}_i^{'}(t)$ the corresponding first derivatives. The vector $\mathbf{d}_i = (d_{i1}, ..., d_{iq})^T$, $\mathbf{d}_i \in \mathbb{R}^q$, for $i = 1, \ldots, n$, contains the traditional covariates that are possibly available for the i-th statistical unit. Moreover, $\boldsymbol{\delta}(t) : I \mapsto \mathbb{R}^8$ and $\boldsymbol{\delta}_d(t) : I \mapsto \mathbb{R}^8$ are 8-variate functional parameters to be estimated as well as $\boldsymbol{\gamma} \in \mathbb{R}^q$ is a vector of parameters to be estimated.
Thanks to the dimensional reduction driven by the selection of the number of MFPC basis obtained using (2), the linear predictor in (3) may be approximated with the following expression:

$$\int_I \sum_{k=1}^{K} \xi_i^k \boldsymbol{\delta}(\boldsymbol{t})^T \mathbf{e}^k(t)dt + \int_I \sum_{k=1}^{K_d} \tilde{\xi_i^k} \boldsymbol{\delta}(\boldsymbol{t})_d^T \mathbf{e}_d^k(t)dt \sum_{h=1}^{q} d_{ih}\gamma_h,$$

where $\xi_i^k = \langle \mathbf{Z}_i, \mathbf{e}^k \rangle$, $\tilde{\xi_i^k} = \langle \mathbf{Z}_i^{'}, \mathbf{e}_d^k \rangle$, and $\{\mathbf{e}_d^k\}_{k\in\mathbb{N}}$ is the basis expansion given by principal components of derivatives. We can represent also $\boldsymbol{\delta}(t)$ and $\boldsymbol{\delta}_d(t)$ using the correspondent Karhunen-Loève expansions, i.e., $\boldsymbol{\delta}(t) = \sum_{l=1}^{K} \zeta^l \mathbf{e}^l$, where $\zeta^l = \langle \boldsymbol{\delta}, \mathbf{e}^l \rangle$ and $\boldsymbol{\delta}_d(t) = \sum_{l=1}^{K_d} \zeta_d^l \mathbf{e}_d^l$, where $\zeta_d^l = \langle \boldsymbol{\delta}_d, \mathbf{e}_d^l \rangle$. Thanks to the orthonormality

of $\{\mathbf{e}^k\}_{k\in\mathbb{N}}$ and of $\{\mathbf{e}_d^k\}_{k\in\mathbb{N}}$ we obtain

$$\begin{aligned}\theta_i &= \sum_{k=1}^{K}\xi_i^k\zeta^k + \sum_{k=1}^{K_d}\tilde{\xi}_i^k\zeta_d^k + \sum_{h=1}^{q}d_{ih}\gamma_h \\ &= \underbrace{\boldsymbol{\xi}_i^T\boldsymbol{\zeta}}_{\substack{data\\contribution}} + \underbrace{\tilde{\boldsymbol{\xi}}_i^T\boldsymbol{\zeta}_d}_{\substack{derivatives\\contribution}} + \underbrace{\mathbf{d}_i^T\boldsymbol{\gamma}}_{\substack{covariates\\contribution}}, \quad i=1,\dots,n.\end{aligned}$$

The model is then reduced to a classical logistic regression, in which the unknowns are represented by the parameters $\boldsymbol{\zeta} = (\zeta^1,\dots,\zeta^K)^T$, $\boldsymbol{\zeta}_d = (\zeta_d^1,\dots,\zeta_d^{K_d})^T$ and $\boldsymbol{\gamma} = (\gamma_1,\dots,\gamma_q)^T$. The same approach can be extended without further difficulties to the more general context of generalized linear models with different responses and link functions.

3 An application to ECG signals

In Ieva et al. (2013), a statistical framework for analysis and classification of ECG curves starting from their sole morphology is proposed. The main goal of that paper is to identify, from a statistical perspective, specific ECG patterns which could benefit from an early invasive approach. In fact, the identification of statistical tools capable of classifying curves using their shape only could support an early detection of heart failures, not based on usual clinical criteria.

The basic statistical unit is the 8-variate function, 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 derivative. There, the 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 LBBB is diagnosed, and 0 if the trace is considered as physiological. The sample from the PROMETEO database we analyse consists of the ECG signals of $n = 149$ subjects, among which 101 are Normal and 48 are affected by LBBB. The main reason for the sample size being relatively small is the incidence of LBBB among all the possible kind of acute myocardial infarction, that is not so high (about 3%). So, among all the diagnosis records we inspected, only fews were eligible. In addiction, among these, only a small part can be retained as "pure LBBB" diagnoses, since it often happens that LBBB arises together with other comorbidities (say arrythmias like Atrial Fibrillation, Atrioventricular Block, Atrial Flutter, Paroxysmal and Supraventricular tachycardia, etc.). Since we want modification on morphological variations in ECGs to be induced only by the presence of LBBB, at least in the training set we compute the MFPC basis upon, excluding traces where LBBB was not the only diagnosis. So doing, we avoided a priori the biases carried by the presence of other comorbidities.

The idea of the analysis presented here is to set up a generalized regression model able to discriminate between patological and physiological traces, explaining the disease probabiity by means of multivariate functional predictors, i.e.,

ECG signals and their first derivatives. The contribution of the multivariate curves is summarized through the dimensional reduction carried out by MFPCA of the covariance matrices of signals and signals' first derivatives proposed in §2.1 and §2.2, in order to take advantage of the functional nature of the data.

In practice, we deal with a noisy and discrete observation of the function describing the ECG trace of each patient. We use the wavelet based smoothing technique for multivariate curves proposed in Pigoli and Sangalli (2012) to obtain the smoothed estimates of 8-dimensional ECG signals and their first derivative. Moreover, since each patient has his own "biological" time, the same event of the heart dynamics may happen at different times for different patients. Since the morphological change due to this difference in timings is misleading from a statistical perspective, we need to register data. It is well known that a correct separation between the different kinds of variability is necessary for a successful analysis (see Ramsay and Silverman, 2005). In particular, as detailed in Ieva et al. (2013), we adopt a registration procedure based on landmarks, which are points of the curve that can be associated with a specific biological time. Five of these landmarks (P_{onset}, QRS_{onset}, QRS_{offset}, T_{onset}, T_{offset}) are provided by Mortara-Rangoni procedure. They identify, for each patient $i = 1, ..., n$, the P wave, the QRS complex and the T wave, i.e., the main segments and waves of the ECG signal. We add one more landmark: the R peak identified on the lead I (I^i_{peak}). We choose the time point identified on this lead as representative for all the leads because only on the lead I both the physiological and pathological ECG traces present a clearly identifiable R peak. Then, since all the leads capture the same heart dynamics, biological time must be the same.

Figure 1 shows denoised and registered data we consider for our analysis. The black solid lines represent the mean functions. Figure 2 shows the corresponding first derivatives. Again the black solid lines represent the mean functions.

We shall now select the components of the MFPC to be considered in the subsequent analysis, both for ECG signals and their first derivatives. In both cases, we choose the first K and K_d components of data and derivative's basis respectively, such that their associated eigenvalues explain a proportion of variance equal to 70%.

Among these, we retained only the first principal components, using the corresponding scores as covariates. The scores are computed projecting data and first derivative on the first elements of the corresponding MFPC basis. We retained only the first MFPC essentially for two reasons: both a stepwise selection based on the AIC as well as a Brier's score minimization criterium selected the scores on the first and tenth components as the most useful ones to explain the illness probability, but since it is known that the efficiency of the estimates of the eigenfunctions and of the corresponding scores is decreasing with respect to the index of eigenfunctions, we decided to focus only on the first ones. Moreover, the results of the risk prediction obtained with the parimonious choice of the first MFPC only remain very robust with respect to those obtained considering

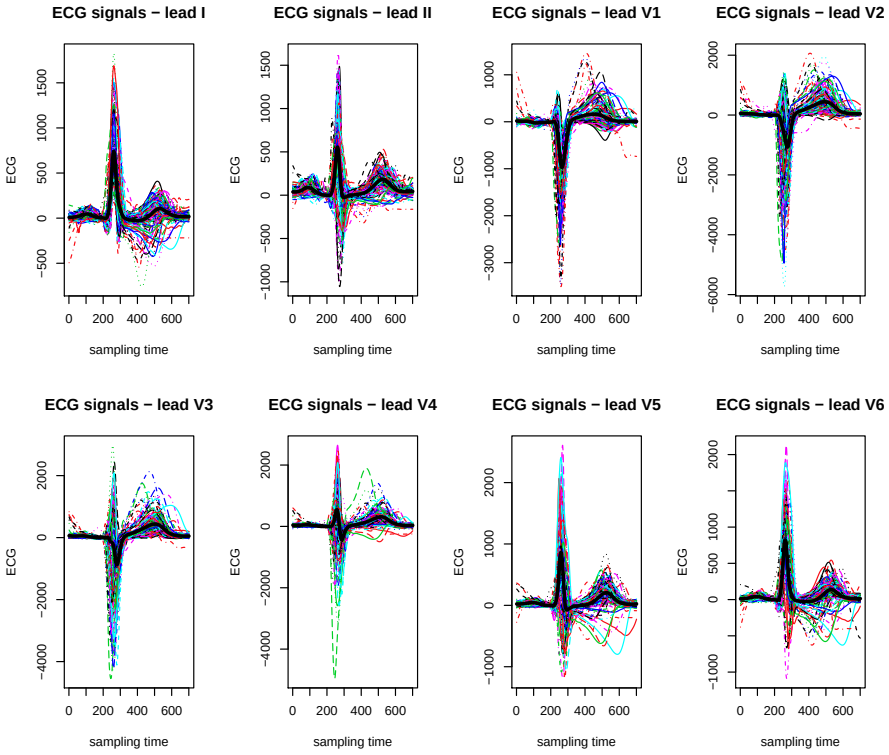

Figure 1: Denoised and registered data (8 leads) for the 149 patients with superimposed the mean functions (black solid lines).

more than one MFPC (as it will be detailed in the following). The scores of the first principal component are then the only ones identified as statistically significant for the generalized regression model, both for the original data and the first derivatives. Figure 3 shows the distributions of the first principal components scores, for the data (left panel) and the first derivatives (right panel) respectively, stratified by the presence/absence of LBBB. The p-values of Wilcoxon tests carried out to compare the distributions of the scores are less than $2 * e^{-16}$ in both cases.
So we fitted the following logistic model: for $i = 1, \ldots, n$,

$$\theta_i = \gamma_0 + \xi_i^k \zeta^1 + \tilde{\xi_i^1} \zeta_d^1 \tag{4}$$

It arises from model (4), where $K = K_d = 1$ and no further patient's covariates are available. The model output is reported in Table 1.

Figure 4 shows the first multivariate functional principal component of the original data. Sample means of each lead are plotted (solid lines), together with two curves obtained by adding (+) and subtracting (-) a suitable multiple C of the principal component. As suggested in Ramsay and Silverman (2005), we

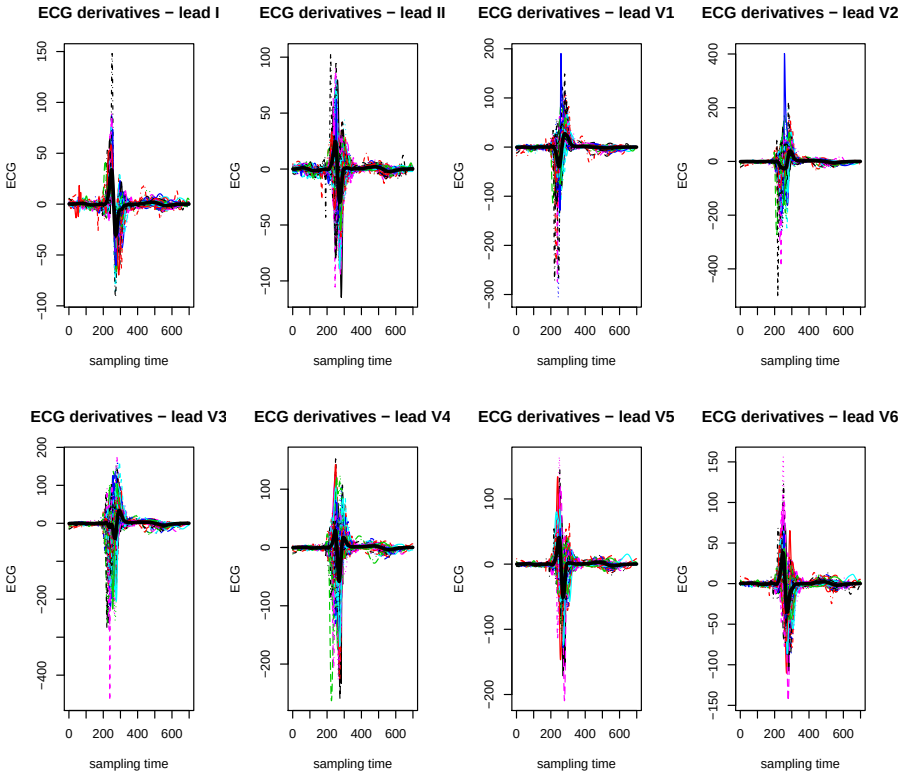

Figure 2: First derivatives (8 leads) for the 149 patients with superimposed the mean functions (black solid lines).

set C as 0.2 times the root-mean-square difference between the estimated mean $(\hat{\mu}_1(t), \ldots, \hat{\mu}_8(t))$ and its overall time average, i.e., $\bar{\mu} = \frac{1}{8}\sum_{i=1}^{8}\int_I \hat{\mu}_i(t)dt$. ¿From Figure 4 it is clear that the first functional principal component speaks about the morphological variability expressed by specific segments of the ECG. These morphological changes are particulary marked in the ST-segment (the part of the ECG curve usually including among time interval between 350 and 600 ms), which in fact is the most useful part of the ECG, apart from the QRS complex, to carry out the LBBB diagnosis, as confirmed by the cardiologists.
The confusion matrix obtained comparing the true and the estimated label of the patients is reported in Table 2. We set the threshold for the classification carried out by the logistic model in (4) equal to 0.5. The mean Cross Validation error of the logistic model is 3.6%, that is not far from the error committed in diagnosing LBBB by physicians, if both false positives and false negatives are considered.

We propose this method as an automatic diagnostic tool to predict the risk

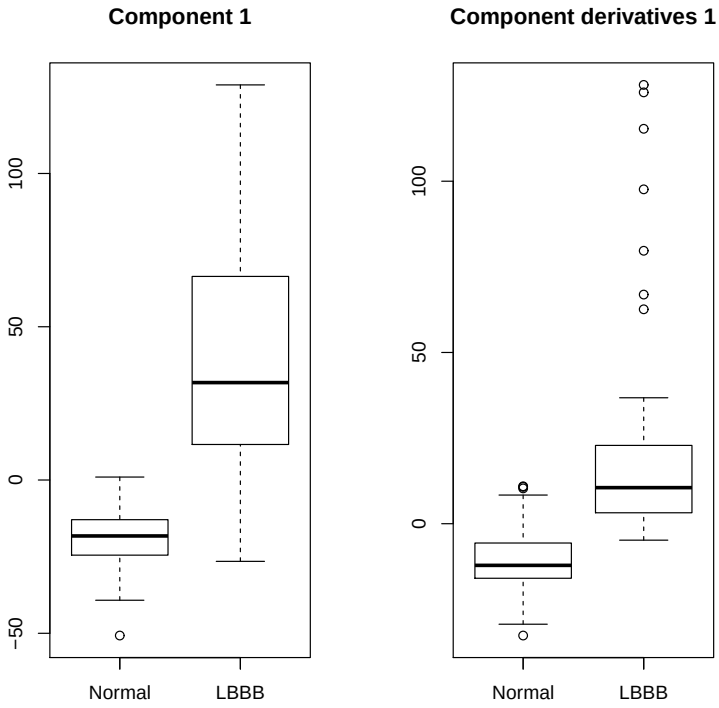

Figure 3: Distributions of first principal components scores, stratified by the presence of disease, for the original data (left panel) and the first derivatives (right panel).

also for new patients entering the study. In fact, although the MFPC basis is a data-driven basis, we check method robustness through a leave-j-out simulation study: we randomly choose a subsample of j patients ($j = 1, 5, 10, 20$), we perform the MFPC analysis and fit the logistic model on the remaining $n - j$ patients, obtaining the estimation of the number of basis components to be retained and the coefficients. This is what we will refer to as the "off-line" step. Then we projects the j isolated ECGs on the basis previously pointed out ("on-line" step), in order to get a real-time computation of the scores corresponding to the new j data and their first derivatives, and the estimated probability of disease. We repeat the experiment 500 times. In Table 3 are reported the mean Actual Error Rate (AER) over the 500 simulations, and the corresponding standard deviation.

In general, the idea is the following: once a reliable and representative dataset of N ECGs is pointed out according to clinical best practice, the procedure we propose computes the "off-line step" described above on the N multivariate curves, selecting a suitable number of components for data and first derivatives

Parameter	Estimate	Std. Error	p-value
γ^0 (Intercept)	-0.07148	0.53112	0.892938
ζ^1 (First PC)	0.16941	0.04695	0.000308
ζ_d^1 (First PC deriv.)	0.16304	0.06352	0.010262

Table 1: Estimates, standard errors and p-values for the parameters of the logistic regression model.

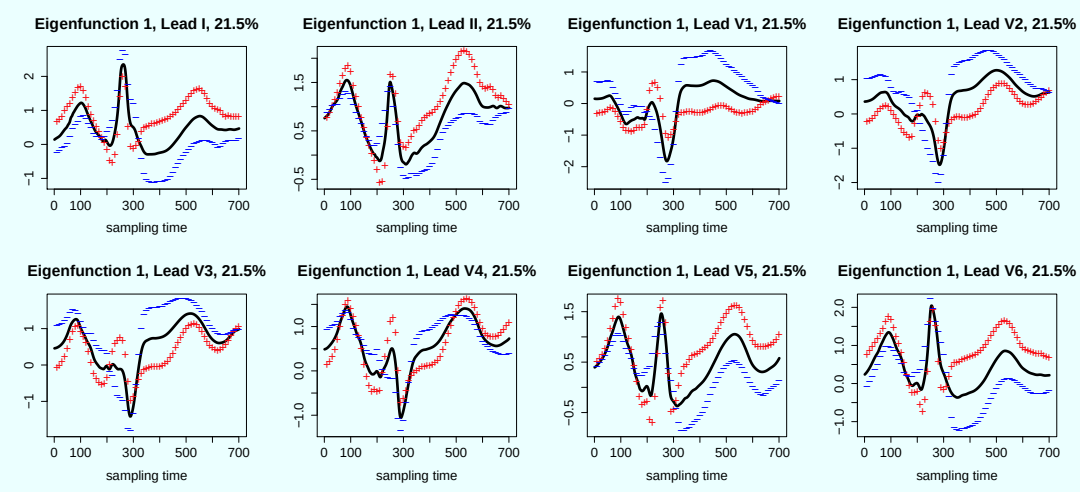

Figure 4: First multivariate functional principal component.

basis and providing the coefficients for the generalised regression model. Then, as long as new patients enter in the study, the semi-automatic diagnosis tool projects their ECGs on eigenfunctions selected in the off-line basis and plugs in the scores estimates as predictors of the logistic model for estimating LBBB risk of the new patients.

4 Conclusions

In this paper, we propose a generalized functional linear regression model for a binary outcome indicating the presence/absence of a cardiac disease, with a multivariate functional data among the relevant predictors. This is an example of data to be necessarily treated in the multivariate functional context; this framework despite its evident interest is quite rarely treated in statistical literature.

The principal aim of this work is then the development of a new semi-automatic diagnostic procedure for classification of the ECG signals generated by telemedicine equipment of the Basic Rescue Units (BRUs). In fact, we set up a framework for carrying out semi-automatic diagnosis of LBBB, starting from

	Normal	LBBB
Classified as Normal	100	4
Classified as LBBB	1	44

Table 2: Confusion matrix.

	Mean	Standard Deviation
$j = 1$	0.062	0.241
$j = 5$	0.047	0.088
$j = 10$	0.055	0.068
$j = 20$	0.056	0.0491

Table 3: Mean and Standard Deviation of AER.

the statistical analysis of the sole curve morphology. The method we propose is then aimed at supporting decisions of people the basic rescue units are equipped by and at identifying specific ECG patterns which could benefit by an early invasive approach, performing a real-time diagnosis.

In particular, we focus our attention on estimating the probability to belong to Left Bundle Branch Block (LBBB) group, using as predictor the 8-leads ECG trace of each patient and its first derivative, which are inserted in a suitable generalized functional regression model. Specifically, we perform a dimensionality reduction by a Multivariate Functional Principal Component Analysis (MFPCA), summarizing the information carried out by the covariance matrices of the signals and their first derivatives by the corresponding scores, obtained projecting data and derivatives on the corresponding Karhunen-Loéve bases. Then we introduce the scores into a generalized regression model where the response is the Bernoulli variable indicating the presence of LBBB. We finally carry out the consequent classification of patients as well as a check for robustness of our method. To this aim, we are actually trying to robustify the estimation method for regression parametera through the use of a wider dataset of numerically simulated ECGs.

The innovative aspect of this paper lies in developing advanced statistical methods aimed at detecting pathological ECG traces (in particular, LBBB), starting only from morphological features of the curves. This allows for diagnoses that are consistent with clinical practice, starting from purely statistical considerations. Further extensions of this work consist to enlarge the spectrum of acute cardiovascular diseases this technique can be applied to. Owing to the extreme generality of the method, this generalization is theoretically straightforward.

Aknowledgements

This work is part of PROMETEO (PROgetto sull'area Milanese Elettrocardiogrammi

Teletrasferiti dall'Extra Ospedaliero). The authors wish to thank 118 Dispatch Centre of Milan. Data are provided by Mortara Rangoni Europe s.r.l..

References

[1] Brown, J.P., Mahmud, E., Dunford, J.V., et al. (2008). Effect of prehospital 12-lead electrocardiogram on activation of the cardiac catheterization laboratory and door-to-balloon time in ST-segment elevation acute myocardial infarction. *American Journal of Cardiology*, 101: 158–161

[2] Canto, J.G., Rogers, W.J., Bowlby, L.J., French, W.J., Pearce, D.J., Weaver, W.D. (1997). The prehospital electrocardiogram in acute myocardial infarction: is its full potential being realized? National Registry of Myocardial Infarction 2 Investigators. *Journal of the American College of Cardiology*, 29(3): 498–505

[3] Diercks, D.B., Kontos, M.C., Chen, A.Y., Pollack, C.V., Wiviott, S.V., Rumsfeld, J.S., Magid, D.J., Gibler, B., Cannon, C.P., Peterson, E.D., Roe, M.T. (2009). Utilization and Impact of prehospital Electrocardiograms for Patients With Acute ST-Segment Elevation Myocardial InfarctionData From the NCDR (National Cardiovascular Data Registry) ACTION (Acute Coronary Treatment and Intervention Outcomes Network) Registry FREE. *Journal of the American College of Cardiology*, 53(2): 161–166

[4] Escabias, M., Aguilera, A.M., Valderrama, M.J. (2004). Principal component estimation of functional logistic regression: discussion on two different approaches. *Nonparametric Statistics*, 16: 365–384

[5] Grieco, N., Ieva, F. and Paganoni, A.M. (2012a). Performance assessment using mixed effects models: a case study on coronary patient care. *IMA Journal of Management Mathematics*, 23(2): 117–131

[6] Grieco, N., Corrada, E., Sesana, G., Ieva, F., Paganoni, A.M., Marzegalli, M. (2012b). Mortality and ST resolution in patients admitted with STEMI: the MOMI survey of emergency service experience in a complex urban area. *European Heart Journal: Acute Cardiovascular Care*, 1(3): 192-199

[7] Ieva, F. and Paganoni, A.M. (2010). Multilevel models for clinical registers concerning STEMI patients in a complex urban reality: a statistical analysis of MOMI^2 survey. *Communications in Applied and Industrial Mathematics*, 1(1): 128–147

[8] Ieva, F., Paganoni, A.M., Pigoli, D., Vitelli, V. (2013). Multivariate functional clustering for the analysis of ECG curves morphology. *Journal of the Royal Statistical Society - Series C*, To appear. doi:10.1111/j.1467-9876.2012.01062.x

[9] James, G.M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society - Series B*, 64: 411–432

[10] Müller, H.G., Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics*, 33(2): 774–805

[11] Pigoli, D. and Sangalli, L.M. (2012). Wavelets in functional data analysis: Estimation of multidimensional curves and their derivatives. *Computational Statistics and Data Analysis*, 56: 1482–1498

[12] 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

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

[14] Ratcliffe, S.J., Heller, G.H., Leader, L.R. (2002). Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression. *Statistics in Medicine*, 21: 1115–1127

[15] Ting, H.H., Krumholz, H.M., Bradley, E.H., Cone, D.C., Curtis, J.P., Drew, B.J., Field, J.M., French, W.J., Gibler, W.B., Goff, D.C., Jacobs, A.K., Nallamothu, B.K., O'Connor, R.E., Schuur, J.D. (2008). Implementation and integration of prehospital ECGs into systems of care for acute coronary syndrome. *Circulation*, 118: 1066–1079

[16] Trivedi, K., Schuur, J.D. and Cone, D.C. (2009). Can paramedics read ST-segment elevation myocardial infarction on prehospital 12-lead electrocardiograms?. *Prehospital Emergency Care*, 13: 207–214.

[17] Zhu, H., Cox, D.D. (2009). A functional generalized linear model with curve selection in cervical pre-cancer diagnosis using fluorescence spectrosopy. *Lecture Notes-Monograph Series*, Optimality: The Third Erich L. Lehmann Symposium, 57: 173–189

MOX Technical Reports, last issues

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

56/2012 Ieva, F.; Paganoni, A.M.
Risk Prediction for Myocardial Infarction via Generalized Functional Regression Models

55/2012 Peng Chen, Alfio Quarteroni, Gianluigi Rozza
Uncertainty quantification of the human arterial network

54/2012 Ettinger, B., Perotto, S.; Sangalli, L.M.
Spatial regression models over two-dimensional manifolds

53/2012 Fumagalli, A.; Scotti, A.
An efficient XFEM approximation of Darcy flows in fractured porous media

52/2012 Perotto, S.
Hierarchical model (Hi-Mod) reduction in non-rectilinear domains

51/2012 Beck, J.; Nobile, F.; Tamellini, L.; Tempone, R.
A quasi-optimal sparse grids procedure for groundwater flows

50/2012 Carcano, S.; Bonaventura, L.; Neri, A.; Esposti Ongaro, T.
A second order accurate numerical model for multiphase underexpanded volcanic jets

49/2012 Migliorati, G.;Nobile, F.;von Schwerin, E.;Tempone, R.
Approximation of Quantities of Interest in stochastic PDEs by the random discrete L2 projection on polynomial spaces

48/2012 Ghiglietti, A.; Paganoni, A.M.
Statistical properties of two-color randomly reinforced urn design targeting fixed allocations

47/2012 Astorino, M.; Chouly, F.; Quarteroni, A.
Multiscale coupling of finite element and lattice Boltzmann methods for time dependent problems