MOX–Report No. 04/2011

Multivariate functional clustering for the analysis of ECG curves morphology

Francesca Ieva, Anna Maria Paganoni,
Davide Pigoli, Valeria Vitelli

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

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

Multivariate functional clustering for the analysis of ECG curves morphology

Francesca Ieva[♯], Anna Maria Paganoni[♯], Davide Pigoli[♯], Valeria Vitelli[♯]

January 14, 2011

[♯] 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,
davide.pigoli@mail.polimi.it, valeria.vitelli@mail.polimi.it

Keywords: ECG signal, Wavelets smoothing, Functional registration, Functional k-means clustering.

Abstract

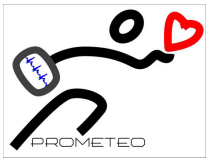
Cardiovascular diseases are one of the main causes of death all over the world. In this kind of pathologies, it is fundamental to be well-timed in order to obtain good prognosis in reperfusive treatment. In particular, an automatic classification procedure based on statistical analyses of tele-transmitted ECG traces would be very helpful for an early diagnosis. This work is a pilot analysis on electrocardiographic (ECG) traces (both normal and pathological ones) of patients whose 12-leads pre-hospital ECG has been sent by life supports to 118 Dispatch Center of Milan. The statistical analysis consists of preliminary steps like reconstructing signals, wavelets denoising and removing the biological variability in the signals through data registration. Then, a multivariate functional k-means clustering of reconstructed and registered ECGs is performed, and performances of classification method are validated. So a semi-automatic diagnostic procedure, based on the sole ECG's morphology, is proposed to classify patients and predict pathologies.

1 Introduction

Cardiovascular ischemic diseases are nowadays one of the main causes of death all over the world. Every year 160.000 persons are affected by an heart failure and 50.000 persons suddenly die for heart attack. In Italy, they are responsible of 44% of overall deaths. Cardiovascular diseases also call for the most part of emergency rescue operations. In fact, almost all events which require rescue operations to the 118 Milan Dispatch Center (the Italian free toll number for emergencies) are classified as "medical

events" and concern cardiovascular system. In case of Coronary Arteries ischemic disease, it is fundamental to be well-timed in order to obtain good prognosis in reperfusive treatment. This result can be obtained only with pre, inter and intra-hospital networks well organized and synchronized.

Since 2001, a working group collecting 23 Cardiology Units of Milan Area and 118 Dispatch Center has been activated on Milan urban area. Starting from 2006, this group perform monthly data collection twice a year on all patients admitted to any hospital belonging to the Milan Cardiological Network with coronary artery disease, stratifying them on mode of admission to Emergency Room (ER), i.e. self-presented, delivered by Basic or Advanced Life Supports with or without tele-ECG transmission. From the analysis of these data, time of first ECG tele-transmission has been pointed out as the most important factor to guarantee a quick access to an effective treatment for patients. The quicker ECG is performed, the higher is the probability of good reperfusion of the treatment the patient undergoes (see Antman et al., 2009; Ieva and Paganoni, 2010; Grieco et al., 2007, 2010).



Then, since 2008, a project has been started with the aim of spreading the intensive use of ECG as pre-hospital diagnostic tool and of constructing a new database of ECGs with features never recorded before in any other data collection on heart diseases. In fact, anticipating diagnostic time, reducing infarction complications and optimizing the number of hospital admissions are the three main goals of PROMETEO (PROgetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall' Extra Ospedaliero). 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 Life Supports (BLSs) of Milan urban area. Up to the start of PROMETEO, just Advanced Life Supports (ALSs) were able to make and send ECG from territory, because physicians were carried on them. To make this project possible, intensive training courses have been carried out to more than 3500 rescuers working on BLSs. Thanks to PROMETEO, planned and realized by 118 Dispatch Center of Milan, it is possible to send quickly the ECG from territory to 118 Dispatch Center itself, and then to the hospital where patient will be admitted to, even when a BLS is sent to the patient.

Persuading people to call 118 Dispatch Center when needed, equipping all Milan life supports with ECG recorder and training rescuers to acquire ECG correctly to all people which call 118 for rescue, regardless of symptoms declared, is the way to strongly reduce delays in treatments and then in reperfusion. In fact this could be the way to obtain early diagnosis and then quicker delivery of patients from territory to Intensive Cardiac Care Units (ICCU) of Hospitals, i.e. a better service for patients affected by Acute Coronary Syndromes (ACS), enabling them to avoid to spend time in the ER and to go directly to Cath-Lab or to Percutaneous Coronary Intervention (PCI).

In this work we analyse a pilot database composed by 48 ECG traces extracted from PROMETEO datawarehouse. Selection criteria will be explained below. The whole datawarehouse contains all ECG traces recorded on Milan urban area by basic life supports since the end of 2008 up to now. Most of them are pathological, showing acute arrhythmias as well as atypical modifications.

Each file contained in PROMETEO datawarehouse is in correspondence to three sub-files. The first one is called *details* and contains technical information, useful for signal processing and analysis, such as times of waves' repolarization and depolarization, landmarks indicating onset and offset instances of main ECG's curves and segments, and automatic diagnoses, established by Mortara-Rangoni VERITAS^{TM} algorithm. We used these automatic diagnoses to label ECG traces we analyzed, in order to validate our unsupervised clustering algorithm's performance. The challenge of this work, in fact, consists of tuning and testing a real time procedure which enables semi automatic diagnosis of the patients' disease based only on ECG traces morphology, then not dependent on clinical evaluations. The second sub-file is called *Rhythm* and contains the ECG signal sampled for 12 seconds (10000 sampled points). The third one is called *Median*. It is built starting from *Rhythm* file, and depicts a *reference* beat lasting 1.2 seconds (1200 points). We carried out the analysis considering only the *Median* files, obtaining 8 curves (one for each lead, see Section 2.1 for details) for each patient, which represents his/her "Median" beat for that lead. Examples of *Rhythm* and *Median* files of a patient are reported in Fig. 1 and 2 respectively.

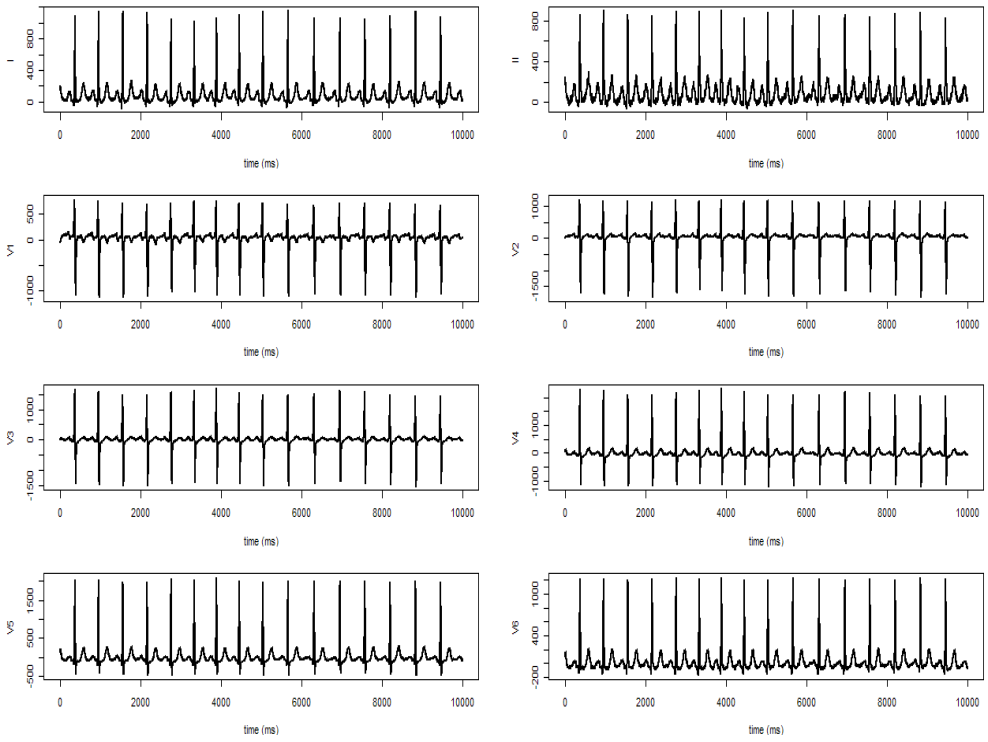


Figure 1: An example of file *Rhythm*.

The main goal of this work is then 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 starting from their shape only could support an early detection of heart failures, not based on usual clinical criteria. To this aim, it is extremely important to understand the link between cardiac physiology and ECG trace shape. As detailed in following sections, we focus on Physiological

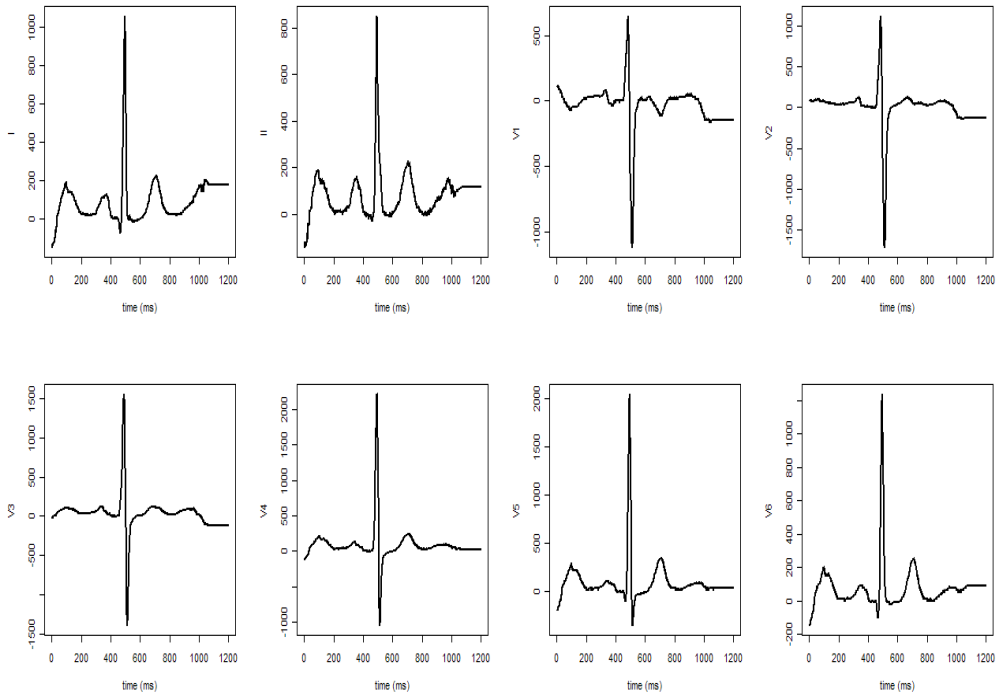


Figure 2: An example of file *Median*.

traces in contrast to Left and Right Bundle Branch Block (LBBB and RBBB respectively) traces. Bundle Branch Block is a cardiac conduction abnormality seen on the ECG. In this condition, activation of the left (right) ventricle is delayed, which results in the one ventricle contracting later than the other.

Details on Bundle Branch Blocks and their connection with non-physiological shape of ECG signal will be treated in Section 2, where also clinical details about ECG signals will be given together with an overview of the pilot database. Procedures of wavelet smoothing and landmarks registration performed on ECG traces are explained in Section 3. In Section 4 data analysis is presented, consisting of a functional k-means clustering of QT-segments of smoothed and registered ECG traces. Finally, in Section 5 results of analysis are discussed, and further developments to be explored in future works are proposed.

2 Electrocardiography and Bundle Branch Block

Before starting with technical details about statistical data analysis, a brief introduction to ECG signal and to electrophysiology of Bundle Branch Block is presented, in order to better understand features of data which will be analyzed in the following.

2.1 The ECG signal

Electrocardiography is a transthoracic recording of the electrical activity of the heart over time captured and externally recorded through skin electrodes. The ECG works mostly by detecting and amplifying the tiny electrical changes on the skin that are caused when the heart muscle "depolarises" during each heart beat (for further inquiry about clinical details, see Lindsay, 2006).

First attempts of measuring ECG signals date back to Willem Einthoven (see Einthoven, 1908; Einthoven et al., 1950). The Einthoven *limb leads* (standard leads) are illustrated in Fig. 3 and are defined in the following way:

$$\text{Lead I: } V_I = \Phi_L - \Phi_R, \quad \text{Lead II: } V_{II} = \Phi_F - \Phi_R, \quad \text{Lead III: } V_{III} = \Phi_F - \Phi_L;$$

where

V_I = voltage of Lead I
V_{II} = voltage of Lead II
V_{III} = voltage of Lead III
Φ_L = potential at the left arm
Φ_R = potential at the right arm
Φ_F = potential at the left foot

These lead voltages satisfy the following relationship:

$$V_I + V_{III} = V_{II}, \qquad (1)$$

hence only two of these three leads are independent. The lead vectors associated with Einthoven's lead system are conventionally found based on the assumption that the heart is located in an infinite, homogeneous volume conductor (or at the center of a homogeneous sphere representing the thorax). If the position of the right arm, left arm, and left foot are at the vertices of an equilateral triangle, having the heart located at its center, then the lead vectors also form an equilateral triangle. A simple model results from assuming that the cardiac sources are represented by a dipole located at the center of a sphere representing the thorax, hence at the center of the equilateral triangle. With these assumptions, the voltages measured by the three limb leads are proportional to the projections of the electric heart vector on the sides of the lead vector triangle. The voltages of the leads are obtained from Equation (1).

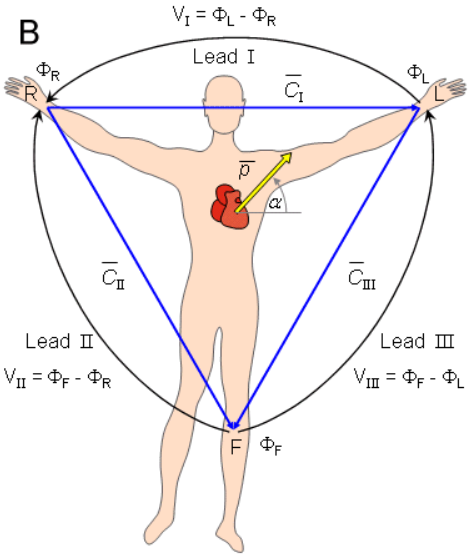


Figure 3: Eithofen limb leads

Nowadays, the most commonly used clinical ECG-system, the 12-lead ECG system, consists of the following 12 leads, which are: I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, V6. The main reason for recording all 12 leads is that it enhances pattern recognition. In fact, this combination of leads gives the clinician an opportunity to compare the projections of the resultant vectors in two orthogonal planes and at different angles, as shown in Fig. 4 and explained in Goldberger (1942a; 1942b); Mason and Likar (1966) and Wilson et al. (1944).

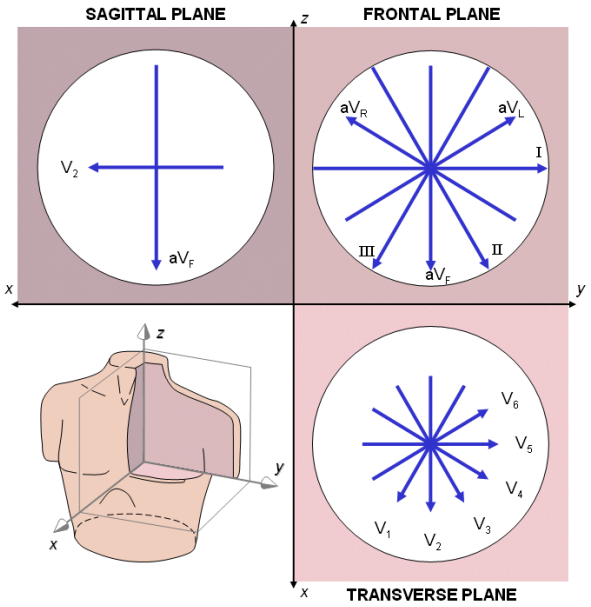


Figure 4: The projections of the lead vectors of the 12-lead ECG system in three orthogonal planes when one assumes the volume conductor to be spherical homogeneous and the cardiac source centrally located.

Of these 12 leads, the first six are derived from the same three measurement points. Therefore, any two of these six leads include exactly the same information as the other four. So, the ECG traces analyzed in the following sections will consist of leads I, II, V1, V2, V3, V4, V5 and V6 only.

The file *Rhythm* of our dataset represents the output of an ECG recorder. From this curve it is possible to trace a representative heartbeat for each patient. As we said before, this is the content of file *Median*, which consists of 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 baseline voltage of the electrocardiogram is known as the isoelectric line and is represented by the PR segment.

Fig. 5 shows a scheme of the typical shape of a physiological single beat, recorded on ECG graph paper; main relevant points, segments and waves are highlighted. The figure also includes definitions of segments and intervals in the trace. Deflections in this signal are denoted in alphabetic order starting with the letter P, which represents atrial depolarization. The ventricular depolarization causes the QRS complex, and repolarization is responsible for the T-wave. Atrial repolarization occurs during the QRS complex and produces such a low signal amplitude that it cannot be detected, with the exception of physiological ECGs (see Scher and Young, 1957).
Finally, in Fig. 6, the connection between relevant waves and segments of ECG trace and mechanic activation of heart is illustrated.
Interpretation of the ECG relies on the idea that different leads "view" the heart from different angles. This has two benefits. Firstly, leads which are showing anomalies due to the presence of pathology (for example Right or Left Bundle Branch Blocks, see Section 2.2) can be used to infer which region of the heart is affected. Secondly,

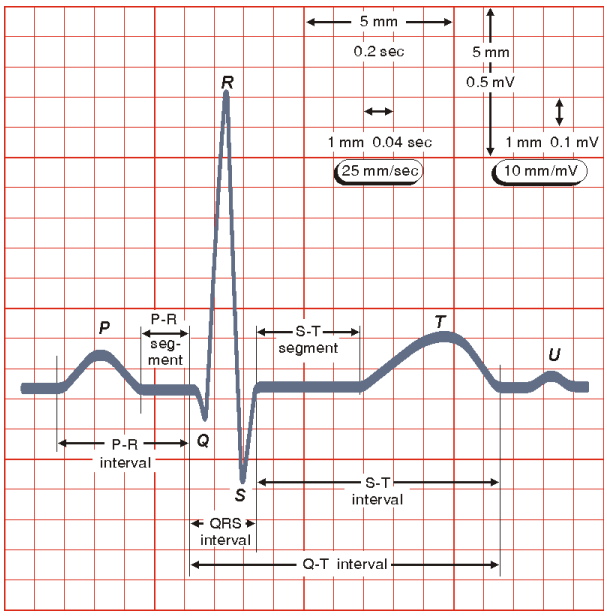


Figure 5: Scheme of the typical shape of a physiological single beat, recorded on ECG graph paper. Main relevant points, segments and waves are highlighted.

the overall direction of travel of the wave of depolarization (which can reveal other problems) can also be inferred. This direction is named the *heart electrical axis*.

2.2 Bundle Branch Blocks

The heart's electrical activity begins in the sinoatrial node (the heart's natural pacemaker, n.1 in Fig. 7), which is situated on the upper right atrium. The impulse travels next through the left and right atria and summates at the AV node (n.2 in Fig. 7). From the AV node the electrical impulse travels down the Bundle of His (n.3 in Fig. 7) and divides into the right and left bundle branches (n.4 and 10 in Fig. 7). The right bundle branch contains one fascicle. The left bundle branch subdivides into two fascicles: the left anterior fascicle and the left posterior fascicle (n.4 and 5 in Fig. 7). Ultimately, the fascicles divide into millions of Purkinje fibres which in turn interdigitise with individual cardiac myocytes, allowing for rapid, coordinated, and synchronous physiologic depolarization of the ventricles.
When a bundle branch or fascicle becomes injured (due to underlying heart disease, myocardial infarction, or cardiac surgery), it may cease to conduct electrical impulses appropriately. This results in altered pathways for ventricular depolarization. Since the electrical impulse can no longer use the preferred pathway across the bundle branch, it may move instead through muscle fibers in a way that both slows the electrical movement and changes the directional propagation of the impulses. As a result, there is a loss of ventricular synchrony, ventricular depolarization is prolonged, and there may be a corresponding drop in cardiac output.

From a clinical perspective a bundle branch block can be diagnosed when the duration of the QRS complex on the ECG exceeds 120 ms. A right bundle branch block typically causes prolongation of the last part of the QRS complex, and may shift the

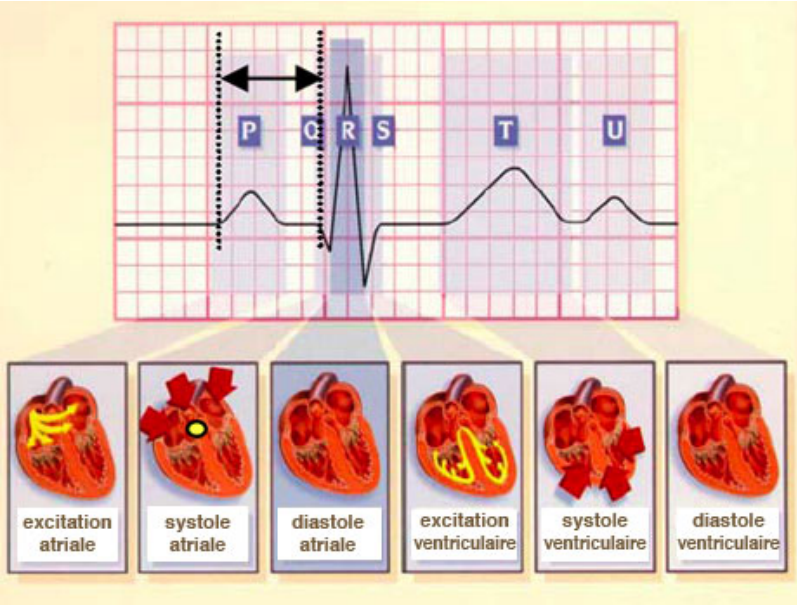


Figure 6: Correspondence between waves in ECG signal and cardiac phases.

heart's electrical axis slightly to the right. The ECG will show a terminal R wave in lead V1 and a slurred S wave in lead I. Left bundle branch block widens the entire QRS, and in most cases shifts the heart's electrical axis to the left. The ECG will show a QS or RS complex in lead V1 and a monophasic R wave in lead I. Another usual finding with bundle branch block is appropriate T wave discordance: this means that the T wave will be deflected opposite the terminal deflection of the QRS complex. Unfortunately, some individuals will exhibit both left and right bundle branch blocks and have a profoundly abnormal QRS interval. This degree of electrical degradation to the myocardium may lead to Ventricular Dyssynchrony, reflected in shape modification of ECG curves, as described above. From a statistical point of view, instead, we will focus our analysis on shape modifications induced on the ECG trace by the effect of the pathology, and we will investigate these shape modifications only in a statistical perspective, i.e. not using clinical criteria to classify ECGs. The exploitation of these morphological modifications in the clustering procedure will be the focus of the following Sections.

3 Data smoothing and registration

As mentioned above, the aim of this work is exploring ECG curves morphology. Thus, the basic statistical unit is the multivariate function which describes heart dynamics, for each patient, on the eight significant leads. However, in practice we have only noisy and discrete observation of this function. Moreover, each patient has his own "biological" time, i.e. the same event of the heart dynamics may happen at different time measurements for different patients: this is only misleading from a morphological point of view. These two problems are common in Functional Data Analysis applications and they can be addressed respectively with data smoothing and registration (see Ramsay and Silverman, 2005).

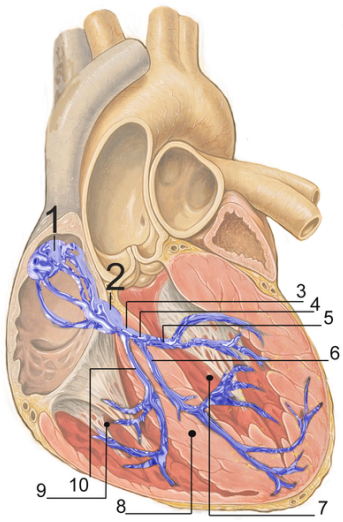


Figure 7: Conduction system of the heart: 1. Sinoatrial node; 2. Atrioventricular node; 3. Bundle of His; 4. Left bundle branch; 5. Left posterior fascicle; 6. Left-anterior fascicle; 7. Left ventricle; 8. Ventricular septum; 9. Right ventricle; 10. Right bundle branch.

3.1 Wavelets smoothing

The first step of the statistical analysis consists in data smoothing starting from noisy measurements; the choice of the functional basis is crucial. Wavelet bases seem suitable for our data because every basis function is localized both in time and in frequency, being therefore able to capture ECG strong localized features (peaks, oscillations...). Every basis function is therefore identified by two indices: the second one identifies time position of the basis function, while the first one indicates the level of wavelet decomposition and corresponds to the frequency. In particular we use a Daubechies wavelet basis with 10 vanishing moments (see Daubechies, 1988 for details).

The wavelet smoothing procedure is illustrated schematically in Fig. 8. The first step consists in changing over to wavelet domain and estimating basis coefficients. True wavelet coefficients are estimated starting from the empirical wavelet coefficients, computed by Discrete Wavelet Transform (DWT) of the original data. To use DWT, it is necessary that the number of observations is a power of two. Thus, in the further analysis we use only the central $2^{10} = 1024$ observation points. There is no loss of significant information: the region on which we focus the analysis contains all the important features of the ECG trace.

Since the eight leads traces (i.e. I, II, V1, V2, V3, V4, V5 and V6) capture the same physical signal, we expect that every significant feature will be reflected on all leads. Therefore, in the attempt to separate the true functional signal from measurement noise we choose a smoothing procedure which takes into account the multivariate nature of the data. Thus, we resort to the method proposed in Pigoli and Sangalli (2010) to obtain the estimate of the vectorial function

$$\mathbf{f}(t) = (I(t), II(t), V1(t), V2(t), V3(t), V4(t), V5(t), V6(t)).$$

for each of the $n = 48$ patients in the database.

Let $\{\mathbf{w}_k \in \mathbb{R}^8; k = 1,\ldots,2^{10}\}$ be a noisy and discrete observation of the 8-dimensional ECG trace $\mathbf{f}$ on a grid of 2^{10} equispaced points. Assume that these data are generated by the model

$$\mathbf{w}_k = \mathbf{f}(t_k) + \boldsymbol{\varepsilon}_k \quad k = 1,\ldots,2^{10}, \tag{2}$$

where the error $\boldsymbol{\varepsilon}_k$ has multivariate normal distribution with mean $\mathbf{0} \in \mathbb{R}^8$ and variance-covariance matrix $\sigma^2\mathbb{I}_8$. Our goal is to accurately estimate the 8 - dimensional curve $\mathbf{f}$. We thus consider the corresponding model on the space of wavelet coefficients. Thanks to the orthogonality of the wavelet transform, this is given by

$$\mathbf{d}_{j,k} = \mathbf{d}^0_{j,k} + \boldsymbol{\rho}_{j,k}, \tag{3}$$

with $\mathbf{d}_{j,k}, \mathbf{d}^0_{j,k}, \boldsymbol{\rho}_{j,k} \in \mathbb{R}^8$, where

$$\mathbf{d}_{j,k} = (d^{I_i}_{j,k}, d^{II_i}_{j,k}, d^{V1_i}_{j,k}, d^{V2_i}_{j,k}, d^{V3_i}_{j,k}, d^{V4_i}_{j,k}, d^{V5_i}_{j,k}, d^{V6_i}_{j,k})$$

are the vectors of the empirical wavelet coefficients corresponding to the data, $\mathbf{d}^0_{j,k}$ are the vectors of the true wavelet coefficients of $\mathbf{f}$ and $\boldsymbol{\rho}_{j,k}$ are the wavelet transforms of the noise, having multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\sigma^2_d\mathbb{I}_8$. The estimation is based on a soft thresholding approach: vector of empirical coefficients is considered coming from noise if its square euclidian norm is below a threshold $t_8 = \hat{\sigma}^2_d(3\log(2^{10}))$. The standard deviation $\hat{\sigma}_d$ is estimated using the median of the absolute deviation from the median (MAD) on wavelet coefficients of level 9, which are supposed to be pure noise (see e.g. Donoho et al., 1995). If the square norm of coefficients vector exceeds this threshold, a shrinkage is applied. Thus, estimated coefficients vector will be

$$\hat{\mathbf{d}}_{j,k} = \left(1 - \frac{\sqrt{t_8}}{||\mathbf{d}_{j,k}||_2}\right)_+ \mathbf{d}_{j,k}\,. \tag{4}$$

Finally, estimated functional coordinates of $\mathbf{f}$ can be obtained through wavelet reconstruction of $L^2(\mathbb{R})$. For details on this smoothing procedure, see Pigoli and Sangalli (2010).

Fig. 9 shows raw data and functional estimates obtained with this wavelet smoothing procedure for a normal subject. Observations are now in a functional form and thus we can use functional data analysis techniques.

3.2 Landmark registration

Functional observations usually show both phase and amplitude variation, i.e. each curve has its own biological time so that same features can appear at different times among the patient. It is well known that a correct separation between these two kind of variability is necessary for a successful analysis (see Ramsay and Silverman, 2005). We address this problem through 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 are provided by Mortara-Rangoni measurement procedure and identify the P wave (P_{onset}, P_{offset}), QRS complex (QRS_{onset}, QRS_{offset}) and T wave (T_{offset}). We add one more landmark corresponding with the R peak on the I lead ($Ipeak$). We choose this

$$I_{ik}, II_{ik}, V1_{ik}, V2_{ik}, V3_{ik}, V4_{ik}, V5_{ik}, V6_{ik}$$
$$\text{for } i = 1, \ldots, n, \quad \text{for } k = 1, \ldots, 2^J$$

$\Downarrow$ **Discrete Wavelet Transform**

$$d_{j,k}^{I_i}, d_{j,k}^{II_i}, d_{j,k}^{V1_i}, d_{j,k}^{V2_i}, d_{j,k}^{V3_i}, d_{j,k}^{V4_i}, d_{j,k}^{V5_i}, d_{j,k}^{V6_i}$$
$$\text{for } i = 1, \ldots, n, \quad \text{for } j = 1, \ldots, J, \quad \text{for } k = 1, \ldots, 2^j$$

$\Downarrow$ **Coefficients Estimation**

$$\widehat{d}_{j,k}^{I_i}, \widehat{d}_{j,k}^{II_i}, \widehat{d}_{j,k}^{V1_i}, \widehat{d}_{j,k}^{V2_i}, \widehat{d}_{j,k}^{V3_i}, \widehat{d}_{j,k}^{V4_i}, \widehat{d}_{j,k}^{V5_i}, \widehat{d}_{j,k}^{V6_i}$$

$\Downarrow$ **Functional Reconstruction**

$$I_i(t), II_i(t), V1_i(t), V2_i(t), V3_i(t), V4_i(t), V5_i(t), V6_i(t)$$
$$\text{for } i = 1, \ldots, n$$

Figure 8: Steps of the wavelet smoothing procedure. In our case $n = 48$ (number of patients), $J = 10$ (since we have $2^{10} = 1024$ points for each track).

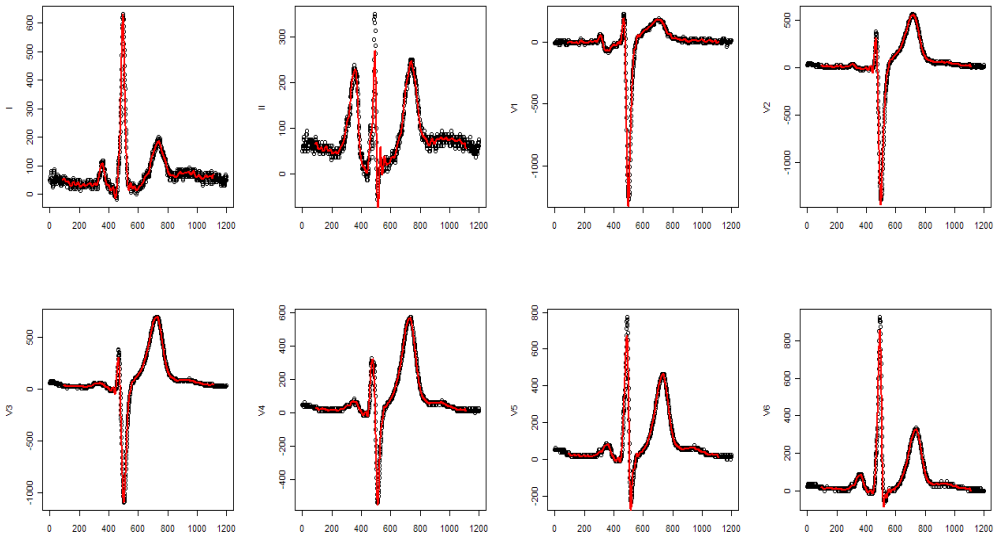


Figure 9: Raw data of the eight leads (black points) and wavelet functional estimates (red) for a normal subject.

landmark because only on the I lead both normal and pathological ECG traces present a clearly identifiable R peak. Since all the leads capture the same heart dynamics,

biological time must be the same. Thus, these landmarks can be used to register all the leads. For each patient i we look for a warping function h_i such that

$$h_i(P_{onset}) = P^0_{onset} \qquad h_i(P_{offset}) = P^0_{offset}$$

$$h_i(QRS_{onset}) = QRS^0_{onset} \quad h_i(Ipeak) = Ipeak^0$$

$$h_i(QRS_{offset}) = QRS^0_{offset} \quad h_i(T_{offset}) = T^0_{offset}$$

where $P^0_{onset}, P^0_{offset}, QRS^0_{onset}, Ipeak^0, QRS^0_{offset}, T^0_{offset}$ are the mean values of the correspondent landmarks (see Table 2). We solve this problem using spline interpolation of order 3. Thus, the registered vectorial function will be

$$\mathbf{F}_i(t) = \mathbf{f}_i(h_i(t)).$$

for every patient $i = 1, \ldots, 48$. Fig. 10 shows both unregistered and registered I leads for all the 48 patients.

The registration procedure separates morphological information (i.e. amplitude variability) and duration of ECG intervals (i.e. phase variability). The former is captured by the registered ECG traces, while the latter is described by warping functions, determined by landmarks. In clinical practice the duration of ECG interval and particularly the QRS complex length is the most important parameter to identify pathological situations. However, this kind of information is not able to distinguish among different pathologies, such as Right and Left Bundle Branch Blocks. This can be seen also in our exploratory dataset. If we perform a multivariate 3-means algorithm on interval lengths ($P_{offset} - P_{onset}$, $QRS_{onset} - P_{offset}$, $QRS_{offset} - QRS_{onset}$ and $T_{offset} - QRS_{offset}$), with the aim of identifying the existing 3 groups, we obtain the result shown in Table 1: this method correctly separates physiological traces from pathological ones but it gives no information on the pathology.

For this reason, we focus our analysis on the registered curves, in the attempt to extract other diagnostic information from ECG morphology. The final diagnostic procedure should of course consider both information coming from morphology analysis and from segments lengths.

	Normal	RBBB	LBBB
Cluster 1	25	0	0
Cluster 2	0	9	5
Cluster 3	0	4	5

Table 1: Confusion matrix related to patients disease classification. Results are obtained performing 3-means clustering algorithm on interval lengths.

4 Data analysis

In this section we propose the use of functional data analysis techniques to perform clustering of smoothed and registered ECG traces. Aim of the analysis is the development of a proper classification procedure, able to distinguish the grouping structure

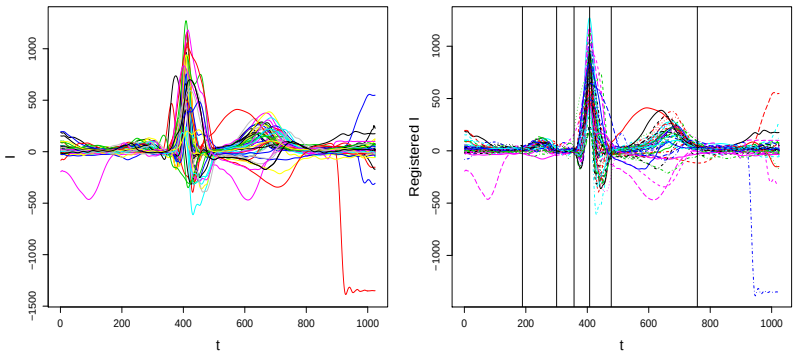


Figure 10: Original I leads for the 48 patients (left) and registered ones (right). Vertical lines indicate landmarks positions.

induced in the sample of ECGs by the presence of different pathologies, on the basis of the sole shape of the considered curves.

4.1 Data selection

As previously discussed in Section 2, ECG traces are very complex functional data, in which different portions of the domain can be analyzed in order to detect different pathologies. The main focus of our analysis stands in the investigation of BBB pathology, which mainly expresses in the ECG trace through a lengthening of the QRS complex and a modification of the T wave (see Section 2.2). In fact, the diagnosis of BBB is not concerned with modifications in P wave, since this portion of the ECG curve deals with cardiac rithm dysfunctions our patients are not affected by. We thus focus our classification analysis on the QT-segment.

In particular, the analyzed dataset consists of the ECG signals of $n = 48$ patients, among which 25 are Normal and 23 are affected by BBB (13 on the right hand side of the heart, and 10 on the left hand side). All the raw ECG traces of these patients have been smoothed and registered according to the procedures described in Section 3.

The landmarks used in the registration procedure are fixed as the mean of the landmarks of all the curves in the dataset; this means that all registered curves show relevant features at the same time points, corresponding to these reference landmarks common to the whole dataset (see Section 3.2): this fact allows us to select, for all the registered curves of the dataset, only the portion of ECG trace belonging to the interval $[P^0_{offset}, T^0_{offset}]$, which is relevant to our diagnostic purposes. The reference values (mean over patients) of landmarks are reported in Table 2, together with the associated standard deviations.

In particular, we select only the portion of

$$\mathbf{F}(t) = \{F^r(t)\}_{r=1}^8 = (I(t),\ II(t),\ V1(t),\ V2(t),\ V3(t),\ V4(t),\ V5(t),\ V6(t))$$

such that $t \in [P^0_{offset}, T^0_{offset}]$, where P^0_{offset} and T^0_{offset} are the values reported in the first line, second and sixth columns of Table 2. The final dataset employed in the subsequent classification analysis is shown in Fig. 11.

	P^0_{onset}	P^0_{offset}	QRS^0_{onset}	$Ipeak^0$	QRS^0_{offset}	T^0_{offset}
mean	188.9	300.2	356.9	407.7	478.0	758.1
standard deviation	38.7	36.7	16.9	17.2	21.3	38.4

Table 2: Landmarks obtained at the end of the registration procedure, as the mean of landmarks of all the curves, and used to select the portion of smoothed and registered ECG curves relevant to our analysis (first line of the table); in the second line, landmarks standard deviations. Landmarks values are referred to a registered time in ms.

4.2 Functional classification

We analyze the n patients according to a functional k-means clustering procedure, in which all the eight leads $\mathbf{F}_i(t) : \mathbb{R} \to \mathbb{R}^8$, for patients $i = 1, \ldots, n$, are simultaneously clustered. To develop this clustering procedure we suppose that $\mathbf{F}_i(t) \in L^2(\mathbb{R}; \mathbb{R}^8)$. Since we consider all the eight leads simultaneously in the analysis, we name the employed clustering procedure *multivariate functional k-means*.

A proper definition of functional k-means procedure and an introduction to its consistency properties can be found in Tarpey and Kinateder (2003). We develop the same k-means procedure, choosing the following distance between ECG traces

$$d(\mathbf{F}_i(t), \mathbf{F}_j(t)) = \sqrt{\sum_{r=1}^{8} \int_{P^0_{offset}}^{T^0_{offset}} (F_i^r(t) - F_j^r(t))^2 dt}, \quad \text{for } i, j = 1, \ldots, n. \tag{5}$$

Note that the measure defined in (5) is the natural distance in the Hilbert space $L^2(\mathbb{R}; \mathbb{R}^8)$.

The k-means clustering algorithm is an iterative procedure, which alternates a step of *centroid calculation*, in which a relevant functional representative (the centroid) for each cluster is identified, and a step of *cluster assignment*, in which all curves are assigned to a cluster, and in particular to the cluster whose centroid is nearer according to the measure in (5). The identification of centroids $\varphi_l(t)$, for $l = 1, \ldots, k$, should find the solution to the following optimization problem

$$\varphi_l(t) = \underset{\varphi \in L^2(\mathbb{R}; \mathbb{R}^8)}{\operatorname{argmin}} \sum_{i: C_i = l} d(\mathbf{F}_i(t), \varphi(t))^2,$$

where C_i is the cluster assignment of the i^{th} patient at the current iteration. The solution to this infinite dimensional optimization problem corresponds (due to the definition of $d(\cdot, \cdot)$) to a functional mean of data belonging to the same cluster. For another implementation of functional k-means algorithm, which integrates registration procedure in the classification steps, see Sangalli et al., (2010); here, instead, we chose to separate the two procedures of registration and clustering, since the latter doesn't use any information beside morphology of the ECG traces, while the former is based on landmarks provided by the Mortara-Rangoni VERITAS$^{\text{TM}}$ algorithm.

Cluster centroids can be obtained via *local polynomial regression* (*loess*, see Kohler (2002) for further details on consistency), with the benefit of keeping the variance of the estimator of the mean constant also at the boundaries of the domain, thanks to the locally varying neighbourhood of data used in the estimation process. More precisely, this technique corresponds to a polynomial regression (we chose polynomial degree $r = 2$), in which the fitting is done locally, since for every point t in the abscissa of the

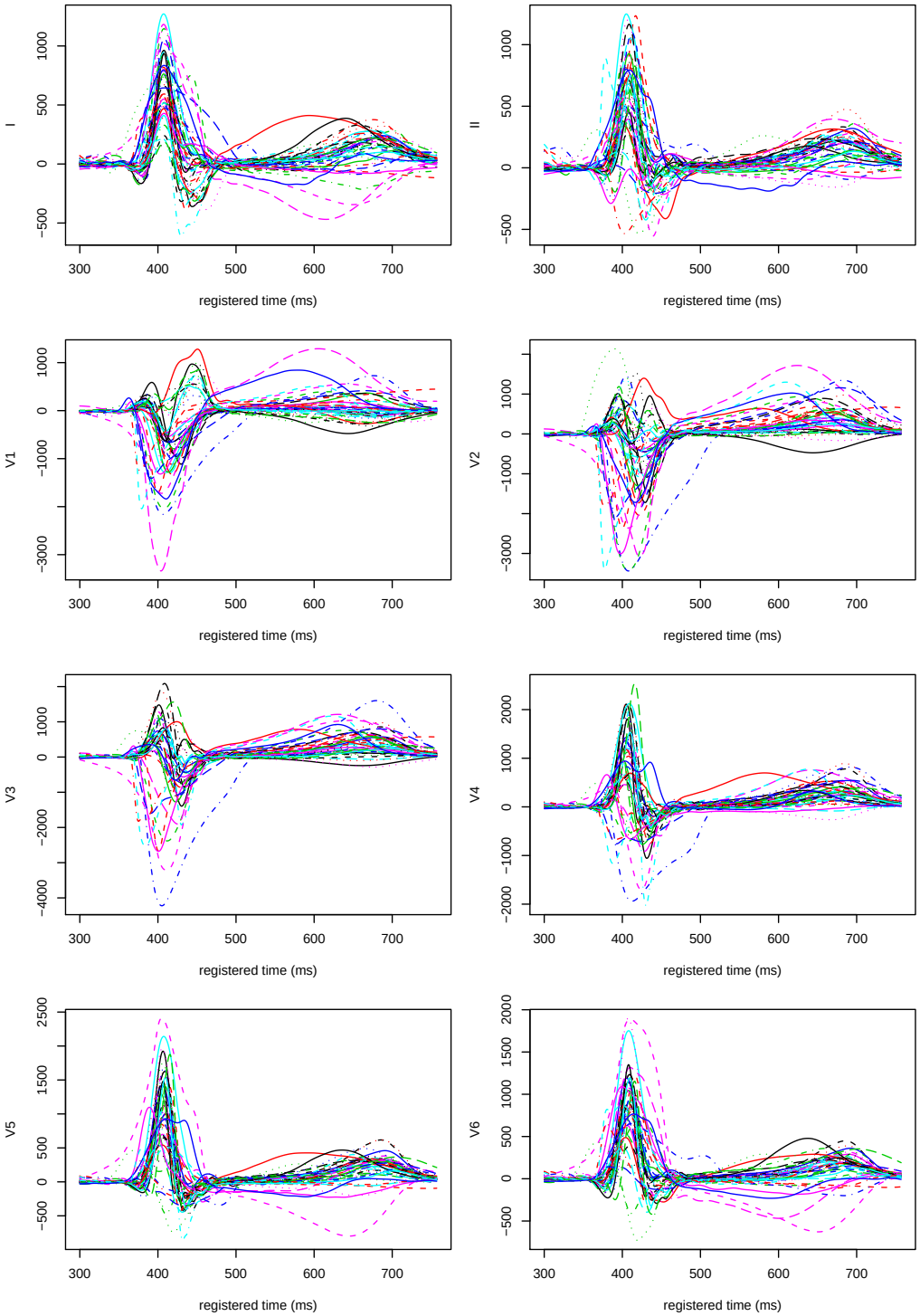


Figure 11: Final dataset used in the classification analysis: 48 smoothed and registered ECG traces, selected over the portion of the abscissa including QRS complex and T wave for each patient; each panel corresponds to a different lead.

curves the fit is made using points in a neighborhood of t, weighted by their distance from t. The size of the neighborhood is controlled by the parameter α, in terms of proportion of points in the abscissa to be considered in the estimation; we set $\alpha = 5\%$.

The k-means clustering procedure clearly depends not only on the choice of the metric $d(\cdot,\cdot)$, but also on the number of clusters k. Being the number of clusters a–priori unknown, we also consider a way to select the optimal number of clusters k^* via a silhouette plot of the final classification (see Struyf et al., 1997). In particular, the silhouette plot of a final classification consists in a bar plot of the *silhouette values* s_i, obtained for each patient $i = 1,\ldots,n$ as

$$s_i = \frac{b_i - a_i}{max\{a_i, b_i\}},$$

where a_i is the average distance, according to (5), of the i^{th} patient to all other patients assigned to the same cluster, while

$$b_i := \min_{l=1,\ldots,k; l\neq C_i} \frac{\sum_{j:C_j=l} d(\mathbf{F}_i(t), \mathbf{F}_j(t))}{\#\{j : C_j = l\}}$$

is the minimum average distance of the i^{th} patient from another cluster. Clearly s_i always lies between -1 and 1, the former value indicating a misclassified patient, while the latter a very well classified one. Note that a patient which alone constitutes a cluster, has silhouette value equal to 1, but he is not considered in the silhouette plot for choosing k^*.

4.3 Results and discussion

In Fig. 12 are shown the final silhouette plots obtained by clustering the sample of 48 ECG traces according to a multivariate functional k-means procedure, and setting $k = 2,3,4,5$. As we can appreciate from the picture, the grouping structure obtained setting $k = 3$ seems the best one, both in terms of silhouette profile, and in terms of wrong assignments. Thus we set $k^* = 3$.

The final classification obtained setting $k = 3$ is shown in Fig. 13, 14 and 15, where the whole functional dataset is shown in lightgray, and only curves assigned respectively to the first, second and third cluster are superimposed in a different color (black, red and green respectively). From inspection of these pictures a different shape of ECGs assigned to different clusters can be immediately appreciated.

Being aware of the different pathologies of the patients included in the sample, we could also analyze the confusion matrix associated to the final cluster assignments, with respect to the Mortara-Rangoni algorithm disease classification (Normal, RBBB and LBBB). The confusion matrix is shown in Table 3; we remark that the final cluster assignments are based on the sole shape of the smoothed and registered ECG curves, analyzed via a unsupervised classification procedure. The results seem appreciable: the final grouping structure traces out quite coherently the patients disease classification, with only few cases wrongly assigned. In particular, we compute *sensitivity* and *specificity* to quantify the effectiveness of the clustering procedure: the two values are respectively 91.3% for sensitivity, and 100% for specificity.

Finally, it seems interesting to visualize the smoothed original ECG trace of a representative for each cluster, so that one can characterize the final grouping structure by

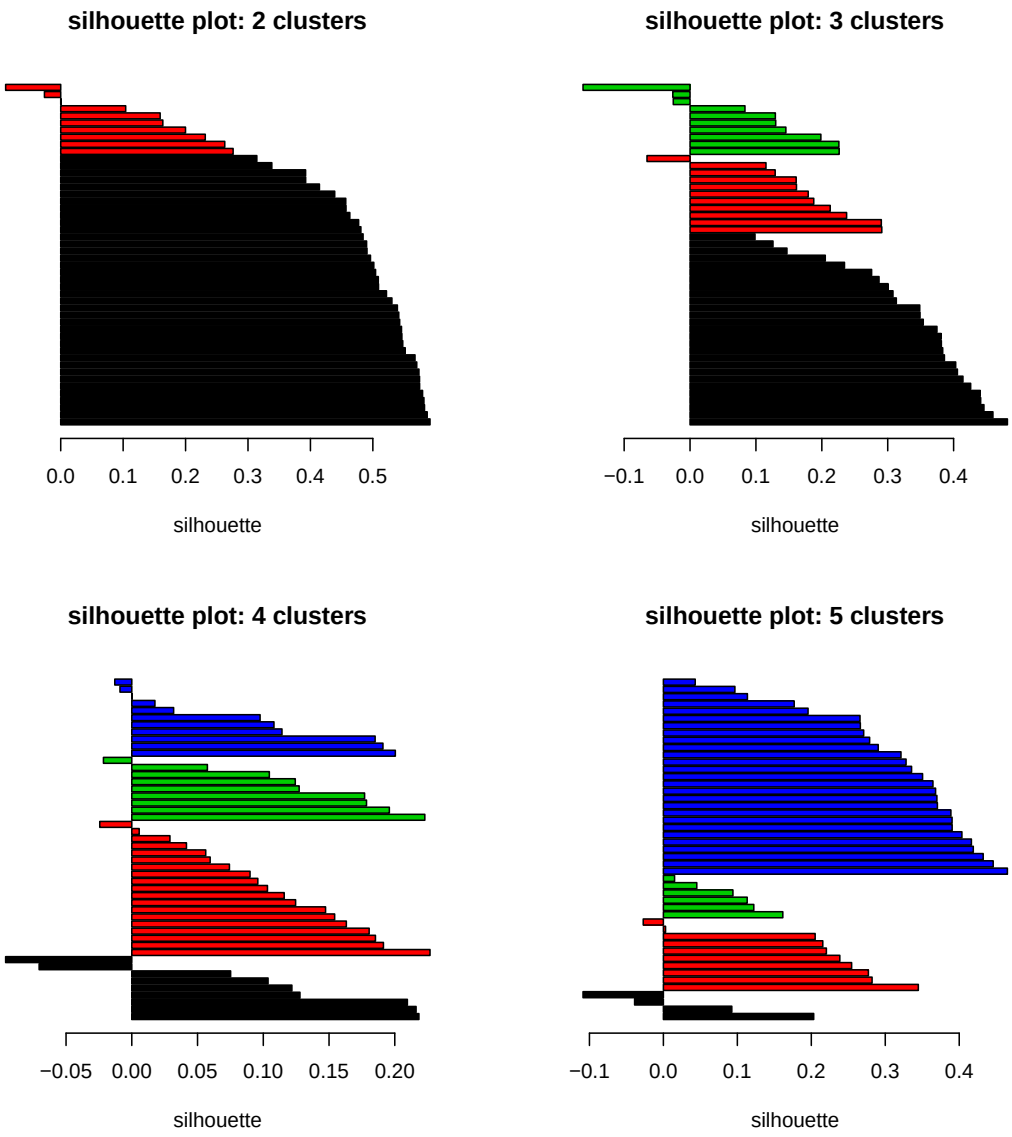


Figure 12: Silhouette plots of the clustering result obtained via multivariate functional k-means procedure, setting $k = 2,3,4,5$; data are ordered according to an increasing value of silhouette within each cluster, and are coloured according to the cluster assignment.

means of a "typical shape" of ECGs assigned to each cluster. This cluster representative can be obtained by selecting, among all curves assigned to a given cluster, the one that according to the measure $d(\cdot,\cdot)$ defined in (5) is more similar to the final estimated centroid of the cluster. The representative ECG traces for each cluster are shown in Fig. 16; we remark that the disease classification of these patients is coherent with the result shown in Table 3, since the one associated to the first cluster is Normal, the one associated to the second one suffers from RBBB, and the one assigned to the third suffers from LBBB. Moreover, supposing not to take into account the Mortara-Rangoni algorithm disease classification, also from clinical inspection of the three selected ECG

	Normal	RBBB	LBBB
Black	25	2	0
Red	0	9	2
Green	0	2	8

Table 3: Confusion matrix related to patients disease classification. Results are obtained by multivariate functional 3-means clustering algorithm to smoothed and registered QT-segment of ECG curves.

traces in Fig. 16 the diagnosis of the physician would have been physiological, affected by RBBB and by LBBB respectively for the first, second and third patient.

This final consideration might lead to the definition of a semi–automatic diagnostic procedure based on the previously described functional clustering technique of ECG traces: in fact, the final result of our clustering procedure is a set of k centroids representative of each cluster, which can be used as reference signals to compare a new ECG trace. Suppose a new ECG signal is available: we could have an immediate hint on the new patient's pathology by smoothing its ECG trace, registering it and finally assigning it to the group characterized by the nearest centroid, again according to the measure defined in (5).

Further refinements of our clustering procedure could help in its integration in the cardiovascular context, possibly for the diagnosis of different kinds of pathologies (not only BBB); due to the extreme generality of the algorithm, which is based only on morphological characteristics of the curves, this generalization can be based on a proper definition of a measure of the distance between functional data.

5 Conclusions and further developments

In this work we proposed a statistical framework for analysis and classification of ECG curves starting from their sole morphology.

We analyzed a pilot database composed by 48 ECG traces - 25 of them were *Normal*, 13 were *Right Bundle Branch Blocks* and 10 were *Left Bundle Branch Blocks* - extracted from PROMETEO datawarehouse.

The strongly localized features (peaks, oscillations…) of ECG curves makes them particulary suited to be smoothed via wavelets decomposition, since every basis function is localized both in time and in frequency; to this aim, we used a Daubechies wavelet basis with 10 vanishing moments. Moreover, being ECGs functional observations, they show both phase and amplitude variation, i.e. the same features can appear at different times among the patients. Since a correct separation between these two kind of variability is necessary for a successful analysis, we perform landmark–based registration of ECG traces, choosing as landmarks those time points that can be associated with a specific biological event: five of them are provided by the measurement procedure, identifying the P wave, the QRS complex and the T wave; we add one more landmark corresponding to the peak of R wave on the I lead, an easily localized feature on each ECG.

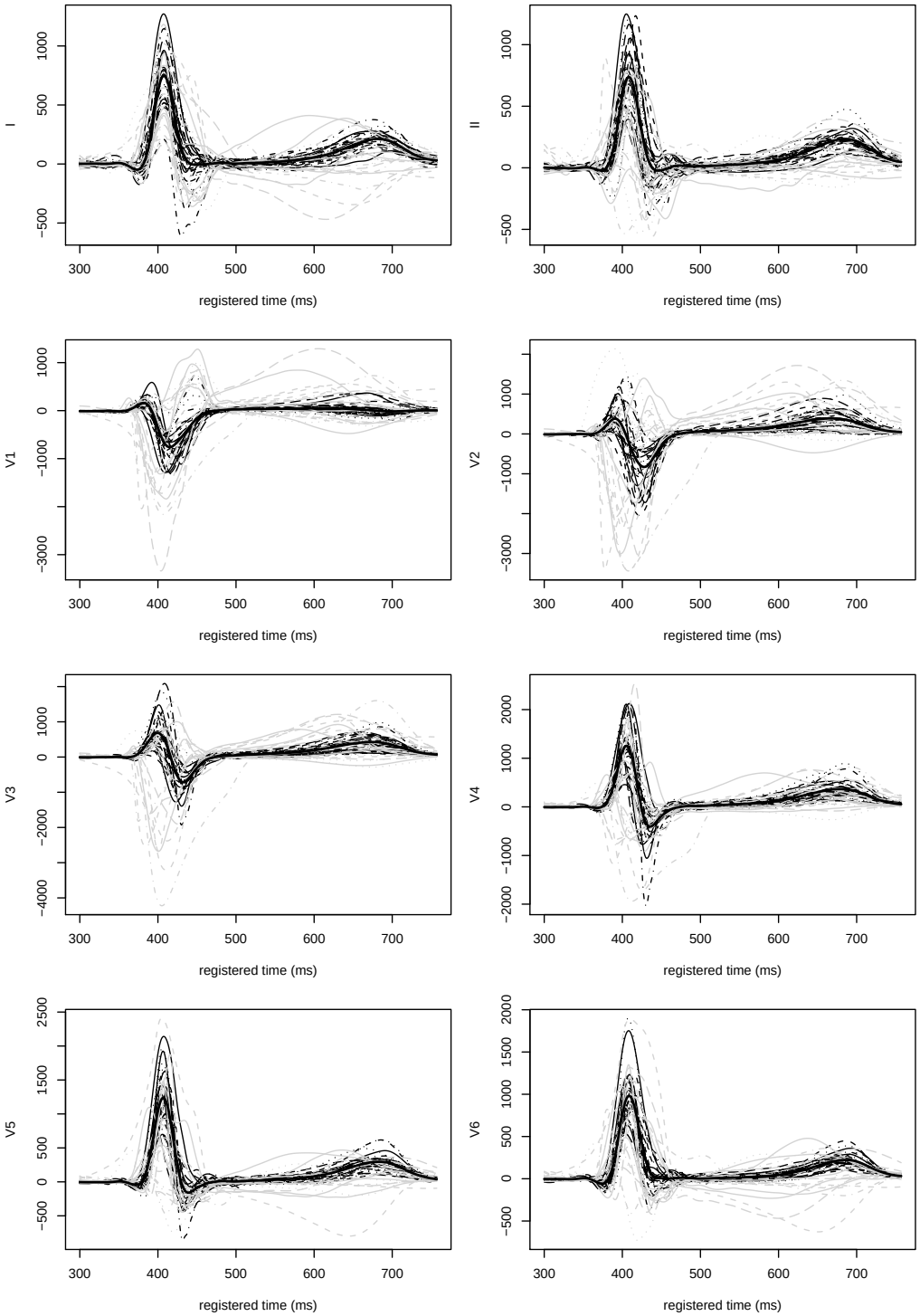


Figure 13: Smoothed and registered ECG traces (QT-segment): the whole dataset is plotted in lightgray, except for the curves assigned to the first cluster pointed out by multivariate functional 3-means procedure, which are shown in black.

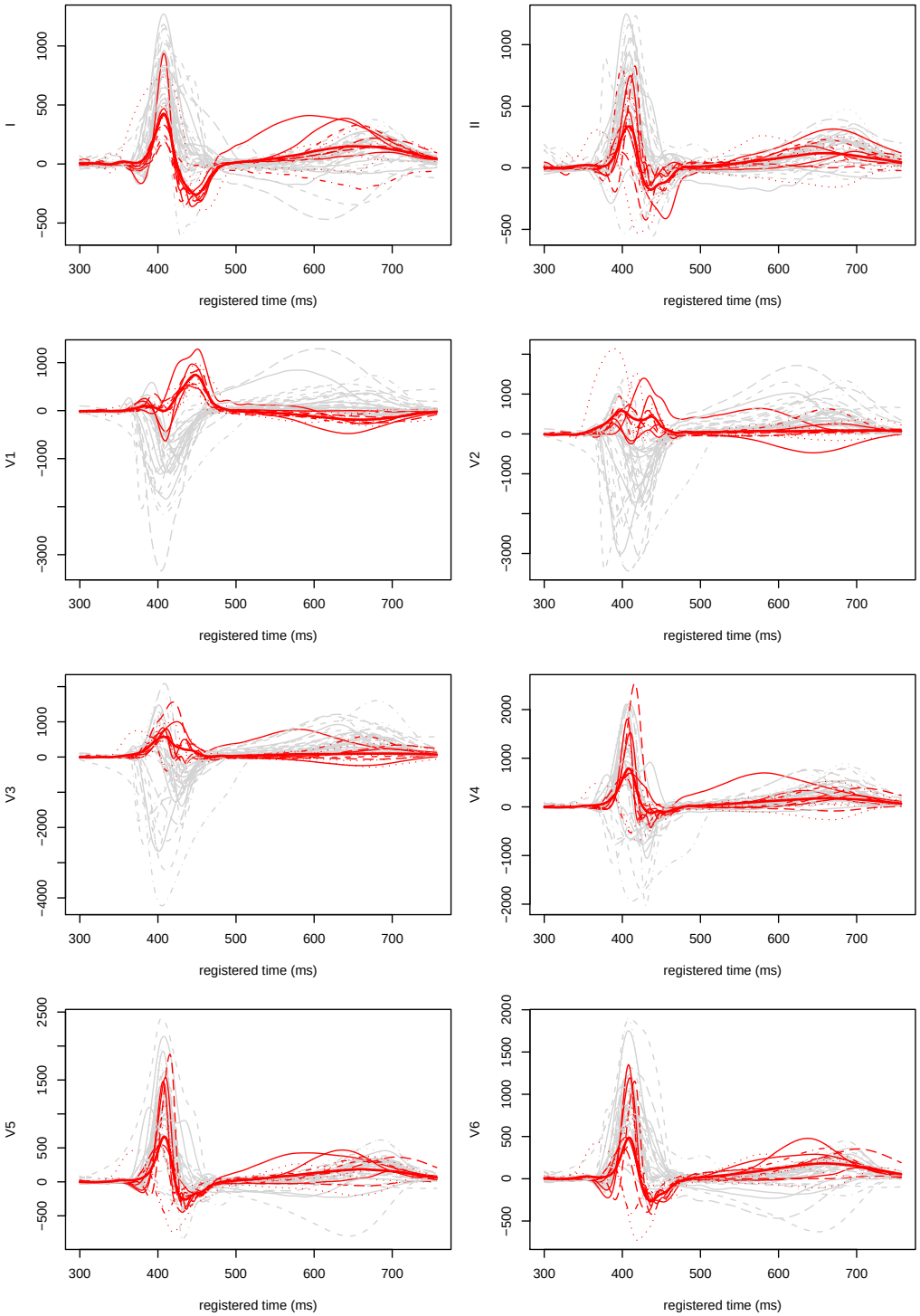


Figure 14: Smoothed and registered ECG traces (QT-segment): the whole dataset is plotted in lightgray, except for the curves assigned to the second cluster pointed out by multivariate functional 3-means procedure, which are shown in red.

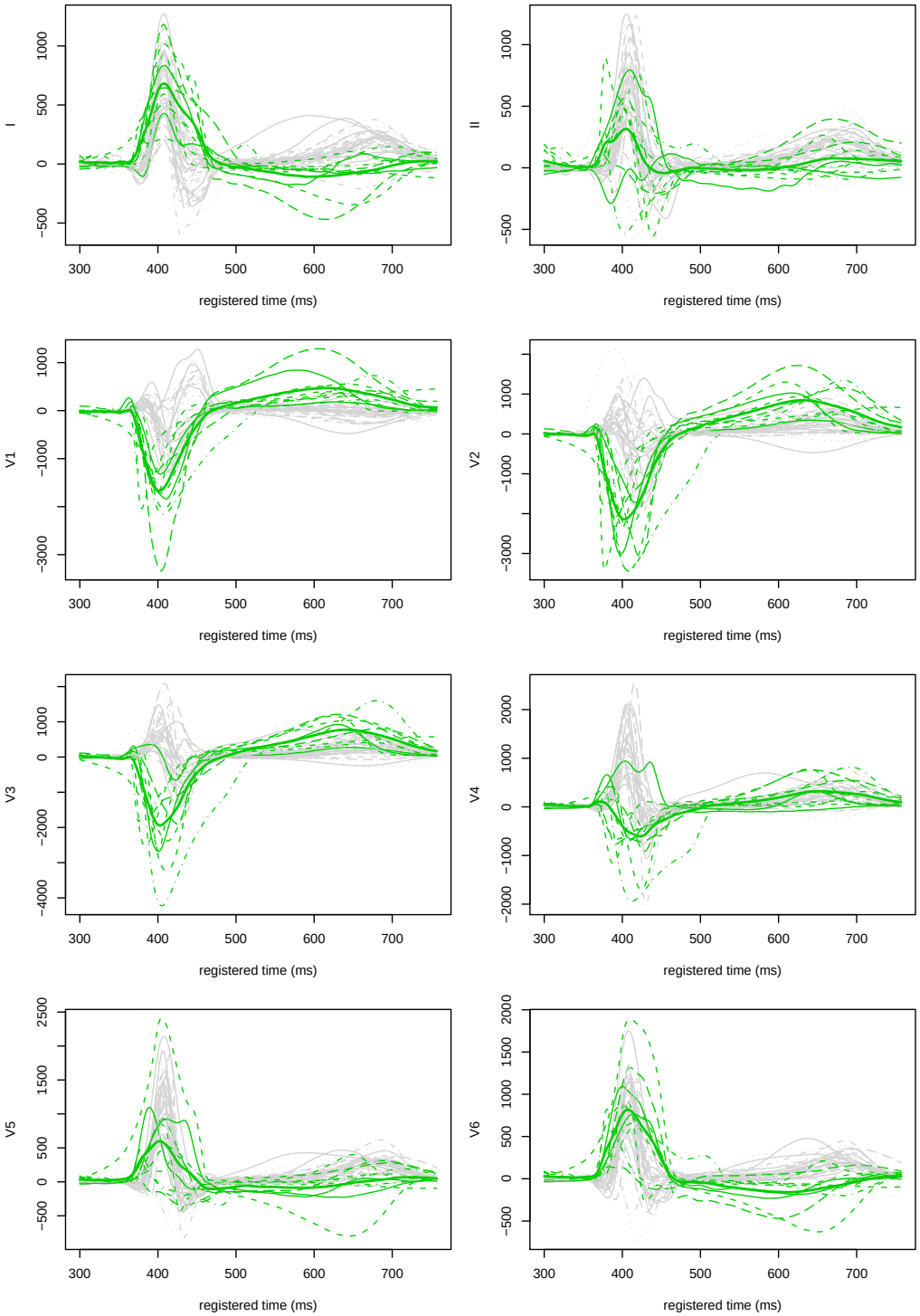


Figure 15: Smoothed and registered ECG traces (QT-segment): the whole dataset is plotted in lightgray, except for the curves assigned to the third cluster pointed out by multivariate functional 3-means procedure, which are shown in green.

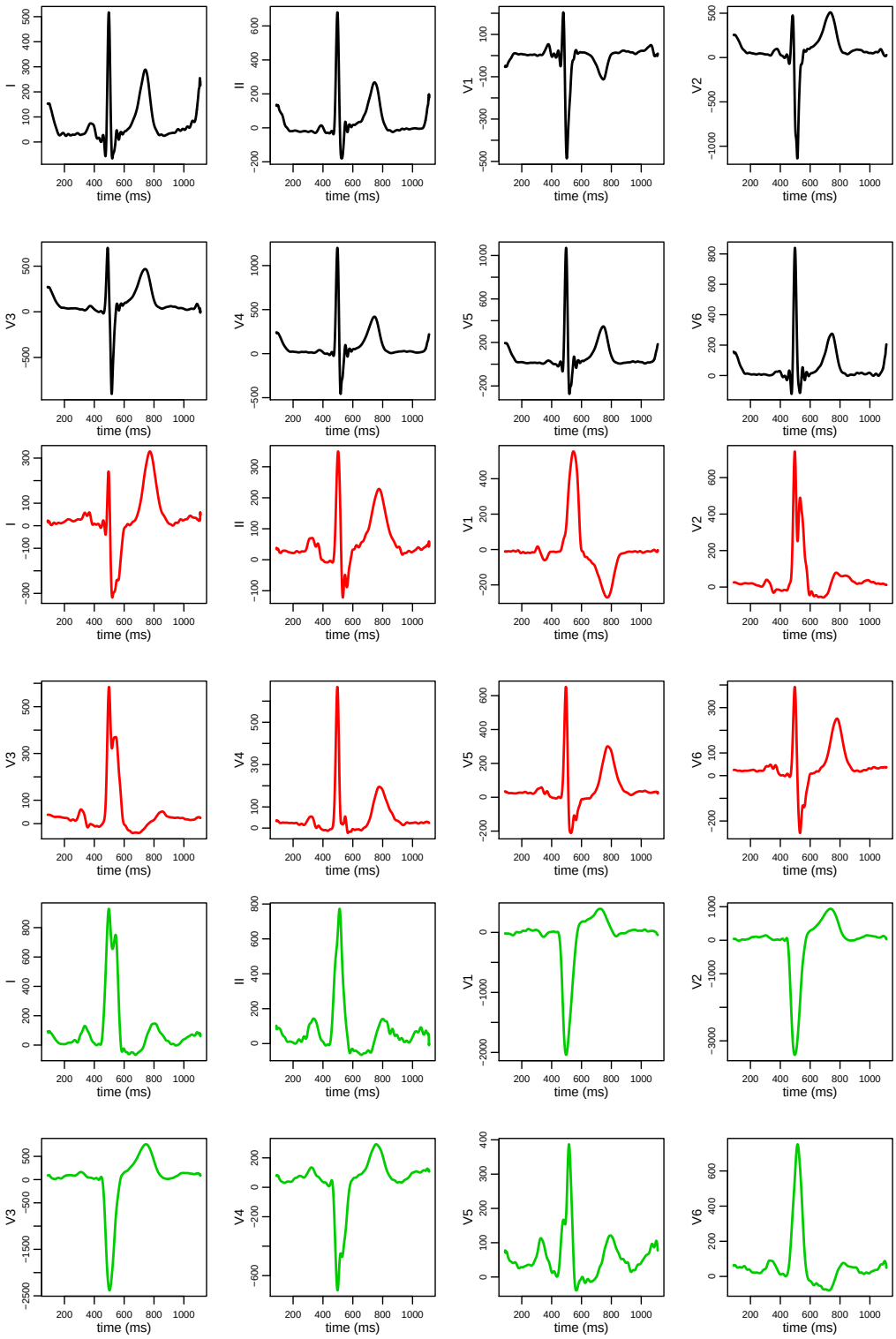


Figure 16: Smoothed original ECG signals of representative patients of three groups pointed out by multivariate functional 3-means procedure (Black curves refer to physiological trace, Green ones to LBBB trace, Red ones to RBBB trace).

Through smoothing and registration we managed to separate morphological information of the curves (i.e. amplitude variability), and duration of each ECG interval (i.e. phase variability); we can thus focus on each of them to perform clustering of smoothed and registered ECG traces. We chose to analyze morphological information via a multivariate functional k-means, thus simultaneously clustering all 8 leads of each patient, and to treat phase information via a multivariate k-means on the ECG intervals duration of each patient. The optimal number of clusters can be chosen via a silhouette plot of the final classification. The confusion matrix resulting from our classification framework shows appreciable results, with low misclassification rate. Moreover, this technique could also help in the semi–automatic diagnosis of BBB–related anomalies of ECGs, with an extreme generality of the classification procedure due to the flexibility in the definition of a proper measure of the distance between functional data.

The innovative aspect of this proposal consists in developing advanced statistical methods aimed at detecting pathological ECG traces (in particular Bundle Branch Blocks), starting only from morphological features of the curves. This allows for diagnoses consistent with clinical practice, starting from purely statistical considerations.

Aknowledgement

This work is within PROMETEO (PROgetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall'Extra Ospedaliero). The authors wish to thank in particular dr. Niccolò Grieco for clinical counselling and Ing. Johan DeBie for technical support. Finally, thanks to AREU, 118 Milan Dispatch Center and Mortara Rangoni Europe s.r.l. for having provided data.

References

[1] Antman, E.M., Hand, M., Amstrong, P.W., Bates, E.R., Green L.A. & al. (2008) Update of the ACC/AHA 2004 Guidelines for the Management of Patients with ST Elevation Myocardial Infarction, *Circulation*, 117, 269-329.

[2] Daubechies, I. (1988), Orthonormal basis of compactly supported wavelets, *Communictions on Pure and Applied Mathematics*, **41**, 909–996.

[3] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995), Wavelet Shrinkage: Asymptopia. *Journal of the Royal Statistical Society, Ser. B* , **57**, 301–369.

[4] Einthoven, W. (1908), Weiteres ber das Elektrokardiogram, *Pflger Arch. ges. Physiol.*, 122, 517–48.

[5] Einthoven, W., Fahr, G. and de Waart, A. (1950), On the direction and manifest size of the variations of potential in the human heart and on the influence of the position of the heart on the form of the electrocardiogram. *American Heart Journal*, 40(2), 163–211.

[6] Goldberger, E. (1942a), The aVL, aVR, and aVF leads: a simplification of standard lead electrocardiography. *American Heart Journal*, 24, 378–96.

[7] Goldberger, E. (1942b), A simple indifferent electrocardiographic electrode of zero potential and a technique of obtaining augmented, unipolar extremity leads. *American Heart Journal*, 23, 483–92.

[8] Grieco, N., Sesana, G., Corrada, E., Ieva, F., Paganoni, A.M., Marzegalli M. (2007). The Milano Network for Acute Coronary Syndromes and Emergency Services. *MESPE journal*, First Special Issue 2007.

[9] Grieco, N., Ieva, F., Paganoni, A.M. (2010). Provider Profiling Using Mixed Effects Models on a Case Study concerning STEMI Patients *Mox Report n. 21/2010*, Dipartimento di Matematica, Politecnico di Milano.
[Online] http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/021-2010.pdf

[10] Ieva, F., 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.

[11] Kohler, M. (2002), Universal Consistency of Local Polynomial Kernel Regression Estimates, *Ann. Inst. Statist. Math.*, 54(4), 879–899.

[12] Lindsay, A.E. (2006), ECG learning centre, [online] http://library.med.utah.edu/kw/ecg/index.html

[13] Mason, R., Likar, L. (1966), A new system of multiple leads exercise electrocardiography, *American Heart Journal*, 71(2), 196–205.

[14] Pigoli, D. and Sangalli, L.M. (2010), "Wavelets in Functional Data Analysis: estimation of multidimensional curves and their derivatives", Tech. Rep. MOX, Dipartimento di Matematica, Politecnico di Milano.

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

[16] Sangalli, L.M., Secchi, P., Vantini, S., and Vitelli, V. (2010), k-mean alignment for curve clustering, *Computational Statistics and Data Analysis*, 54, 1219–1233.

[17] Scher, A.M. and Young, A.C. (1957), Ventricular depolarization and the genesis of the QRS, *Annals of New York Academy of Science*, 65, 768–78.

[18] Struyf, A., Hubert, M., and Rousseeuw, P. (1997), Clustering in an Object-Oriented Environment, *Journal of Statistical Software*, 1, 4, 1–30.

[19] Wilson, F.N., Johnston, F.D., Rosenbaum, F.F., Erlanger, H., Kossmann, C.E., Hecht, H., Cotrim, N., Menezes de Olivieira, R., Scarsi, R., Barker, P.S. (1944), The precordial electrocardiogram, *American Heart Journal*, 27, 19–85.

[20] Tarpey, T., and Kinateder, K. K. J. (2003), Clustering Functional Data, *Journal of Classification*, 20, 93–114.

MOX Technical Reports, last issues

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

04/2011 Francesca Ieva, Anna Maria Paganoni, Davide Pigoli, Valeria Vitelli:
Multivariate functional clustering for the analysis of ECG curves morphology

03/2011 Giulia Garegnani, Giorgio Rosatti, Luca Bonaventura:
Mathematical and Numerical Modelling of Fully Coupled Mobile Bed Free Surface Flows

02/2011 Toni Lassila, Alfio Quarteroni, Gianluigi Rozza:
A reduced basis model with parametric coupling for fluid-structure interaction problems

01/2011 M. Dalla Rosa, Laura M. Sangalli, Simone Vantini:
Dimensional Reduction of Functional Data by means of Principal Differential Analysis

43/2010 Giancarlo Pennati, Gabriele Dubini, Francesco Migliavacca, Chiara Corsini, Luca Formaggia, Alfio Quarteroni, Alessandro Veneziani:
Multiscale Modelling with Application to Paediatric Cardiac Surgery

42/2010 Stefano Baraldo, Francesca Ieva, Anna Maria Paganoni, Valeria Vitelli:
Generalized functional linear models for recurrent events: an application to re-admission processes in heart failure patients

41/2010 Davide Ambrosi, Gianni Arioli, Fabio Nobile, Alfio Quarteroni:
Electromechanical coupling in cardiac dynamics: the active strain approach

40/2010 Carlo D'Angelo, Anna Scotti:
A Mixed Finite Element Method for Darcy Flow in Fractured Porous Media with non-matching Grids

39/2010 Carlo D'Angelo:
Finite Element Approximation of Elliptic Problems with Dirac Measure Terms in Weighted Spaces. Applications to 1D-3D Coupled Problems

38/2010 Nancy Flournoy, Caterina May, Piercesare Secchi:
Response-adaptive designs in clinical trials for targeting the best treatment: an overview