Appl. Statist. (2013)
62, *Part* 3, *pp.* 401–418

Multivariate functional clustering for the morphological analysis of electrocardiograph curves

Francesca Ieva, Anna M. Paganoni and Davide Pigoli

Politecnico di Milano, Italy

and Valeria Vitelli

Ecole Centrale, Paris, and Supélec, Gif-sur-Yvette, France

[Received February 2011. Final revision July 2012]

Summary. Cardiovascular ischaemic diseases are one of the main causes of death all over the world. In this class of pathologies, a quick diagnosis is essential for a good prognosis in reperfusive treatment. In particular, an automatic classification procedure based on statistical analysis of teletransmitted electrocardiograph ('ECG') traces would be very helpful for an early diagnosis. This work presents an analysis of ECG traces, either physiological or pathological, of patients whose 12-lead prehospital ECG has been sent to the 118 Dispatch Center in Milan by life-support personnel. The statistical analysis starts with a preprocessing step, where functional data are reconstructed from noisy observations and biological variability is removed by a non-linear registration procedure. Then, a multivariate functional *k*-means clustering procedure is carried out on reconstructed and registered ECGs and their first derivatives. Hence, a new semi-automatic diagnostic procedure, based solely on the ECG morphology, is proposed to classify ECG traces; finally, the performance of this classification method is evaluated.

Keywords: Electrocardiograph signal; Functional *k*-means clustering; Functional registration; Wavelets smoothing

1. Introduction

Cardiovascular ischaemic diseases are nowadays one of the main causes of death all over the world. In Italy, they are responsible for 44% of overall deaths and for most of the emergency rescue operations. In fact, almost all calls to the 118 Dispatch Center in Milan (the Italian toll-free number for emergencies) that require rescue operations are due to cardiovascular failures. In the case of coronary artery ischaemic disease, a quick diagnosis is essential for a good prognosis in reperfusive treatment. This requires well-organized and synchronized prehospital, interhospital and in-hospital networks.

In 2001, a working group was formed comprising 23 cardiology units in the Milanese urban area and the 118 Dispatch Center. Since 2006, this group has been performing data collection twice a year (two monthly sessions per year) on all patients who are affected by coronary artery diseases, admitted to any hospital belonging to the Cardiological Network in Milan, within a project called 'Month monitoring myocardial infarction in Milan' (see Ieva and Paganoni (2010) and Grieco *et al.* (2007, 2011)). The analysis of these data identified the time of first

Address for correspondence: Davide Pigoli, Dipartimento di Matematica, Politecnico di Milano, via Bonardi 9, I-20133, Milano, Italy.
E-mail: davide.pigoli@mail.polimi.it

0035–9254/13/62401

electrocardiograph ('ECG') teletransmission as the most important factor to guarantee quick access to an effective treatment for patients (see also Antman *et al.* (2008)).

In 2008, a project called the Progetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall' Extra Ospedaliero (PROMETEO) started with the aim of spreading the intensive use of ECGs as a prehospital diagnostic tool and constructing a new database of ECGs with features which had not been recorded in any other data collection on heart diseases. All basic life-support units in the urban area of Milan carry ECG recorders with global system for mobile communications transmission.

In this work we analyse a sample ($n = 198$) of data from the PROMETEO data warehouse, which contains all the ECG traces recorded in the Milanese urban area by basic life-support units since the end of 2008. Our main goal is to identify, from a statistical perspective, specific ECG patterns which could benefit from an early invasive approach. The identification of statistical tools for classifying curves by using their shape helps in an early detection of heart failure, replacing a traditional clinical observation. For this, it is extremely important to understand the link between cardiac physiology and ECG trace shape. As detailed in the following sections, we focus on physiological traces in contrast with right bundle branch block (RBBB) and left bundle branch block (LBBB) traces. A bundle branch block (BBB) is a cardiac conduction abnormality that is seen in the ECG. In this condition, the activation of the right (or left) ventricle is delayed, which results in one ventricle contracting later than the other.

Details on BBBs and their connection with non-physiological shapes of ECG signals will be treated in Section 2, where also clinical details about ECG signals will be given. Wavelet smoothing of ECG traces and their first derivatives and a procedure of landmark registration are explained in Section 3. In Section 4 data analysis is presented, consisting of a multivariate functional k-means clustering of QT-segments of smoothed and registered ECG curves and first derivatives. In Section 5, the results of the analysis are discussed, and further developments to be explored in future work are proposed. All the analyses were carried out by using R statistical software (see R Development Core Team (2009)).

2. Electrocardiography and bundle branch blocks

Electrocardiography is a transthoracic recording of the electrical activity of the heart over a period of time as it is captured and externally recorded through skin electrodes. The ECG works mostly by detecting and amplifying the tiny electrical changes in the skin caused by the depolarization of the heart muscle during each heartbeat (for further clinical details, see Lindsay (2006)). Nowadays, the most commonly used clinical ECG system, the 12-lead ECG system (see Einthoven (1908) and Einthoven *et al.* (1950)), consists of the following 12 leads: I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5 and V6. The main reason for recording all the 12 leads is that it enhances pattern recognition (see Goldberger (1942a, b), Mason and Likar (1966) and Wilson *et al.* (1944)). 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 that are analysed in the following sections will consist of leads I, II, V1, V2, V3, V4, V5 and V6 only.

Fig. 1(a) shows the stylized shape of a physiological single beat together with lead I of a true physiological ECG (Fig. 1(b)). The main relevant points, segments and waves are highlighted. Deflections in the stylized signal are listed alphabetically 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 is masked by ventricular

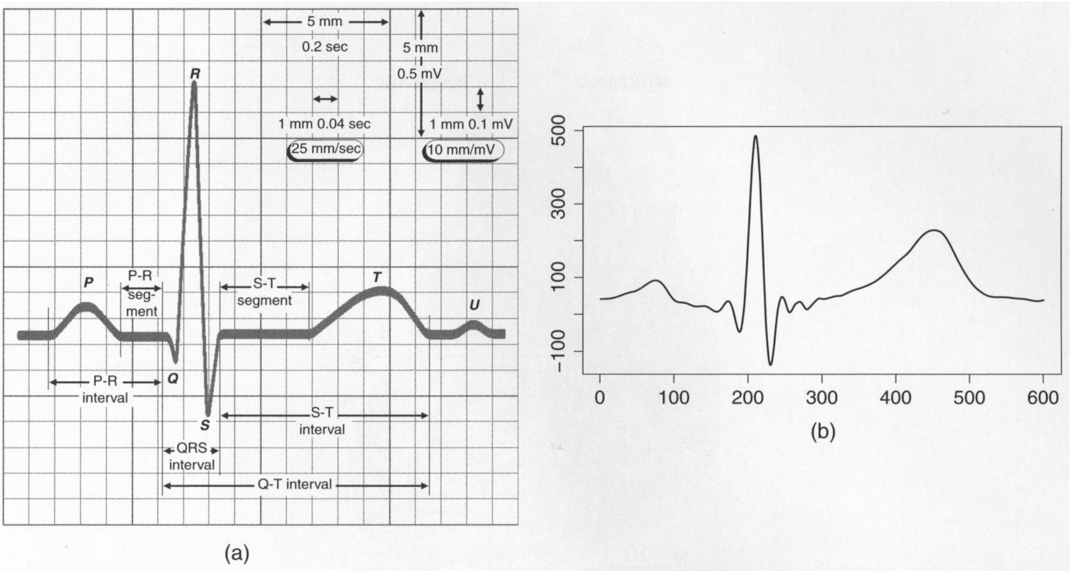

Fig. 1. (a) Stylized shape of a physiological single beat, recorded on ECG graph paper (the main relevant points, segments and waves are highlighted) and (b) lead I of a true physiological ECG

repolarization activity (see Scher and Young (1957)). The direction of the wave of depolarization is named the *heart electrical axis*.

As mentioned before, in this work we analyse 198 ECG traces from the PROMETEO data warehouse. Each file in PROMETEO can be associated with three subfiles. The first is called 'Details' and contains technical information, which is useful for signal processing and analysis, such as times of waves' repolarization and depolarization, landmarks indicating onset and offset times of main ECG subintervals and automatic diagnoses, established by the commercial Mortara–Rangoni VERITAS^{TM} algorithm (`http://www.mortara.com/products/healthcare/veritas-algorithm`). We used these automatic diagnoses to label the ECG traces that we analysed, to validate the performance of our unsupervised clustering algorithm. The challenge of this work consists of tuning and testing a realtime procedure which enables semi-automatic diagnosis of the patients' disease based only on the morphology of ECG traces, making clinical evaluations not strictly necessary. The second subfile is called 'Rhythm' and contains the output of an ECG recorder. Specifically, it registers 10 s (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 s (1200 points). We carried out the analysis using only the median files, obtaining eight curves (one for each ECG lead) for each patient, representing a patient's 'Median' beat for that lead. This representative heartbeat is a trace of a single cardiac cycle (heartbeat), i.e. of a P-wave a QRS-complex, a T-wave and a U-wave, which are normally visible in 50–75% of ECGs.

The heart's electrical activity begins in the sinoatrial node (the heart's natural pacemaker; 1 in Fig. 2), which is set on the upper right atrium. The impulse travels next through the left and right atria and summates at the atrioventricular node (2 in Fig. 2). From the atrioventricular node the electrical impulse travels down the bundle of His (3 in Fig. 2) and divides into the right and left bundle branches (4 and 10 in Fig. 2). The right bundle branch contains one fascicle. The left bundle branch subdivides into two fascicles: the left anterior fascicle and the left posterior fascicle (6 and 5 in Fig. 2). Again, the fascicles divide into millions of Purkinje fibres which in

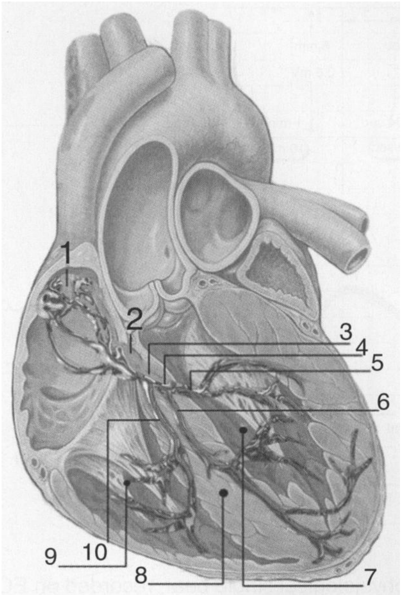

Fig. 2. Conduction system of the heart: 1, sinoatrial node; 2, atrioventricular node; 3, bundle of His; 4, left bundle branch; 5, left posterior fascicle; 6, left anterior fascicle; 7, left ventricle; 8, ventricular septum; 9, right ventricle; 10, right bundle branch

turn interdigitize with individual cardiac myocytes, allowing for rapid, co-ordinated, and synchronous physiologic depolarization of the ventricles. Bundle branch or fascicle injuries result in altered pathways for ventricular depolarization. In this case, 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, an RBBB typically causes prolongation of the last part of the QRS-complex and may shift the heart electrical axis slightly to the right. An LBBB widens the entire QRS-complex and in most cases shifts the heart electrical axis to the left. Another usual finding with BBBs is an appropriate T-wave discordance: this means that the T-wave will be deflected in the opposite direction to the terminal deflection of the QRS-complex.

From a statistical point of view, we shall focus our analysis on shape modifications that are induced on the ECG curves and their first derivatives by the BBB pathology, and we shall investigate these shape modifications only from 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

The data set from the PROMETEO data warehouse consists of the ECG signals of $n = 198$ subjects, among which 101 are normal and 97 are affected by BBBs (49 RBBBs and 48 LBBBs). As mentioned above, the aim of this work is to explore the morphology of the ECG curves. Thus, the basic statistical unit is the multivariate function which describes, for each patient, the heart dynamics on the eight leads.

However, in practice we have only a noisy and discrete observation of the function describing the ECG trace of each patient. Moreover, each patient has his own 'biological' time, i.e. the same event of the heart dynamics may happen at different times for different patients: that is why the morphological change due to this difference in timings is misleading from a statistical perspective. These two problems are common in functional data analysis applications and they can be addressed respectively by data smoothing and registration (see Ramsay and Silverman (2005)).

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,...). Since the eight leads of interest (i.e. I, II, V1, V2, V3, V4, V5 and V6) jointly describe the complex heart dynamics, the smoothing technique should take into account all eight leads simultaneously. This helps in detecting relevant features affecting more than one lead. We use the wavelet-based smoothing technique for multivariate curves proposed in Pigoli and Sangalli (2012). This technique is used to obtain the smoothed estimates of eight-dimensional ECG signals. 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. Thus, starting from the vectorial raw signal, we can estimate the vectorial function

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

and its derivatives, for each patient $i = 1, \ldots, n$. Fig. 3 shows raw data and functional estimates obtained with this wavelet smoothing procedure for a normal subject. The smoothing procedure is essential also for an accurate derivative reconstruction, as shown in Fig. 4, where the estimate of the first derivative is superimposed on the first central finite difference (i.e. a rough indication of first-derivative behaviour). In particular, a Daubechies wavelet basis with 10 vanishing moments (see Daubechies (1988)) is used, as this basis is sufficiently smooth to allow the computation of the first derivative of the estimated functional data (see Pigoli and Sangalli (2012) for details).

As in most smoothing methods based on wavelet expansion, it is necessary to deal with a grid of 2^J points, $J \in \mathbb{N}$. Thus, in the further analysis we use only the central $2^{10} = 1024$ observation points. There is no loss of significant information: the portion of the signal that we focus on contains all the important features of the ECG trace.

This smoothing method relies on the decimated discrete wavelet transform, which is not translation invariant. This is suitable for our procedure, since the problem of the misalignment of different wave shapes is dealt with by landmark registration.

3.1. Landmark registration

Functional observations usually show variability both in phase and amplitude, i.e., apart from morphological variability, each curve has its own biological time so the same feature can appear at different times among the patients. It is well known that a correct separation between these two kinds of variability is necessary for a successful analysis (see Ramsay and Silverman (2005)). We face this problem with 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 the Mortara–Rangoni procedure and can be found in the details file. They identify, for each patient $i = 1, \ldots, n$, the P-wave $(P^i_{\text{onset}}, P^i_{\text{offset}})$, the QRS-complex $(QRS^i_{\text{onset}}, QRS^i_{\text{offset}})$ and the T-wave (T^i_{offset}). We add one more landmark: the R-peak identified on lead I (I^i_{peak}). We choose this landmark because on lead I only both the physiological 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 time warping function h_i such that

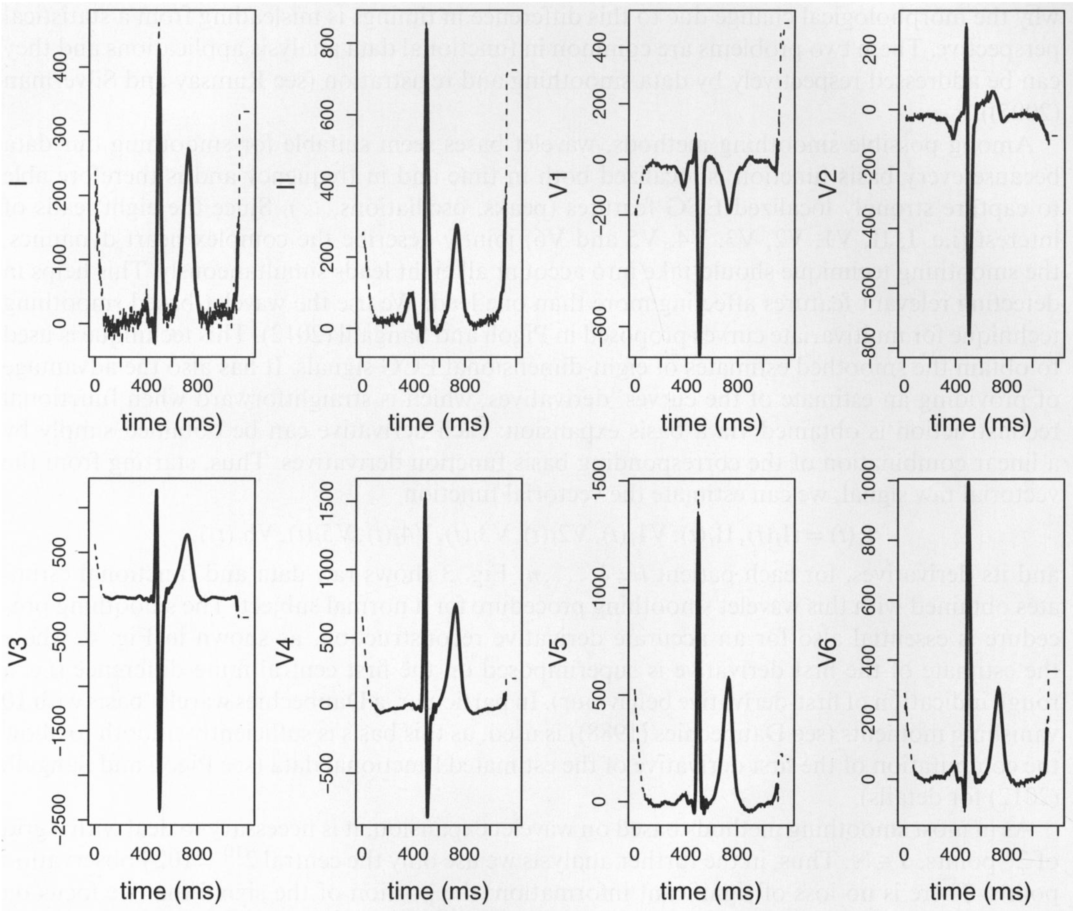

Fig. 3. Raw data of the eight leads (-------) and wavelet functional estimates (———) for a normal subject

$$h_i(P^i_{\text{onset}}) = \bar{P}_{\text{onset}}, \qquad h_i(P^i_{\text{offset}}) = \bar{P}_{\text{offset}},$$
$$h_i(QRS^i_{\text{onset}}) = \overline{QRS}_{\text{onset}}, \qquad h_i(I^i_{\text{peak}}) = \bar{I}_{\text{peak}},$$
$$h_i(QRS^i_{\text{offset}}) = \overline{QRS}_{\text{offset}}, \qquad h_i(T^i_{\text{offset}}) = \bar{T}_{\text{offset}}$$

where $\bar{P}_{\text{onset}}$, $\bar{P}_{\text{offset}}$, $\overline{QRS}_{\text{onset}}$, $\bar{I}_{\text{peak}}$, $\overline{QRS}_{\text{offset}}$ and $\bar{T}_{\text{offset}}$ are the mean values of the corresponding landmarks over all the patients. These values are reported in Table 1, together with the associated standard deviations. The ith warping function h_i is actually obtained by a cubic spline interpolation of the pairs

$$(P^i_{\text{onset}}, \bar{P}_{\text{onset}}), (P^i_{\text{offset}}, \bar{P}_{\text{offset}}), (QRS^i_{\text{onset}}, \overline{QRS}_{\text{onset}}), (I^i_{\text{peak}}, \bar{I}_{\text{peak}}),$$
$$(QRS^i_{\text{offset}}, \overline{QRS}_{,\text{offset}}), (T^i_{\text{offset}}, \bar{T}_{\text{offset}}).$$

To solve this problem, we resort to the monotone interpolation method that was proposed in Fritsch and Carlson (1980). Thus, the registered vectorial function will be

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

for every patient $i = 1, \ldots, n$. Fig. 5 shows both unregistered and registered I-leads for all the 198 patients. As mentioned above, biological time is the same for all the leads of the same patient.

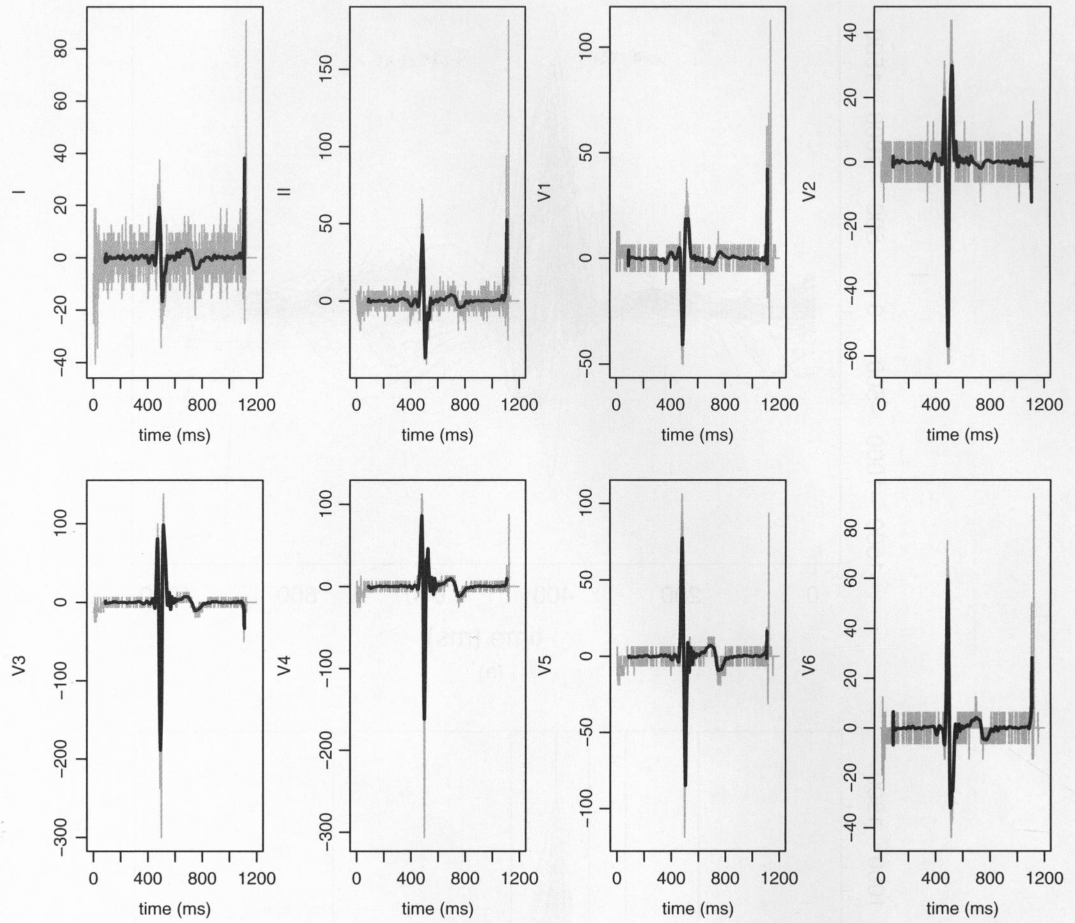

Fig. 4. First central finite difference of the eight leads (——) and wavelet estimates of the first derivatives (——) for a normal subject

Table 1. Mean of landmarks of all the curves, used to select the portion of smoothed and registered ECG curves and landmark standard deviations†

	$\bar{P}_{\text{onset}}\,(ms)$	$\bar{P}_{\text{offset}}\,(ms)$	$\overline{QRS}_{\text{onset}}\,(ms)$	$\bar{I}_{\text{peak}}\,(ms)$	$\overline{QRS}_{\text{offset}}\,(ms)$	$\bar{T}_{\text{offset}}\,(ms)$
Mean	184.3	298.2	354.8	407.2	476.9	755.8
Standard deviation	39.7	37.4	18.9	15.4	21.4	44.2

†Landmarks values are referred to a registered time.

This is a non-linear registration procedure, since in this framework there is no simple affine transformation which can take into account the subject-specific variability.

The registration procedure separates morphological information (i.e. amplitude variability) from duration of the different segments of the ECG (i.e. phase variability). The former is captured by the registered ECG traces, whereas the latter is described by warping functions, determined by landmarks. In clinical practice the duration of different segments of ECGs

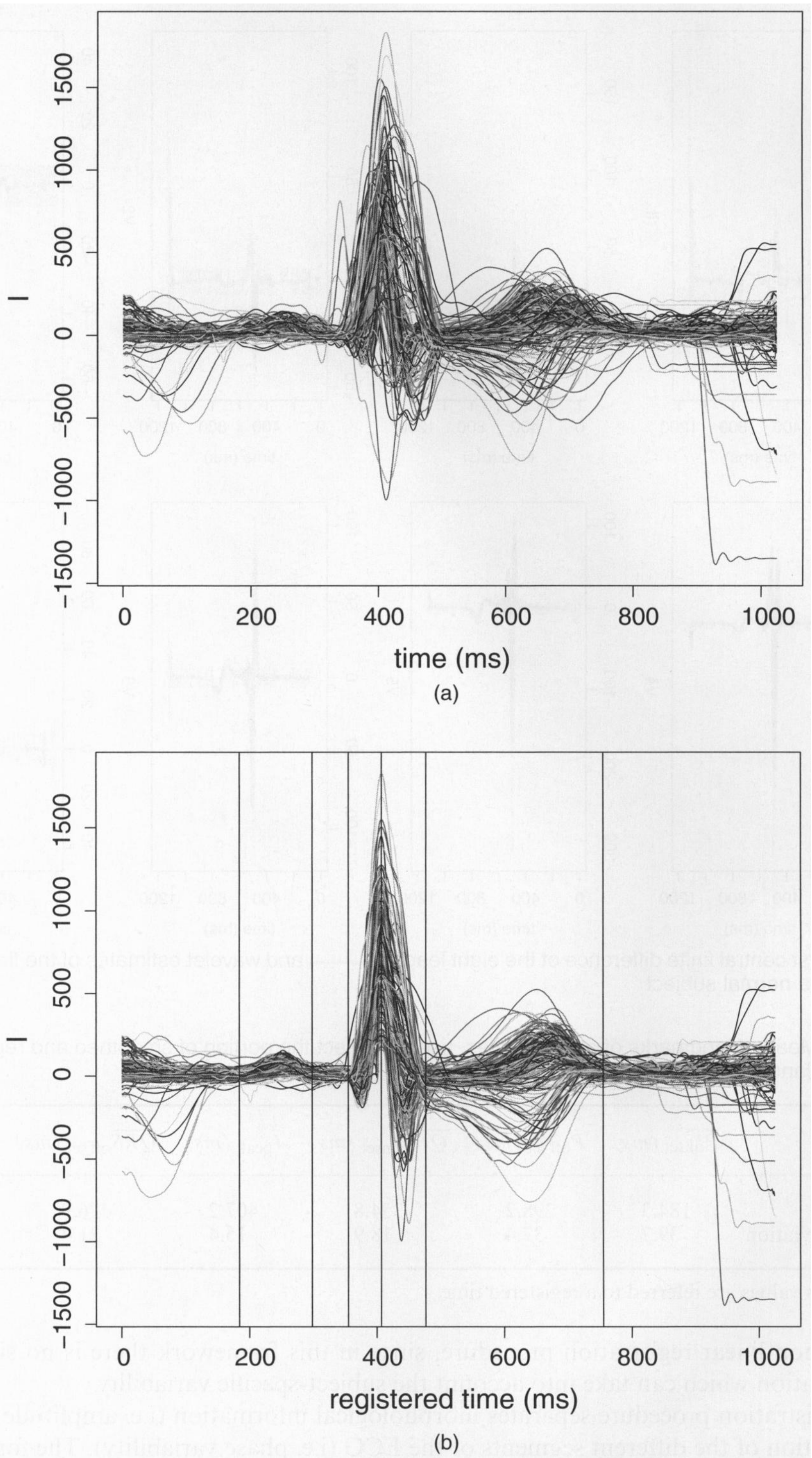

Fig. 5. Original I-leads for (a) the 198 patients and (b) the registered patients: |, position of mean landmarks $\bar{P}_{\text{onset}}$, $\bar{P}_{\text{offset}}$, $\overline{QRS}_{\text{onset}}$, $\bar{I}_{\text{peak}}$, $\overline{QRS}_{\text{offset}}$, $\bar{T}_{\text{offset}}$

Table 2. Confusion matrix related to patients' disease classification†

Cluster	*Normal*	*RBBB*	*LBBB*
1	96	6	0
2	2	17	25
3	3	26	23

†Results are obtained performing the three-means clustering algorithm on interval lengths.

and particularly the QRS-complex length is one of the most important parameters to identify pathological situations. However, this kind of information does not distinguish between different pathologies, such as RBBBs and LBBBs. This can also be seen in our data set. If we perform a multivariate three-means algorithm on the sample of interval lengths $\{(P^i_{\mathrm{offset}} - P^i_{\mathrm{onset}}), (QRS^i_{\mathrm{onset}} - P^i_{\mathrm{offset}}), (QRS^i_{\mathrm{offset}} - QRS^i_{\mathrm{onset}})$ and $(T^i_{\mathrm{offset}} - QRS^i_{\mathrm{offset}})\}$, for $i = 1, \ldots, n$, with the aim of identifying the existing three groups, we obtain the result that is shown in Table 2: 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, to extract other diagnostic information from ECG morphology. In clinical practice, the result of our analysis might be considered together with traditional diagnostic tools based on segment lengths.

4. Data analysis

In this section we propose the use of functional data analysis techniques to perform a clustering of smoothed and registered ECG traces. The aim of the analysis is to develop a proper classification procedure, to distinguish the grouping structure that is induced in the sample of ECGs by the presence of different pathologies, solely on the basis of the shape of the curves considered.

As previously discussed in Section 2, ECG traces are very complex functional data, where different portions of the domain can be analysed to detect different pathologies. The main focus of our analysis is the investigation of BBB pathology, which mainly expresses itself in the ECG trace through a lengthening of the QRS-complex and a modification of the T-wave. In fact, the diagnosis of BBBs has nothing to do with modifications in the P-wave, since this portion of the ECG curve deals with cardiac rhythm dysfunctions that do not affect our patients. We thus focus our classification analysis on the QT-segment. Since we have already registered the ECG signals, all the curves show relevant features at the same time points, corresponding to the reference landmarks $\bar{P}_{\mathrm{onset}}$, $\bar{P}_{\mathrm{offset}}$, $\overline{QRS}_{\mathrm{onset}}$, $\bar{I}_{\mathrm{peak}}$, $\overline{QRS}_{\mathrm{offset}}$ and $\bar{T}_{\mathrm{offset}}$ (see Section 3.1): this allows us to select, for all the registered curves of the data set, only the portion of ECG trace belonging to the interval $[\bar{P}_{\mathrm{offset}}, \bar{T}_{\mathrm{offset}}]$, which is relevant to our diagnostic purposes.

In particular, we select the portion of

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

such that $t \in T := [\bar{P}_{\mathrm{offset}}, \bar{T}_{\mathrm{offset}}]$, where $\bar{P}_{\mathrm{offset}}$ and $\bar{T}_{\mathrm{offset}}$ are the values reported in the first row of Table 1, in the third and last columns respectively.

4.1. Functional classification

We analyse the n patients according to a functional k-means clustering procedure, where all the eight leads $\mathbf{F}_i(t): T \rightarrow \mathbb{R}^8$, for patients $i=1,\ldots,n$, are simultaneously clustered. To develop this clustering procedure we assume that $\mathbf{F}_i(t) \in H^1(T;\mathbb{R}^8)$. Since we consider all the eight leads simultaneously in the analysis, we name the clustering procedure employed *multivariate functional k-means*, to distinguish it from *standard functional k-means*, which would treat each lead separately.

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

$$d_1\{\mathbf{F}_i(t),\mathbf{F}_j(t)\} = \sqrt{\left[\sum_{r=1}^{8}\int_T \{F_i^r(t)-F_j^r(t)\}^2\,\mathrm{d}t + \int_T \{\mathrm{D}F_i^r(t)-\mathrm{D}F_j^r(t)\}^2\,\mathrm{d}t\right]}, \qquad (1)$$

for $i,j=1,\ldots,n$, where $\mathrm{D}F_i^r(t)$ is the wavelet estimate of the first derivative of the rth lead in the ECG trace of the ith patient. Note that the distance that is defined in equation (1) is the natural distance in the Hilbert space $H^1(T;\mathbb{R}^8)$.

To perform comparisons and to test the robustness of our clustering procedure we have considered two more distances between a pair of ECG traces:

$$\tilde{d}_1\{\mathbf{F}_i(t),\mathbf{F}_j(t)\} = \sqrt{\sum_{r=1}^{8}\int_T \{\mathrm{D}F_i^r(t)-\mathrm{D}F_j^r(t)\}^2\,\mathrm{d}t}, \qquad (2)$$

$$d_2\{\mathbf{F}_i(t),\mathbf{F}_j(t)\} = \sqrt{\sum_{r=1}^{8}\int_T \{F_i^r(t)-F_j^r(t)\}^2\,\mathrm{d}t}. \qquad (3)$$

The distance that is defined by equation (2) is the natural seminorm in the Hilbert space $H^1(T;\mathbb{R}^8)$, whereas the distance that is defined in equation (3) is the norm in the Hilbert space $L^2(T;\mathbb{R}^8)$. They are both considered in the clustering procedure not only to compare the performance of multivariate functional k-means under different specifications of the distance, but also to have an insight on the role of the first derivative of the curves: we claim that both the ECG trace and its first derivative are essential to distinguish between different morphologies.

The functional k-means clustering algorithm is an iterative procedure, alternating a step of *cluster assignment*, where all curves are assigned to a cluster, and a step of *centroid calculation*, where a relevant functional representative (the centroid) for each cluster is identified. More precisely, the algorithm is initialized by fixing the number k of clusters, and by randomly selecting between the curves in the data set a set of k initial centroids $\{\varphi_1^{(0)}(t),\ldots,\varphi_k^{(0)}(t)\}$. Initializing the means by randomly selecting k different data is a standard procedure also in the multivariate k-means algorithm; that is why we follow this strategy to select the initial mean curves. Given this initial choice, the algorithm iteratively repeats two basic steps. At the mth iteration of the algorithm, $m>0$, the two steps are performed as follows.

Step 1 (cluster assignment step): each curve is assigned to the cluster whose centroid at the $(m-1)$th iteration is the nearest according to the distance d, which is chosen between the distances that are defined in equations (1), (2) or (3). This means that the choice of the mth cluster assignment of the ith patient $C_i^{(m)}$, for $i=1,\ldots,n$, is

$$C_i^{(m)} = \underset{l=1,\ldots,k}{\arg\min}\, d\{\mathbf{F}_i(t),\varphi_l^{(m-1)}(t)\}.$$

Step 2 (centroid calculation step): the identification of centroids $\{\varphi_1^{(m)}(t),\ldots,\varphi_k^{(m)}(t)\}$ at the mth iteration is performed by solving the optimization problem

$$\varphi_l^{(m)}(t) = \arg \min_{\varphi \in \Omega_d} \sum_{i:C_i^{(m)}=l} d\{\mathbf{F}_i(t), \varphi(t)\}^2,$$

where $C_i^{(m)}$ is the cluster assignment of the ith patient at the current iteration, d is one of the three distances defined in equations (1), (2) or (3) and Ω_d is the Hilbert space where the chosen distance d is natural.

The solution to the infinite dimensional optimization problem that is expressed in the centroid calculation step obviously depends on the choice of the distance: it is possible to prove that, both when the distance is measured with d_1 or d_2, the minimizer $\varphi_l^{(m)}(t)$ corresponds to the functional mean of curves belonging to the same cluster. An immediate consequence of this result is that, when the seminorm in $H^1(\tilde{d}_1)$ is used, the centroid is the functional mean of the first derivatives of curves belonging to the same cluster.

The algorithm is stopped when the same cluster assignments are obtained at two subsequent iterations, i.e. the set of cluster assignments $\{C_1^{(\bar{m})}, \ldots, C_n^{(\bar{m})}\}$ and the set of centroids $\{\varphi_1^{(\bar{m})}(t), \ldots, \varphi_k^{(\bar{m})}(t)\}$ are considered final solutions of the algorithm if we obtain $C_i^{(\bar{m}+1)} \equiv C_i^{(\bar{m})}$ for all $i = 1, \ldots, n$. In the case that we consider, with this criterion, convergence is ensured in 8–10 iterations.

There are many different implementations of the functional k-means algorithm in the literature on functional data analysis, among which some procedures integrate registration in the classification steps (e.g. the k-means alignment algorithm that is described in Sangalli *et al.* (2010), the core shape modelling approach in Boudaoud *et al.* (2010), the non-parametric time-synchronized iterative mean updating technique in Liu and Müller (2003) or the simultaneously aligning and clustering K-centres model in Liu and Yang (2009)). Here, instead, we chose to separate registration and clustering in two consecutive steps of the analysis, since the latter does not use any information besides the morphology of the ECG traces, whereas the former is based on a strong clinical indication provided by landmarks supplied by the Mortara–Rangoni VERITAS$^{\text{TM}}$ algorithm.

Obviously, the k-means clustering procedure depends not only on the choice of the distance, but also on the number of clusters k. Since the number of clusters is unknown, we also consider a way to select the optimal number of clusters k^* via silhouette values and a plot of the final classification (see Struyf *et al.* (1997)). In particular, the silhouette plot of a final classification consists of 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 one of the three distances defined in equations (1)–(3), of the ith patient to all other patients assigned to the same cluster, whereas

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

is the minimum average distance of the ith patient from another cluster, where d is one of the three distances defined in equations (1)–(3). Clearly s_i always lies between -1 and 1, the former value indicating a misclassified patient, whereas the latter indicates a well-classified patient. Note that a patient who alone constitutes a cluster has silhouette value equal to 1 but is not considered in the silhouette plot for choosing k^*.

4.2. Results and discussion

The aim of the analysis is to detect the underlying grouping structure in our sample of 198 ECG

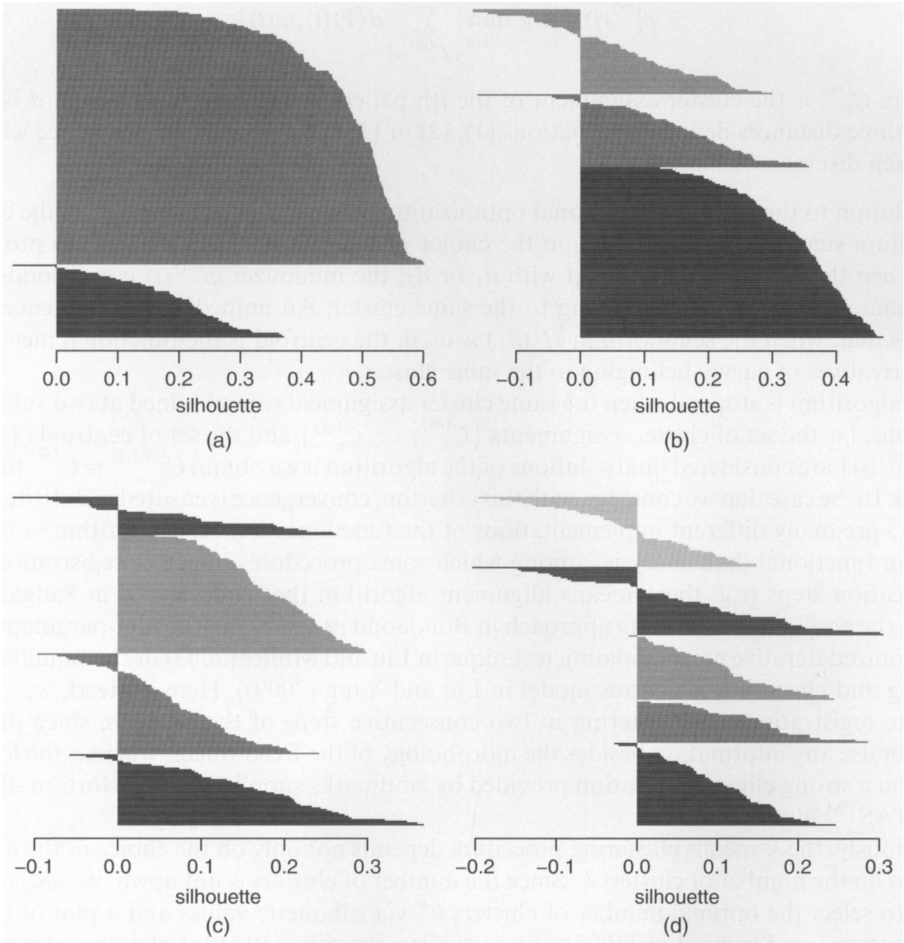

Fig. 6. Silhouette plots of the clustering result obtained via the multivariate functional k-means procedure, setting (a) $k=2$, (b) $k=3$, (c) $k=4$ and (d) $k=5$ and with distance given by equation (1): the data are ordered according to an increasing value of silhouette within each cluster and are coloured according to the cluster assignment

traces. So, we perform clustering of the whole data set via the multivariate functional k-means algorithm that was previously described, using the different distances between curves given in equations (1)–(3).

The final silhouette plots obtained by clustering the sample of 198 ECG traces according to a multivariate functional k-means procedure with distance d_1 (1), and setting $k=2,3,4,5$, are shown in Fig. 6. As we can see in the picture, the grouping structure that is obtained by setting $k=3$ seems the best, both in terms of silhouette profile, and in terms of wrong assignments. A similar result is obtained by measuring the distance between curves via equations (2) or (3); however, the procedure seems to detect the best grouping structure when both the curves and their derivatives are considered in the distance. We thus set $k^*=3$.

The final classification that is obtained by choosing this distance, and setting $k=3$, is shown in Fig. 7, where the whole functional data set is coloured according to cluster assignments; each panel corresponds to a different lead. Given the final cluster assignments, the cluster of healthy

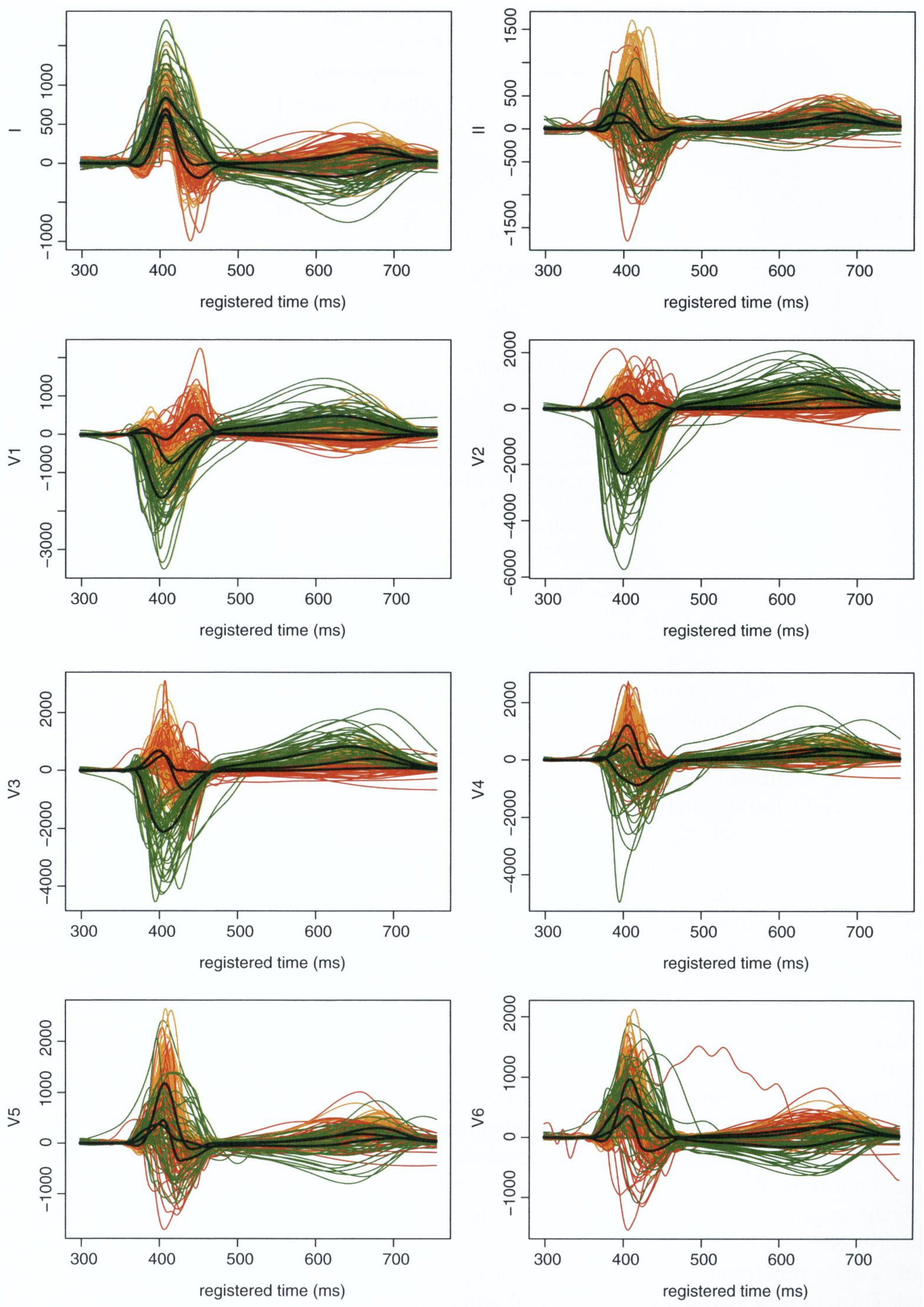

Fig. 7. Smoothed and registered ECG traces (*QT*-segment): the whole data set is coloured according to the final cluster assignment of the multivariate functional three-means clustering, with distance given by equation (1); the superimposed black curves are the three final cluster centroids (functional means); each panel corresponds to a different lead of the ECG trace

Table 3. Confusion matrices related to patients' disease classification†

Cluster	*Normal*	*RBBB*	*LBBB*
H^1-norm (equation (1))			
1	95	7	1
2	6	42	3
3	0	0	44
H^1-seminorm (equation (2))			
1	71	12	0
2	30	36	5
3	0	1	43
L^2-norm (equation (3))			
1	94	6	2
2	7	43	3
3	0	0	43

†The results are obtained by the application of a multivariate functional three-means clustering algorithm to smoothed and registered QT-segments of ECG curves, with different choices of the distance between ECGs. Clusters 1, 2 and 3 in the top panel respectively correspond to orange, green and red in Fig. 7.

patients is detected as the cluster that includes the most physiological traces. The pathological clusters are subsequently chosen, by selecting first the cluster that contains the maximum number of pathological traces of the same kind, and consequently the cluster that remains. Analysing this picture a different shape of ECGs assigned to different clusters can be immediately appreciated, especially looking at the final centroids (functional means) of each group, which are drawn in black in each panel. We shall now verify whether this difference in the ECGs' morphology across clusters is due to the different pathologies.

Since we have an indication of the different pathologies of the patients who are included in the sample, we can analyse the confusion matrix that is associated with the final cluster assignments, compared with the Mortara–Rangoni algorithm classification (normal, RBBB and LBBB). The confusion matrices that are obtained via multivariate functional k-means with different choices of the distance between curves (1)–(3) are shown in Table 3. We remark that the final cluster assignments are based solely on the shape of the smoothed and registered ECG curves and their first derivatives, analysed via an unsupervised classification procedure.

Choosing the H^1-norm and the L^2-norm, both results seem significant: the final grouping structure identifies out quite effectively the patients' disease classification, with few wrongly assigned cases. Moreover, we note the improvement in the results that are obtained via multivariate functional three-means compared with the results of the three-means clustering algorithm on interval lengths (see Table 2): we can now not only detect pathological subjects but also distinguish between the two different pathologies in the data set. The result that is obtained via multivariate functional three-means clustering with H^1 seminorm, instead, is not so positive, since clusters 1 and 2 apparently merge physiological traces with ECGs of patients affected by RBBBs.

The effectiveness of the clustering procedure in detecting the grouping structure in data suggests the definition of a semi-automatic diagnostic tool based on the multivariate functional

Table 4. Mean misclassification costs and standard deviations computed via equation (4) over 20 repetitions of the cross-validation procedure

	Results for the following distances:		
	d_1	$\tilde{d}_1$	d_2
Mean cost_{CV}	0.1227	0.2213	0.1281
Standard deviation cost_{CV}	0.1113	0.09794	0.1236

k-means algorithm: 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. We could have an immediate hint on a new patient's diagnosis by smoothing his ECG trace, registering it and finally assigning it to the group that is characterized by the nearest centroid.

It is important to evaluate the *misclassification cost* for this procedure, with the choice of the different functional distances. For this, we perform a *cross-validation analysis*. We randomly choose between ECGs a training set of 80 normal subjects, 40 RBBBs and 40 LBBBs, for a total of $n_{\text{training}} = 160$ curves. A multivariate functional three-means clustering is performed on the training set selected; we then consider the remaining $n_{\text{test}} = 38$ curves, and we assign each of them to the cluster whose centroid is nearest, according to distances (1)–(3). Given the patients' disease classification, we compute the misclassification cost by using the index

$$\text{cost}_{\text{CV}} = \frac{\lambda_1 \, \text{misc}_{\text{N}} + \lambda_2(\text{misc}_{\text{RN}} + \text{misc}_{\text{LN}}) + \lambda_3(\text{misc}_{\text{RL}} + \text{misc}_{\text{LR}})}{n_{\text{test}}}, \tag{4}$$

where misc_{N} is the number of healthy patients assigned to a pathological cluster, misc_{RN} and misc_{LN} are the number of patients respectively affected by RBBB and LBBB who are assigned to the cluster of healthy patients, and misc_{RL} and misc_{LR} are the number of patients whose ECGs are detected as pathological, but whose pathology is wrong. The parameters λ_1, λ_2 and λ_3 are misclassification weights: they are chosen according to the suggestion of the clinicians, who believe that assigning a BBB patient to the cluster of healthy patients is approximately four times more serious than treating a normal subject as pathological, which indeed is twice as serious as assigning an RBBB patient to the LBBB cluster (or vice versa); to determine the values of the weights we introduce a further request, i.e. we ask cost_{CV} to be equal to 1 in the worst case when all normal subjects are classified as BBB and all BBB subjects are classified as normal. This leads to the choices $\lambda_1 = 0.4270$, $\lambda_2 = 1.7079$ and $\lambda_3 = 0.2135$.

We repeat this procedure 20 times, computing each time the misclassification cost according to equation (4): the mean and standard deviation computed along the 20 cross-validation repetitions are shown in Table 4. Distance d_1 seems to give best results, thus confirming our initial claim: we need both the registered curves and first derivatives to compare ECG morphology accurately.

We also show the average confusion matrices for each of the distances considered, which have been obtained by taking the mean of the confusion matrices along the cross-validation repetitions. The three average confusion matrices obtained for the different distances are shown in Table 5. The result is very good both for equation (1) and for equation (3), even if we can note

Table 5. Cross-validation results: average confusion matrices related to patients' disease classification†

Cluster	*Normal*	*RBBB*	*LBBB*
H^1-norm (equation (1))			
1	19.7	2.10	0.25
2	1.3	6.85	0.45
3	0	0	7.35
H^1-seminorm (equation (2))			
1	19.05	3.6	0.55
2	1.95	3.95	1
3	0	1.3	6.6
L^2-norm (equation (3))			
1	19.8	2	0.5
2	1.2	6.75	0.4
3	0	0	7.35

†The results are obtained on a test set of 38 patients, by application of the multivariate functional three-means clustering algorithm, with different choices of the distance between ECGs.

that the L^2-distance assigns slightly more LBBB patients to the cluster of the healthy patients than to the other pathological cluster. Moreover, these results confirm the considerations that we have made while looking at the results shown in Table 4.

5. Conclusions

In this work we proposed a statistical framework for the analysis and classification of ECG curves starting from their morphology only. We have analysed a database composed of 198 ECG traces—101 of them were normal, 49 were RBBBs and 48 were LBBBs—extracted from the PROMETEO data warehouse. The strongly localized features (peaks, oscillations, ...) of ECG curves make them particulary suited to be smoothed via wavelet methods, since every basis function is localized both in time and in frequency; for this, and to reconstruct smoothed curves together with their first derivatives, we used a Daubechies wavelet basis with 10 vanishing moments. Moreover, as the ECGs are 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 kinds of variability is necessary for a successful analysis, we register ECG traces, choosing a landmark-based procedure, which identifies as landmarks those time points that can be associated with a specific biological event. Five of them are provided by the Mortara-Rangoni VERITAS$^{\mathrm{TM}}$ algorithm, identifying the P-wave, the QRS-complex and the T-wave; we add one more landmark corresponding to the peak of R-wave on lead I, which is an easily localized feature on each ECG. In this way, we managed to separate morphological information of the curves (i.e. amplitude variability) from the duration of each ECG interval (i.e. phase variability).

We chose to analyse morphological information via multivariate functional k-means, thus simultaneously clustering all eight leads of each patient, with three different choices for the distance between ECGs, involving curves and/or first derivatives; our claim is that both the ECG trace and its first derivative are necessary to capture the morphological characteristics of

ECGs fully. The optimal number of clusters can be chosen via a measure of the goodness of the clustering results, and in all cases considered which have been taken into account it is set equal to 3. The confusion matrix resulting from our classification framework shows effective results, especially when the distance considers both curves and first derivatives, confirming our initial claim. Thus, we propose a classification procedure which uses group centroids as reference signals. This technique could help in the semi-automatic diagnosis of BBB-related pathologies. We perform a cross-validation analysis to evaluate the misclassification cost that is associated with this procedure: our algorithm performances seem very encouraging, especially when functional distance considers both ECG curves and their first derivatives. The classification procedure proposed is very general, owing to the flexibility in the definition of distance between functional data.

The innovative aspect of this paper lies in developing advanced statistical methods aimed at detecting pathological ECG traces (in particular, BBBs), starting only from morphological features of the curves. This allows for diagnoses that are consistent with clinical practice, starting from purely statistical considerations.

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 BBBs). Owing 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 distance between functional data, e.g. including higher order derivatives.

Acknowledgements

This work is part of PROMETEO. The authors thank in particular Dr Niccolò Grieco for clinical counselling and Ing Johan DeBie for technical support. Finally, we thank the Azienda Regionale Emergenza Urgenza, the 118 Milan Dispatch Center and Mortara Rangoni Europe s.r.l. (`http://www.mortara.com`) for supplying the data.

References

Antman, E. M., Hand, M., Amstrong, P. W., Bates, E. R. and Green, L. A. (2008) Update of the ACC/AHA 2004 guidelines for the management of patients with ST elevation myocardial infarction. *Circulation*, **117**, 269–329.

Boudaoud, S., Rix, H. and Meste, O. (2010) Core Shape modelling of a set of curves. *Computnl Statist. Data Anal.*, **54**, 308–325.

Daubechies, I. (1988) Orthonormal basis of compactly supported wavelets. *Communs Pure Appl Math.*, **41**, 909–996.

Einthoven, W. (1908) Weiteres über das Elektrokardiogram. *Pflüg. Arch.*, **122**, 517–548.

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. *Am. Hrt J.*, **40**, 163–211.

Fritsch, F. N. and Carlson, R. E. (1980) Monotone piecewise cubic interpolation. *SIAM J. Numer. Anal.*, **17**, 238–246.

Goldberger, E. (1942a) The aVL, aVR, and aVF leads: a simplification of standard lead electrocardiography. *Am. Hrt J.*, **24**, 378–396.

Goldberger, E. (1942b) A simple indifferent electrocardiographic electrode of zero potential and a technique of obtaining augmented, unipolar extremity leads. *Am. Hrt J.*, **23**, 483–492.

Grieco, N., Ieva, F. and Paganoni, A. M. (2011) Performance assessment using mixed effects models: a case study on coronary patient care. *IMA J. Mangmnt Math.*, **23**, 117–131.

Grieco, N., Sesana, G., Corrada, E., Ieva, F., Paganoni, A. M. and Marzegalli, M. (2007) The Milano Network for Acute Coronary Syndromes and Emergency Services. *Medit. Soc. Pacng Electphysiol. J.*, specl iss. 1.

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. *Communs Appl. Industrl Math.*, **1**, 128–147.

Lindsay, A. E. (2006) ECG learning centre. (Available from `http://library.med.utah.edu/kw/ecg/index.html`.)

Liu, X. and Müller, H.-G. (2003) Modes and clustering for time-warped gene expression profile data. *Bioinformatics*, **19**, 1937–1944.

Liu, X. and Yang, M. (2009) Simultaneous curve registration and clustering for functional data. *Computnl Statist. Data Anal.*, **53**, 1361–1376.

Mason, R. and Likar, L. (1966) A new system of multiple leads exercise electrocardiography. *Am. Hrt J.*, **71**, 196–205.

Pigoli, D. and Sangalli, L. M. (2012) Wavelets in functional data analysis: estimation of multidimensional curves and their derivatives. *Computnl Statist. Data Anal.*, **56**, 1482–1498.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis*, 2nd edn. New York: Springer.

R Development Core Team (2009) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Sangalli, L. M., Secchi, P., Vantini, S. and Vitelli, V. (2010) k-mean alignment for curve clustering. *Computnl Statist. Data Anal.*, **54**, 1219–1233.

Scher, A. M. and Young, A. C. (1957) Ventricular depolarization and the genesis of the QRS. *Ann. New Yrk Acad. Sci.*, **65**, 768–778.

Struyf, A., Hubert, M. and Rousseeuw, P. (1997) Clustering in an object-oriented environment. *J. Statist. Softwr.*, **1**, no. 4, 1–30.

Tarpey, T. and Kinateder, K. K. J. (2003) Clustering functional data. *J. Classificn*, **20**, 93–114.

Wilson, F. N., Johnston, F. D., Rosenbaum, F. F., Erlanger, H., Kossmann, C. E., Hecht, H., Cotrim, N., Menezes de Olivieira, R., Scarsi, R. and Barker, P. S. (1944) The precordial electrocardiogram. *Am. Hrt J.*, **27**, 19–85.