POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell'Informazione

Corso di Laurea in
Ingegneria Gestionale

Novel health monitoring technologies for the ageing society

Relatore: Prof. Bianca Maria COLOSIMO

Co-relatore: Ing. Marco GRASSO

Tesi di Laurea di:

Morena Lucia Emma ZECCHILLO Matr. 796182

Anno Accademico 2014 - 2015

Ringraziamenti

Desidero innanzi tutto ringraziare la Prof.ssa Bianca Maria Colosimo per avermi dato l'opportunità di poter realizzare un lavoro di ricerca in questo ambito.
Ringrazio fortemente anche l'Ing. Marco Grasso per la disponibilità, le competenze che mi ha trasmesso e per la stimolo che mi ha dato per migliorare sempre di più il mio lavoro.
Un ringraziamento doveroso va anche all'ing. Matteo Maggioni che mi ha supportato durante gran parte di questo lavoro.
Un ultimo, ma non per importanza, ringraziamento va a mia mamma, mio papà e a tutta la mia famiglia che mi è sempre stata accanto anche nei momenti più difficili.
Col termine famiglia mi riferisco anche a Mattia, Camilla, Federica, Marta, Christian, Ilaria, Laura e tutte quelle persone fantastiche che hanno reso questi miei anni di vita universitaria (e non) speciale.

Abstract

There is a continuously increasing demand for novel medical technologies, motivated by the challenges of the ageing societies and the need for more sustainable and innovative healthcare systems. An increasing interest is devoted to wearable sensors and the deployment of sensor networks aimed at collecting large amounts of data from patients in their everyday life, in order to guarantee a continuous monitoring of their health conditions. In this framework, a bulk of literature has been focused on ECG signal monitoring, in order to detect possible anomalies in a fast and reliable way. Mainstream methods involve the identification of salient features of interest and the computation of synthetic indexes that are compared with predefined thresholds. Those thresholds rely on gender and age group of the patients, but they are applied regardless of the specific pattern of each single patient. This yields non optimal monitoring results in most cases, in terms of high false alarm rates (Type I error) and missed detection of actual anomalies (Type II error). This thesis proposes a novel idea for ECG signal monitoring. It consists of determining the "signature" of the single individual under analysis, allowing a custom monitoring system that search for deviations from a specific "in control" condition instead of searching for deviations from standard conditions. The study is focused on the "design phase" of control charts, i.e., the phase during which in-control signal should be collected in order to estimate the control limits to be applied for future signal monitoring. In this phase, it is very important to understand if the patient conditions were actually in-control during the data collection period. Possible anomalous observations (hereafter denoted by "dataset contaminations") must be detected and removed, because they may have a detrimental impact on the estimation of control limits. A control chart scheme that couples traditional Statistical Process Control (SPC) methods and clustering techniques is studied and developed. The results show that it is suitable for profile monitoring of ECG signals.

Keywords: ECG, profile monitoring, SPC, design phase, decontamination, clustering

Sommario

Il progressivo invecchiamento della popolazione e la necessità di avere un sistema sanitario sostenibile e all'avanguardia ha portato ad una domanda di dispositivi medici innovativi in continua crescita.
Sempre maggiore è l'interesse legato sia a sensori indossabili, che allo sviluppo di reti di sensori, capaci di raccogliere grandi quantità di dati durante lo svolgimento delle attività quotidiane, in modo da garantire un monitoraggio continuo delle condizioni di salute dei pazienti.
In questo contesto, un considerevole numero di ricerche si è focalizzato sul monitoraggio del segnale ECG, in modo da identificare eventuali anomalie in modo veloce e affidabile.
I metodi tradizionali si basano sull'identificazione di alcuni parametri di interesse e sul calcolo di indici sintetici, che dovranno essere poi confrontati con soglie predefinite. Queste soglie variano col sesso e con l'età del paziente, ma non tengono conto di peculiarità legate allo specifico caso in analisi.
Ciò porta ad avere risultati spesso non ottimali, sia in termini di falsi allarmi (errore di I tipo) che di mancata identificazione di anomalie esistenti (errore di II tipo).
In questa tesi si propone quindi un'idea innovativa per il monitoraggio del segnale ECG. Essa consiste nella determinazione della "firma" del singolo individuo in analisi, permettendo di ottenere un sistema di monitoraggio personalizzato che lavori cercando deviazioni da una condizione standard di riferimento. Lo studio è focalizzato sulla "fase di progettazione" di carte di controllo, ovvero la fase in cui i segnali in controllo devono essere raccolti per poter stimare i limiti di controllo da applicare alla successiva fase di monitoraggio. In questa fase è molto importante capire se i dati raccolti provengono da una condizione del paziente effettivamente in controllo o meno. Osservazioni anomale, da qui in poi chiamate contaminazioni, devono essere identificate e rimosse, dal momento che possono avere un impatto negativo sulla stima dei limiti di controllo. Verrà quindi studiata e sviluppata una carta di controllo che unisca ai metodi tradizionali del Controllo Statistico di Processo (SPC) alcune tecniche di clustering. I risultati mostreranno che questa carta è idonea allo scopo di monitorare un segnale ECG.

Keywords: ECG, profile monitoring, SPC, design phase, decontamination, clustering

Index

Figures and Tables

Chapter 1

Introduction

1.1 Sensing technologies for ageing population

As Pollack described in [1], today approximately 10 percent of the world's population is over the age of 60; by 2050 this proportion will have more than doubled. Moreover, the greatest rate of increase is amongst the "oldest old" people aged 85 and over. While many older adults remain healthy and productive, this segment of the population is largely subject to physical and cognitive impairment at higher rates than younger people.
New technologies to support older adults and help them cope with the changes of aging started to be developed for this purpose in the last years, and there is a continuously increasing demand for advanced solutions in this field.
It is important to keep in mind that not only the absolute number of older adults is grown, but also the proportion of the population that is over the age of 65; thus, there will be fewer young people to help older adults cope with the challenges of ageing. While human caregiving cannot and will not be replaced, assistive technologies that can supplement human caregiving have the potential to improve the quality of life for both older adults and their caregivers. In particular, assistive technologies now being developed may enable older adults to "age in place," that is, remain living in their homes for longer periods of time. A large body of research, analysed and reviewed by Hareven in [2], has shown that older people prefer to maintain independent households as long as possible. Additionally, institutionalization has an enormous financial cost, not only for elders and their caregivers, but also for governments. Thus technology that can help seniors live at home longer provides a "win-win" effect, both improving quality of life and potentially saving enormous amounts of money. Other technologies can help those elders, who are in assisted living or skilled nursing care facilities, maintain more independence there. A range of remote monitoring, diagnosis and real-time signal analysis techniques has been used in the design of advanced assistive technologies: a great number of systems have been developed to help people compensate for the physical and sensory deficits that may accompany aging ([3], [4]).
Maintaining functional independence, in fact, is a high priority for many older adults and often, staying in their own homes is key to such independence.

Sensing technology, i.e., wearable sensors and sensor networks, has the potential to assist in this goal by supporting the everyday tasks of older individuals, as well as by aiding caregivers and family members ([5], [6]).

1.2 Statistical monitoring of ECG signals

The development of new technological sensors enables the realization of truly wearable and wireless instrumented garments capable of recording physiological signals; thereby, these particular clothing can be used by patients during everyday activities. Breathing pattern, electrocardiogram, activity sensors, pressure, temperature, can be listed as physiological variables which are likely to be monitored through these proposed devices. A miniaturized short-range wireless system, then, can be integrated in the sensitive garment and used to transfer the collected information.

In this frame, the focus of this thesis regards electrocardiogram (ECG) signals, which are easily recordable also through a simple holter.

The heart is comprised of muscle (*myocardium*) that is rhythmically driven to contract and hence drive the circulation of blood throughout the body. Before every normal heartbeat, or *systole*, a wave of electrical current passes through the entire heart, which triggers myocardial contraction. The pattern of electrical propagation is not random, but spreads over the structure of the heart in a coordinated pattern which leads to an effective, coordinated systole. This results in a measurable change in potential difference on the body surface of the subject. The resultant amplified (and filtered) signal is known as an electrocardiogram (ECG, or sometimes EKG).
A broad number of factors affect the ECG, including abnormalities of cardiac conducting fibers, metabolic abnormalities (including a lack of oxygen, or *ischemia*) of the myocardium, and macroscopic abnormalities of the normal geometry of the heart. ECG analysis is a routine part of any complete medical evaluation, due to the heart's essential role in human health and disease, and the relative ease of recording and analysing the ECG in a non-invasive manner.
Understanding the basis of a normal ECG requires appreciation of four phenomena:

- the electrophysiology of a single cell;
- how the wave of electrical current propagates through myocardium;
- the physiology of the specific structures of the heart through which the electrical wave travels;
- how that leads to a measurable signal on the surface of the body, producing the normal ECG.

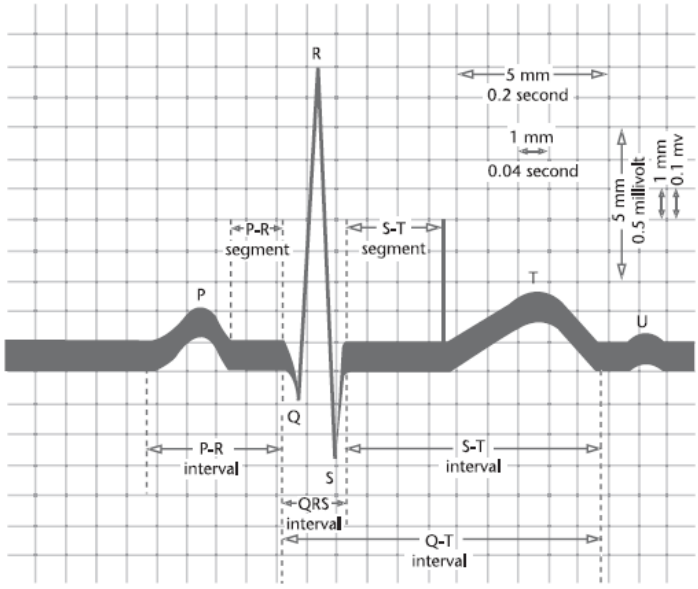

Figure 1.1: Representation of an heartbeat composition (From: [7]).

Fig. 1.1 illustrates the normal clinical features of a single heartbeat, which include wave amplitudes and inter-wave timings. Fig. 1.2, on the other hand, show the same waves in a real ECG signal.

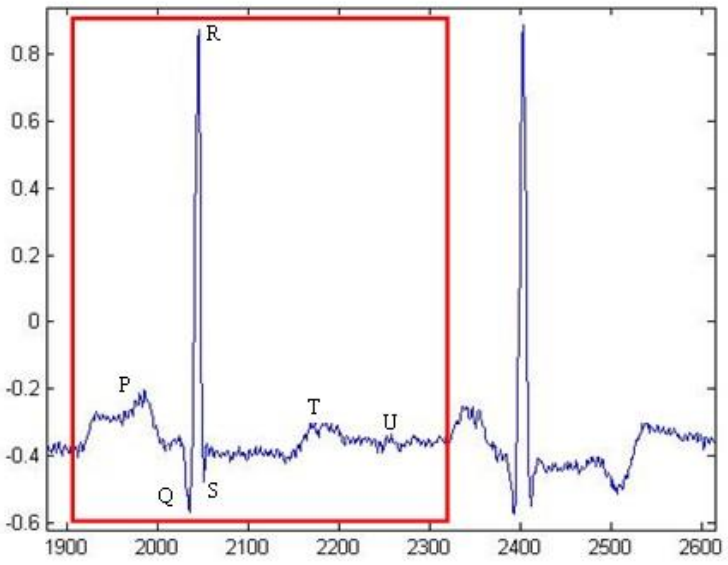

Figure 1.2 : A real heartbeat composition.

Some characteristic points can be find:

- P wave: during normal atrial depolarization, the main electrical vector is directed from the sinoatrial (SA) node towards the atrioventricular (AV)

node and spreads from the right atrium to the left atrium, which turns into the P wave on the ECG;
- Q-R-S waves: a Q wave is any downward deflection after the P wave, an R wave follows as an upward deflection, and the S wave is any downward deflection after the R wave. They are useful to determine the QRS complex, which reflects the rapid depolarization of the right and left ventricles;
- T wave: it represents the repolarization (or recovery) of the ventricles. The interval from the beginning of the QRS complex to the apex of the T wave is referred to as the absolute refractory period. The last half of the T wave is referred to as the relative refractory period (or vulnerable period);
- U wave: it is hypothesized to be caused by the repolarization of the interventricular septum; it normally has a low amplitude, and even more often is completely absent. If present, it always follows the T wave, and also follows the same direction in amplitude. If too prominent, it can lead to some pathology's suspects.

Also some time intervals, such as ST, QT and RR (which are the distances between the points characterized by those letters) can be observed, considering their usefulness in identifying some pathologies.
To a first approximation, electrical problems come in two forms: those which make the heart pump too slowly or infrequently (*bradycardias*), and those which make the heart pump too quickly (*tachycardias*). If the pumping is too slow, the cardiac output of life-sustaining blood can be dangerously low. If too quick, the cardiac output can be too low since the heart does not have time to fill; moreover, the heart can suffer damage (e.g., *demand ischemia*) when it tries to pump too rapidly.
Other anomalies that can be noticed are arrhythmia and metabolic abnormalities. An *arrhythmia* is any abnormal cardiac rhythm. One category of arrhythmias occurs when the trigger to depolarize originates outside of the SA node, in another part of the myocardium (common causes of ectopy include a drug effect, such as caffeine, or a viral infection of the myocardium, or other inflammation or damage of part of the heart , e.g. ischemia). When the ectopic beat originates in the atria, it leads to a *premature atrial beat*, also known as an *atrial premature contraction* (APC); when it originates in the ventricles, it leads to a *premature ventricular beat* or *ventricular premature contraction* (VPC). For what concern the metabolic abnormalities, a characteristic change is provoked by high serum potassium levels (*hyperkalemia*) which cause a high, pointed T wave, and ultimately, loss of the P wave and distortion of the QRS. *Hypokalemia*, on the other hand, causes an undulation after the T wave called a U wave.
Also abnormalities of the geometry of the heart can be revealed by an ECG analysis.

Hence, it is clear why finding a method to monitor the ECG in order to detect some of these conditions promptly assumes great importance.

While analysing the clinical electrocardiogram, it is important to use a systematic approach.
The steps followed by clinicians to identify abnormalities in the ECG are:
- Identify the QRS complexes;
- Identify the P waves;
- Examine the QRS complex in each lead;
- Examine the ST-T segments;
- Examine the T waves;
- Examine the QT interval.

Once an abnormality is identified, there are often several potential explanations, many of which lead to several ECG pathologies; hence, it may not be possible to determine the significance of the abnormality with certainty. To confirm a potential diagnosis from the ECG, other characteristic abnormalities are often sought.

Unfortunately, the ECG is often contaminated by noise and artifacts that can be within the frequency band of interest and can manifest with similar morphologies as the ECG itself. Some of the disturbance sources can be classified as:
- Power line interference;
- Loss of contact between the electrode and the skin;
- Movement of the electrode away from the contact area on the skin;
- Electrical activity due to muscle contractions;
- Baseline drift (usually from respiration);
- Data collecting device noise;
- Noise generated by other medical equipment present in the patient care;
- Quantization noise and aliasing;
- Signal processing artifacts.

Although each of these disturbance sources can be reduced by judicious use of hardware and experimental setup, it is impossible to remove all of them. Therefore, it is important to quantify the nature of the noise in a particular data set and choose an appropriate algorithm suited to the disturbance sources as well as the intended application.[1]

Novel technologies driven by evolving societal challenges demand of advanced capabilities aimed at translating “human judgements” into “computer

[1] (For more information about these topics: [7].)

judgements", i.e. trying to let the software evaluate an ECG on the basis of what a human being, the doctor, would have done in its place.

1.3 Challenges and goals

ECG signals are complex and need pre-processing operations before starting any kind of analysis.
After that, the general focus in the literature is mainly on the identification of some features of interest in order to make a comparison with some theoretical/empirical thresholds generally used to decide whether an heartbeat has anomalies or not.
These features can be generally referred to some interval duration or some mutation in the shape/amplitude of the waves that create each beat.
A challenging way of analysing this kind of data, proposed in this work, is treating each heartbeat as a profile in order to perform the so called "profile monitoring" ([8], [9]), which is the use of control charts for cases in which the quality of a process can be characterized by a functional relationship between a response variable and one or more explanatory variables.
In this case, the response variable is the ECG value and the explanatory variable is time.
In many cases of profile monitoring the in-control shape of the profile may be described in terms of a parametric model, in order to perform the monitoring task searching for shifts in the parameters of this model. In this case, control charts are based on the estimated parameters of the model from successive profile data observed over time.
In this way, the ECG will be segmented in order to have a general holistic view of all the heartbeats, which are going to be studied with the purpose of finding any deviation from their "regular" shape.
In fact, all the human heartbeats are composed by the same waves and phases, but this does not mean that all the humans have the same identical beat shape.

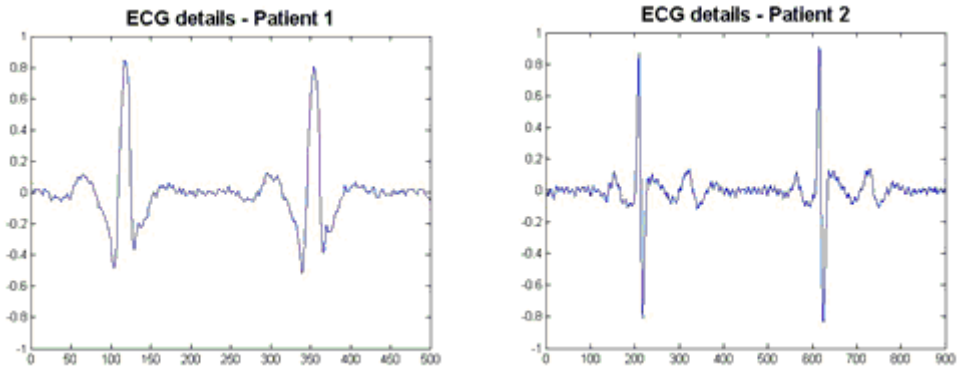

Figure 1.3: Heartbeats belonging to two different patients.

As can be seen in Fig.1.3, not only the shapes are different, but also patient 2 needs a longer time to show two heartbeats (900 points instead of 500. Each point is taken after the same amount of time).

Because of this, the novel idea of this thesis is to exploit the profile monitoring framework to estimate a signature of the ECG for each single individual, and to tune the control limits with respect to the specific signature. This implies a customization of the monitoring strategy, in order to enhance the capability of detecting actual anomalies by limiting the false alarm rates to a desired level.

Profile monitoring, as any other control charting schemes in SPC, involves two sequential phases. A design phase (phase I) aimed at estimating the parameters of the process and the control limits to be applied in the following monitoring phase (phase II). During phase I, one has to determine if the process was actually in-control during the data collection period. If out-of-control data are detected, they should be investigated in order to find assignable causes. If those causes are found, anomalous data (hereafter denoted by "contaminant data", which yield phase I contamination) should be removed and the process parameter estimation must be updated.

Generally speaking, collecting data from a steady process does not mean or assure the fact of always having in control data; real or false error can always occur. This can potentially influence the parameter estimation of phase I, as well as compromise the successive work of phase II.
The following example will clarify this point.

A group of fifteen elements will be used for the phase I construction of a control chart.
In normal conditions, these elements belong to a normal distribution, with mean $\mu = 1$ and standard deviation $\sigma = 0.5$.
The parameters needed to build the chart are the sample mean and sample standard deviation; in this way, it is possible to draw the central line and the two control limits, upper and lower, which are calculated as the sample mean $\pm$ 3 sample standard deviation. The chart obtained from these data can be seen in Fig 1.4(a).

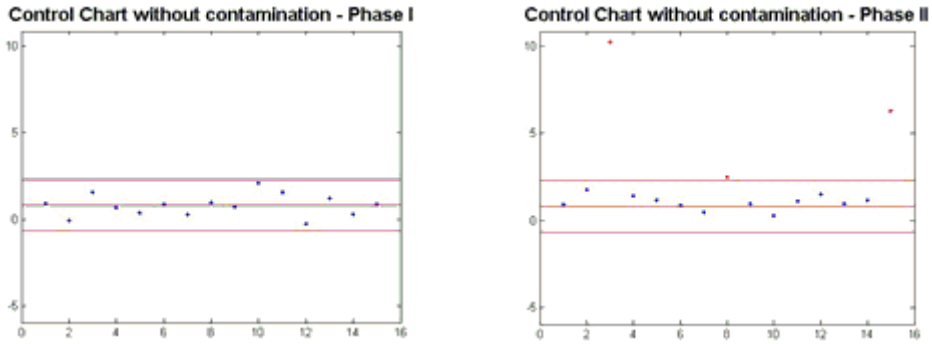

Figure 1.4: Control chart performances on a non-contaminated dataset.
(a) On the left, phase I control chart. (b) On the right, phase II control chart.

These parameters can now be used for the phase II analysis, in order to verify if a new group of elements deviates from the in control situation or not.
Three profiles with different mean and standard deviation, represented in red in Fig. 1.4(b), exceed the upper control limit and, for this reason, are correctly identified as anomalies.

A different conclusion is obtained if the dataset used for the phase I analysis is contaminated by some out-of-control data.
For example, in this case, three profiles, in red in Fig.1.5(a), with mean $\mu_{OOC} = 5$ and standard deviation $\sigma_{OOC} = 3$ contaminate the dataset.
Considering that these contaminating profiles has higher mean and standard deviation, also the sample calculations will be higher than the ones previously obtained; as a result, there is an inflation of the control limits that can be seen in Fig.1.5(a).

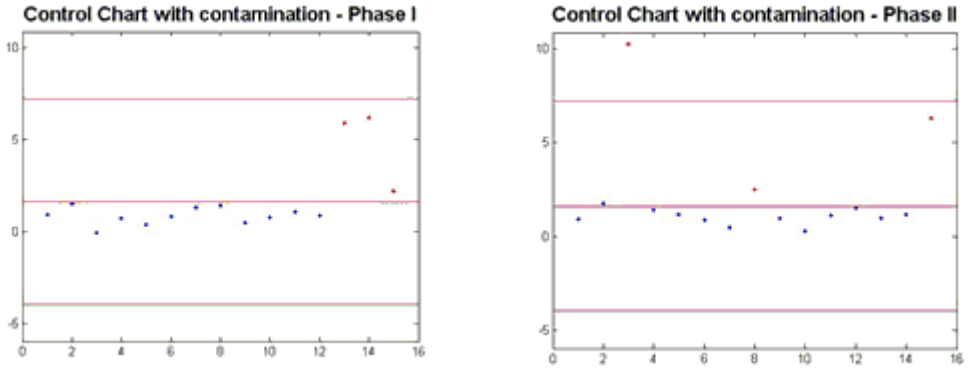

Figure 1.5: Control chart performances on a contaminated dataset.
(a) On the left, phase I control chart. (b) On the right, phase II control chart.

In Fig 1.5(b) it can be seen that this, in case of phase II analysis, prevent from the correct identification of 2/3 anomalies that have been easily identified in the case before, when no contamination was present during the estimation of the parameters needed to build the control chart.

The study is focused on the "design phase" of control charts, and the aim consists of developing a phase I profile monitoring methodology that is suitable to monitor ECG signals. Major challenges include: (i) how to define a signature that characterized the single individual, (ii) how to cope with misalignments and time stretching of profile data, (iii) how to enhance the detection of contaminations in phase I.
A control chart scheme that couples traditional Statistical Process Control (SPC) methods and clustering techniques is studied and developed. The results show that it is suitable for profile monitoring of ECG signals. Some critical issues are discussed, and the aspects that deserve future research are highlighted.

1.4 Outline

This work is structured as follow:

- in Chapter 2 the case study is presented;
- in Chapter 3 the control chart design is described together with some profile monitoring techniques and some methodologies that can be used to complete the decontamination task;
- in Chapter 4 the approach followed in this work will be introduced;
- its performances will be shown in Chapter 5;
- finally, conclusions and further directions will be presented in Chapter 6.

Chapter 2

A motivating case study

The data used in this work come from the MIT-BIH Arrythmia Database[2] , where different series of 30 minutes ECG data referred to different patients are collected.
The source of the ECGs included in the MIT-BIH Arrhythmia Database is a set of over 4000 long-term Holter recordings that were obtained by the Beth Israel Hospital Arrhythmia Laboratory between 1975 and 1979. The database contains 23 records chosen at random from this set, and 25 records selected from the same set to include a variety of rare but clinically important phenomena that would not be well-represented by a small random sample of Holter recordings. Each of the 48 records is slightly over 30 minutes long.
The first group is intended to serve as a representative sample of the variety of waveforms and artifact that an arrhythmia detector might encounter in routine clinical use. Records in the second group were chosen to include complex ventricular, junctional, and supraventricular arrhythmias and conduction abnormalities. Several of these records were selected because features of the rhythm, QRS morphology variation, or signal quality may be expected to present significant difficulty to arrhythmia detectors; these records have gained considerable notoriety among database users.
The subjects were 25 men aged from 32 to 89 years, and 22 women aged from 23 to 89 years.
Each record was given to two cardiologists, who worked on them independently to indicate abnormal beats and add comments.
Once both sets of cardiologists' annotations for a given record had been transcribed and verified, they were automatically compared beat-by-beat, and a chart recording was printed. This chart showed the cardiologists' annotations in the margin, with all discrepancies highlighted. Each discrepancy was reviewed and resolved by consensus.

The patient selected in this work is a 56 years old woman without any regular medication to take. Her regular heartbeat shape can be seen in Fig. 2.1.

[2] http://www.physionet.org/physiobank/database/mitdb/

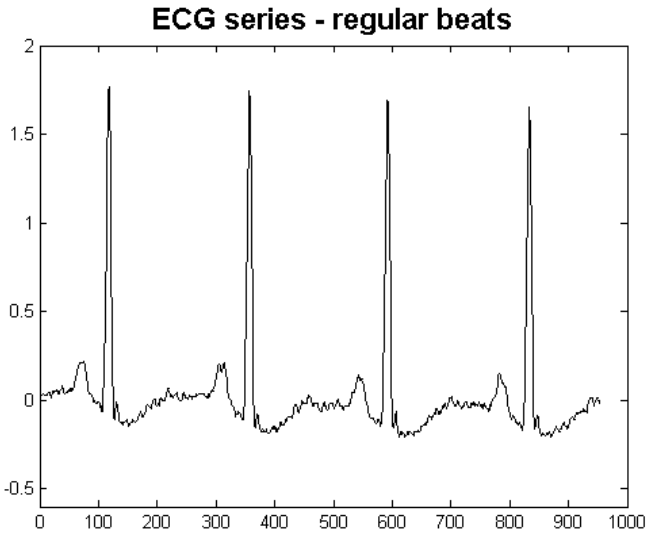

Figure 2.1: Patient ECG - example of regular heartbeats.

These data belong to the second group of records; hence, they contain some anomalies that are described in its annotation file.
Annotations reveal the presence of some heart beats that can be classified as nodal (junctional) premature beat, while others can be considered premature ventricular contraction.

A junctional premature beat is a delayed heartbeat originating not from the atrium but from an ectopic focus somewhere in the atrioventricular (AV) junction. It occurs when the rate of depolarization of the sinoatrial (SA) node falls below the rate of the atrioventricular node. This dysarhythmia also may occur when the electrical impulses from the SA node fail to reach the AV node because of SA or AV block. It is a protective mechanism for the heart, to compensate for the SA node no longer handling the pacemaking activity, and is one of a series of backup sites that can take over pacemaker function when the SA node fails to do so.
In the database under analysis, it affects some heart beats in a row.

Some noticeable signs that can be seen in these cases affect the QRS complex, which is generally narrow, either without a preceding P wave or with an abnormal P wave (usually inverted), and which occurs sooner than expected and is followed by a long compensatory pause.
In Fig. 2.2, both these characteristics are shown: P wave, which precedes the highest peak (R), is abnormal and QRS complex is not as wide as in a normal condition.

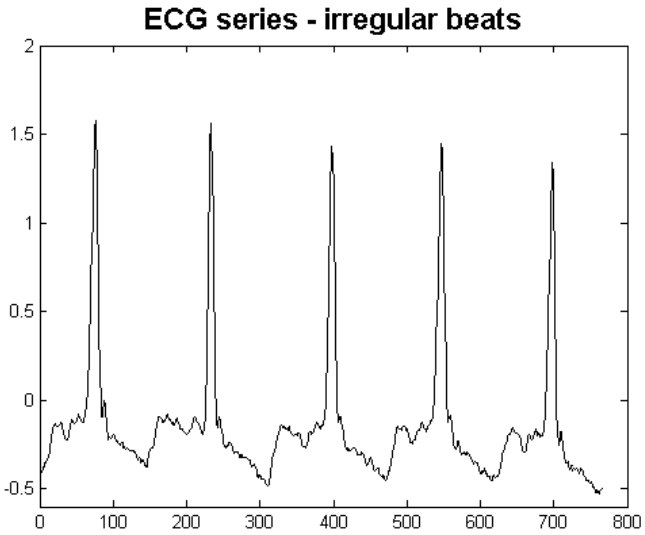

Figure 2.2: Patient ECG - example of junctional premature beats

A premature ventricular contraction (PVC) is a relatively common event where the heartbeat is initiated by Purkinje fibres in the ventricles rather than by the SA node, the normal heartbeat initiator. The electrical events of the heart detected by the electrocardiogram allow a PVC to be easily distinguished from a normal heart beat. Although a PVC can be a sign of decreased oxygenation to the myocardium (cardiac muscle) often PVCs are benign and may even be found in otherwise healthy hearts.
A PVC may be perceived as a "skipped beat" or felt as palpitations in the chest. In a normal heartbeat, the ventricles contract after the atria have helped to fill them by contracting; in this way the ventricles can pump a maximized amount of blood both to the lungs and to the rest of the body. In a PVC, the ventricles contract first and before the atria have optimally filled the ventricles with blood, which means that circulation is inefficient. However, single beat PVC arrhythmias do not usually pose a danger and can be asymptomatic in healthy individuals.
In the database under analysis, it affects some heart beats sporadically.

Some noticeable signs that can be seen in these cases affect the QRS complex, which can be broad with abnormal morphology, and premature. Some signs can be found also in the shape of T wave.

In Fig. 2.3, it is possible to see the strange behaviour of the abnormal beat's QRS complex.

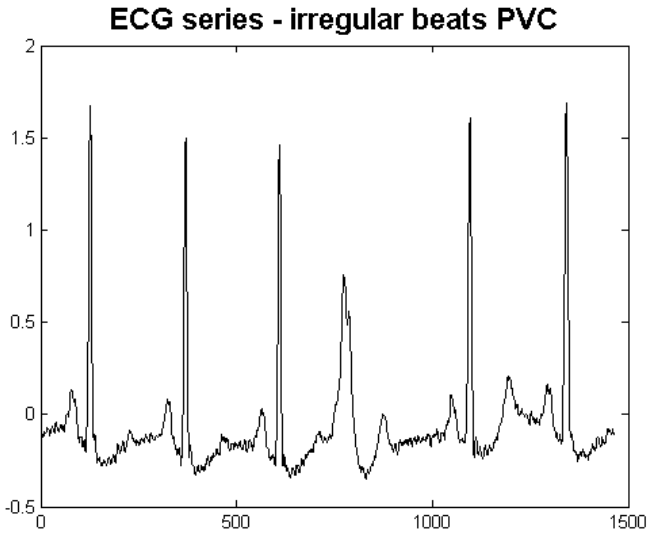

Figure 2.3: Patient ECG - example of premature ventricular contraction.

These annotations will be taken into consideration to verify, at the end of the analysis, if the method implemented is capable of detecting anomalies or not.

As already showed before, every heartbeat is constituted by the same waves, but every person has a peculiar heartbeat shape (Fig. 2.4).

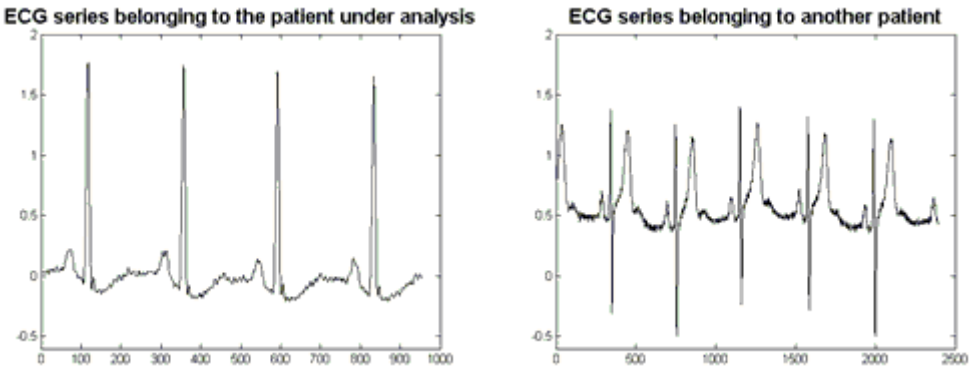

Figure 2.4: Difference between heartbeats belonging to two different patients.

Hence, in order to perform the analysis accordingly to a profile monitoring approach, this information will be exploited to take into consideration not only clinical anomalies that can be found in the ECG annotation file, but also any profile that is particularly different from the in-control pattern (i.e. the

"signature profile") and that is not labelled as "clinical anomaly" within the database.

Some examples of these kind of abnormal heartbeats referred to the patient under analysis can be seen in Fig. 2.5.

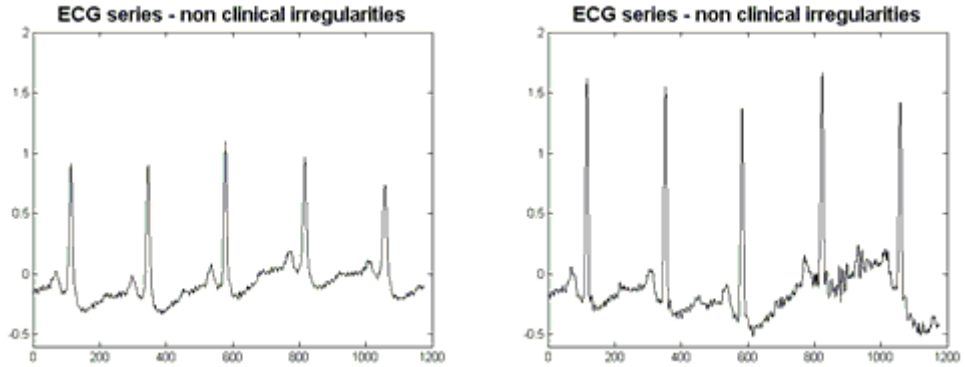

Figure 2.5: Examples of non-clinical irregularities in the patient ECG.

Therefore, considering the impact that abnormal values can have on parameters estimation, as shown in Chapter 1.3, also these beats which significantly deviates from the regular patient's signature, might be recognized and identified as anomalies.

The first thing to do when analysing an ECG data series is a pre-processing phase in order to filter out noises and baselines that represent a disturbance (i.e., not relevant source of information) and may have a detrimental effect on the following analysis (Fig. 2.6).

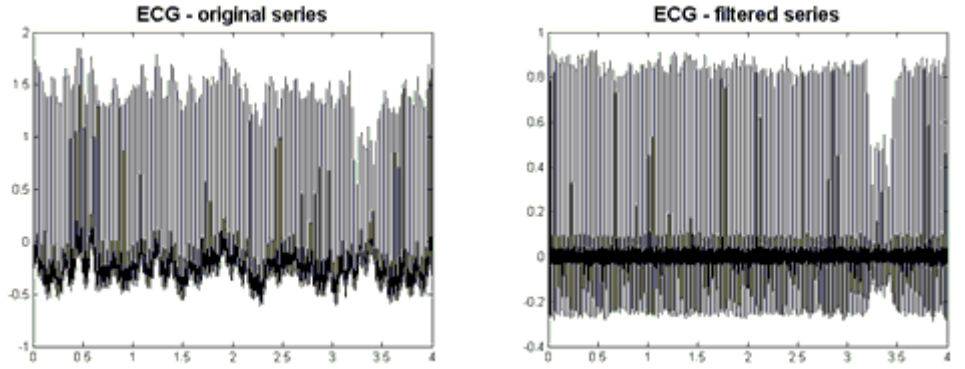

Figure 2.6: ECG series.
(a) On the left, the original ECG record. (b) On the right, the filtered ECG.

After that, the identification of each heartbeat is generally done with the purpose of measuring some characteristic features that can be compared to some medical thresholds in order to determine whether there is any anomaly or not.
In literature, different algorithms have been proposed for the detection of heartbeats, which is generally made identifying QRS complexes, but the key reference work remains the one proposed by Pan and Tompkins [10]. In their approach, the first step is to apply digital filters to reduce the influence of noise sources and, thereby, improve the signal-to-noise ratio. Then, some relevant information are extracted from the signal, such as the slope of the R wave, which is a popular signal feature used to locate the QRS. However, many abnormal QRS complexes with large amplitudes and long durations can be missed because of their relatively low R-wave slopes. Thus, this feature alone is insufficient for proper QRS detection. To achieve reliable performance, they extract also other parameters from the signal such as amplitude, width, and QRS energy. Using a dual-threshold technique, their method is able to locate missed beats and, thereby, false negatives can be reduced. In case of irregular heart rates, these two thresholds are reduced by half in order to increase the sensitivity of detection and to avoid missing valid beats. Once a valid QRS complex is recognized, there is a 200 ms refractory period before the next one can be detected, since QRS complexes cannot occur more closely than this physiologically. This refractory period eliminates the possibility of a false detection such as multiple triggering on the same QRS complex during this time interval. When a QRS detection occurs following the end of the refractory period but within 360 ms of the previous complex, it should be determined if it is a valid QRS complex or a T wave. In this case, the waveform with the largest slope is judged to be the QRS complex.

Some attempts have been done trying to implement alternative QRS detectors, like the methods proposed by Sayadi and Shamsollahi [11] or Madeiro, Cortez et al. [12], but none of them outperforms the Pan-Tompkins algorithm.

In this work, the identification of heartbeats has a final purpose which is completely different from the one generally aimed by traditional ECG monitoring methods; in fact, the goal is to determine whether there is any abnormal behaviour or not without using any medical threshold, but comparing each beat with a custom control limit (by adopting a patient's signature-based approach).
For this reason, heartbeats identification is not the only step that should be completed; another task of great importance is the segmentation of the whole ECG series into single beats that are going to be considered as a group of profiles that needs monitoring.

Each beat will be considered correctly segmented if all the characteristic points described before (P,Q,R,S,T) compare in it (i.e.: the segmentation operation must not yield any relevant information loss). In literature, this step is generally ignored due to the fact that the detection of common characteristic features does not necessarily need an exact segmentation of the data series. In fact, for example, if the height of an R peak or an RR interval is needed, the important thing is to identify these peaks; the fact that all the other points can be referred to a heartbeat or to the previous/following one do not have any particular relevance.

When doing "profile monitoring", however, all the points are relevant and need to be assigned to a specific heartbeat. Considering that the RR distance is generally not constant in the whole ECG, it is impossible to simply cut the beats considering a fixed number of points for each, regardless of the fact that some beats are longer than others.
For this reason, an algorithm capable of taking into account the difference between heartbeats' length will be proposed. This, however, would lead to profiles composed by a different number of points and, thus, not usable to directly perform profile monitoring analysis. Therefore, a profile scaling step will be needed in order to obtain profiles of the same length; this phase will be done by centering each profile on its R peak and considering separately the part that precedes this peak and the part that follows it; in this way, all the profiles will be also aligned with reference to a clearly recognizable landmark point (R).
The methodology will be explained in details in Chapter 4.

Chapter 3

Control chart design for profile monitoring

3.1. Profile Monitoring – a state of the art

In [8] Woodall described what profile monitoring is and how to cope with this kind of data.
Profile monitoring is the use of control charts for cases in which the quality of a process or product can be characterized by a functional relationship between a response variable and one or more explanatory variables. These cases appear to be increasingly common in practical applications.
For each profile it is assumed that p ($p > 1$) values of a response variable (Y) are measured along with the corresponding values of one or more explanatory variables (the X's).
This kind of data are also called functional data, due to the fact that the response variable is a function of all the explanatory variables and the same "functional form" is expected for every replicate, or profile, taken into consideration.
In many cases of profile monitoring it is efficient to summarize the in-control shape of the profile with a parametric model and monitor for shifts in the parameters of this model. The control charts are based on the estimated parameters of the model from successive profile data observed over time. For nonparametric methods one can alternatively monitor metrics that reflect the discrepancies between observed profiles and a baseline profile established using historical Phase I data.
Hence, either parametric models or nonparametric methods can be used in both phase I and II.
In many applications, however, a simple linear regression model is not sufficient to represent the shape of a profile; hence, more complicated methods are needed, such as multiple and polynomial regression, nonlinear regression models, mixed models, and so on.

An example of profile monitoring application with a parametric and nonparametric model approach is provided by Williams, Woodal and Birch in [9].
Starting from a dataset composed by different profiles, an approximating model is selected in order to obtain a set of coefficients used to describe each profile in a more concise way. With these coefficient, a parametric description of each element is obtained and a multivariate control chart can be built. Another way of proceed with the same dataset, is to identify a baseline (generally the mean

profile) and to evaluate the distance between each profile and this baseline. Different metrics can be used to evaluate distances; however, regardless to the metric used, these values will then be used to build a univariate control chart.

Before trying to apply the profile monitoring techniques to ECG data, it is important to go one step backward. The considerations made until this point, in fact, were made making comparisons between the same characteristic points in different profiles, with the implied assumption that the profiles under analysis could be easily overlapped showing a natural alignment along the x-axis.
This would mean that the only variation expected should be mainly on the y-axis, while the position of each point/timing would be the same regardless to the fact of being an IC or OOC profile.
This can be considered a sort of simplification; in fact, in reality, when working with data which are replicates of the same process in time, like heartbeats, some problems of "misalignment" along the x-axis are expected.
Some processes can show accelerations or decelerations, which can mutate the shape of the profile leading to wrong conclusions.

The problem of statistical analysis in function spaces is important in a wide variety of applications arising in nearly every branch of science, ranging from speech processing to geology, biology and chemistry. One can easily encounter a problem where the observations are real-valued functions on an interval, and the goal is to perform their statistical analysis. By statistical analysis is meant to compare, average, and model a collection of such random observations. These problems can, in principle, be addressed using tools from functional analysis. However, a serious challenge arises when functions are observed with flexibility or domain warping along the x axis.

In [13] Ramsay and Li explained this problem with a typical example generally used to study this problem, the Berkeley child's growth study. Ten estimates of the acceleration in height (Fig. 3.1) show, individually, the salient features of growth in children: the large deceleration during infancy is followed by a rather complex but small acceleration phase during late childhood, and then the dramatic acceleration-deceleration pulses of the pubertal growth spurt finally give way to zero acceleration in adulthood. The timing of these salient features obviously varies from child to child. Ignoring this timing variation in computing a mean function (bold broken curve) can result in an estimate of average acceleration that does not resemble any of the observed curves: the mean curve has less variation during puberty than any single curve, and the duration of the mean pubertal growth spurt is rather larger than for any individual curve.

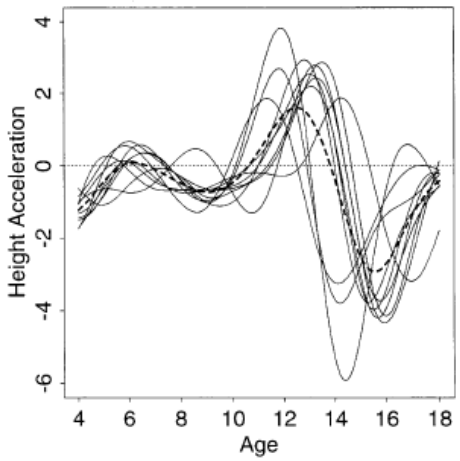

Figure 3.1: 10 height acceleration curve (in centimeters per year squared) for boys. (From [13])

This example illustrate that the rigid metric of physical time may not be directly relevant to the internal dynamics of many real life systems. Rather, there can be a sort of physiological timescale that relates non-linearly to physical time and varies from case to case. Human growth is largely a consequence of a complex sequence of hormonal events that do not happen at the same rate from child to child and also have a variable rate over the growth of a specific child. Put more abstractly, the values of two or more functions, may differ because of two types of variations; the first is the more familiar range variation, or vertical variation, due to the fact that two functions may simple differ at points of time at which they can be compared. But they may also exhibit a domain variation in that the two functions should be compared at two different times at which the two cases are essentially in comparable states. For example, the intensity of the pubertal growth spurts of two children should be compared at their respective ages of peak velocity rather than any fixed age.

The problem of transforming the argument of curves to align various salient features is referred to in different ways, such as curve registration, time warping or structural averaging. The process of aligning curves by identifying the timing of certain salient features in the curve, such as the point of zero acceleration during the pubertal growth spurt, is called marker registration. Using this strategy, curves are aligned by transforming time so that marker events occur at the same values of the transformed time. However, marker registration can present some problems, because marker events may be missing from certain curves, and marker time estimates can often be difficult to obtain.

One way to couple profile monitoring and curve alignment is the one proposed by Grasso et al. in [14], which consists of monitoring, at the same time, the coefficients of a parametric model of the signal and the coefficients of the warping function used for registration. It has been showed that this method may improve the detection performances, as the information loss is minimized.
In this way, in fact, it is possible to keep trace of the time warping effect introduced by the registration (alignment) step and, if a difference between different profile pattern is no longer appreciable by looking at the aligned profiles, it might be evident by looking at the warping coefficients.

The approach advocated by Grasso et al. in [14] will be applied in this study.

In the ECG case, some important characteristic should be present in each beat (i.e. the 5 waves P,Q,R,S,T); however, the most relevant part of the information is related to the QRS complex, which can be identified quite easily thanks to the prominence of the R-peak. Therefore, aligning each profile considering for convenience this peak as the reference point for every beat is a task that should be done before performing any profile monitoring analysis.
Some examples of this alignment phase on ECG data are presented in literature, also without the final aim of using any profile monitoring technique to analyse this kind of signal.

Vullings, Verhaegen, and Verbruggen in [15] proposed the usage of dynamic time warping (DTW [16]), a typical technique used for speech-recognition tasks, to align ECG beats. Their method follows different steps: first, they perform some filtering to remove high frequency noise, and next they approximate the filtered signal with lines. The end points of the lines are considered as a set of possible fiducial points. The segmentation of the ECG only depends on selecting the correct points among the (few) end points of the lines. Using a standard QRS detection algorithm from literature, they divide the ECG signal into separate heartbeats. Finally, every heartbeat is compared using DTW with a set of P, QRS and T waves, and the best matches are selected for the detection of the fiducial points. They demonstrated how it is possible to segment a heartbeat if there is the availability of a previously segmented heartbeat that looks similar. Hence, their alignment phase is based on the recognition of a pattern among all the patterns created putting together different segment of beats.
Some drawbacks show how it is important to perform a correct segmentation in order not to obtain strange results that can significantly distort the analysis; in fact, sometimes, the initial segmentation of an heartbeat can cause problems, such that the onset of a P wave is situated at the end of a period, introducing a large error. To reduce this effect, one could create overlap between successive heartbeats, or one could divide every period between two subsequent QRS

complexes in four instead of three segments (end of QRS complex, T wave, P wave, and begin of QRS complex) in order to perform a better research of typical patterns in those zones.

A review of several algorithms that have been developed to segment the ECG automatically is presented by Zifan, Saberi et al. in [17], who also present a new method based on adaptive piecewise constant approximation (APCA [18]) and piecewise derivative dynamic time warping (PDDTW). First, they perform some pre-filtering to remove high frequency noise and next approximate the filtered signal by the adaptive piecewise constant approximation method. Using a standard peak QRS detection algorithm from the literature (Pan-Tompkins), the signal is then divided into separate heartbeats. Finally, every heartbeat is compared using a set of P, QRS and T wave templates, and the best matches are selected for the detection of the fiducial points.
The initial phase is the typical pre-processing phase common to all the algorithm seen until now, and the final part is very similar to what has been explained in [15]. The difference is in the way to perform the recognition of fiducial points. Adaptive piecewise constant approximation (APCA) approximates each time series by a set of constant value segments of varying lengths such that their individual reconstruction errors are minimal. The PDDTW algorithm, applied after the pre-processing on the large set of adaptive approximations of heartbeats obtained, has the task of selecting those points which are most likely the searched fiducial points. Then, the alignment phase is performed using reference heartbeats.

Finally, in [19], Shorten and Burke focused their attention on different time warping approach that can be used to optimise the time alignment or normalisation of two independent signals in order to generate the most efficient match between them.
Therefore, the steps generally followed in these algorithms related to ECG series are more or less the same and can be described by Fig. 3.3.

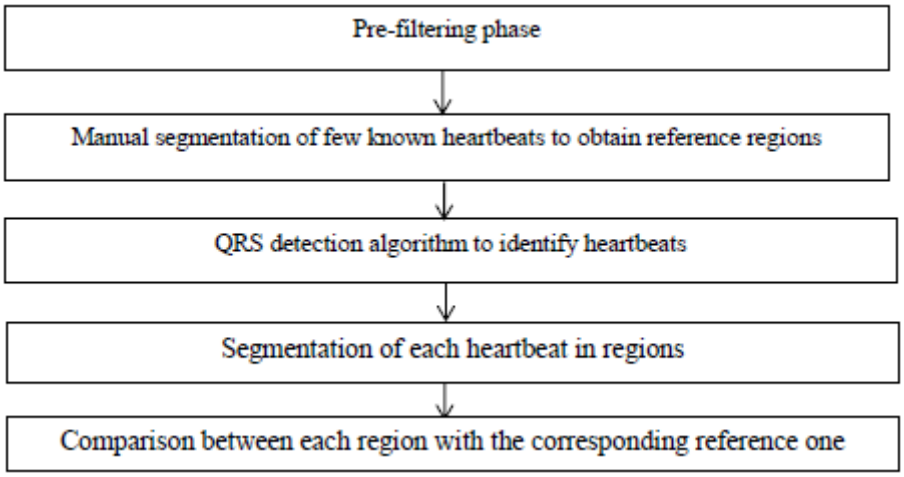

Figure 3.3: Traditional ECG alignment phase

To sum up, what is generally proposed in literature is to identify the R peak, which is the easiest one to find; then, the identification all the other peaks and valleys are strongly related to the reference profiles selected at the beginning of the analysis. Hence, if these reference profiles are not available, it is impossible to conclude the registration task.
For this reason, a more automated way will be followed in this work, in order to take advantage of the R-peak identification to detect a profile without taking into strong consideration all the other characteristic waves which are more difficult to detect with precision. In order not to lose information in the alignment phase, coefficients will be analysed too, as Grasso et al. suggested in [14].

3.2. Phase I decontamination

If the dataset used to design the control limits is contaminated by more than few special cause anomalies, then the monitoring performances can be seriously compromised, as shown in Chapter 1.
For this reason, it is always important to apply a sort of decontamination in order to prevent this degeneration of the control chart performances.

In literature, the task of Phase I decontamination in profile monitoring applications has been investigated by few authors. One reference study in this field is the one authored by Chen, Birch and Woodall [20], which is the baseline approach considered here for ECG profile monitoring. An in-depth analysis of the method is presented in the following section (Chapter 3.3), while some

variations introduced to apply this method to the case study under analysis are presented in Chapter 4.

However, some techniques that can be related to the idea of distinguish a group of abnormal elements from a group of normal ones are the ones linked to cluster analysis.

There are many ways to identify clusters within a group of elements:
- Density-based clustering methods consider clusters as regions with a particular high density of objects, separated by regions with really low density. A well-known algorithm based on this concept is DBSCAN [21];
- Grid-based clustering is based on the construction of a multi-resolution grid. After this grid is constructed, data are divided into different cells; the more populated ones are considered probable clusters. Some algorithms based on this method are: STING [22], WaveCluster [23] and CLIQUE [24];
- Model-based clustering tries to optimize the correspondence between the data and a mathematical model defined by the user; these methods can be used both with a statistical approach and with neural networks. In general, they are based on the EM (Expectation-Maximization) algorithm [25], which starts with a first phase of estimation of parameters and then, iteratively, tries to maximize an objective function which is related to them.

However, independently from the way of observing data, clustering is the task of grouping objects in such a way that the dissimilarity within the group is minimized and the one between groups is maximized.
There are two main ways that be followed to obtain this goal:
- Agglomerative clustering, which starts from different units, and lately groups them in order to obtain a certain number of clusters;
- Divisive clustering, which starts from the whole group and divides it in order to obtain a certain number of clusters.

In both these definitions, there is the necessity to reach a "certain number of clusters"; sometimes this value is known, some others not.

If the number of clusters is known a priori, it is a great hint because there is one less parameter to estimate.
Partitioning methods such as k-means and k-medoids are appropriated for this kind of situation.
Many times, however, there is the need to face the problem of identifying different clusters without knowing how much of them to expect.

Sugar and James [26] proposed a non-parametric method for choosing the number of clusters based on distortion, a quantity that measure the average

distance, per dimension, between each observation and its closest cluster center. A set of candidate cluster centers is first assumed, varying the number of clusters k. With a jump method, a knee in the k-distortion plot is found, and that is considered the ideal number of clusters. In order to reduce problems for selecting k connected to the monotony of the plotted function, a cross-validation or another measure of distortion can be evaluated in order to obtain another k value that will be compared to the first one.

Another typical way to proceed, is maximizing (or minimizing) a function evaluated for each selected number of clusters in order to determine the optimal number of cluster k*. Some indexes, like the gap statistic introduced by Tibshirani, Walther and Hastie in [27], are based on this principle. An application of this method is also proposed by Pedersen and Kulkarni in [28].

Once the number of cluster to search has been identified, it is possible to start the cluster analysis. While working with kmeans algorithms, it is important to identify those k points that will be used as cluster centers and will allow for the correct identification of the k clusters. Hence, if these centers are wrongly identified, the resulting clustering will not reflect the real structure of the data under analysis.
Different methods have been proposed in order to try to mitigate this kind of problem. As examples, Muhr and Granitzer in [29] proposed a k-means approach extended with a split and merge step. Basically, this split and merge k-means creates an initial partitioning with a predefined number of clusters. Afterwards, consecutive split and merge steps are done and the changes on the cluster result are assessed using some internal validity measure, like the Bayesian Information Criterion (BIC). Those split and merge steps are repeated until changes no longer improve this criterion. A similar approach has been proposed also by Pelleg et al. in [30].

Considering that not only the number of cluster is often critical to obtain, but also that the following identification of k clusters can be influenced by the initialization of some parameters, kmeans clustering method are hardly usable for the phase I decontamination task.
In fact, no information about the contaminating sources are available, and, moreover, some abnormal clusters can be made up of few elements, which are difficult to detect.
For this reason, hierarchical clustering methods could be more apt to this purpose.

Hierarchical clustering can be used with the support of a graph, the dendrogram, that can be useful to have a visual image of how many clusters can be obtained

using the grouping technique selected. Therefore, there is no need to have a priori knowledge about the number of cluster to search for.
Generally, this kind of clustering technique is user-driven, in the sense that the user decide how to cut the dendrogram after having visualized it.

Some automatic approaches, however, have been proposed also to face another problem; like k-means methods, in fact, also agglomerative clustering rely on the choices made at the beginning of their implementation. Therefore, once a merging choice has been done, there is no way to return on previous decisions.

Ferraretti, Gamberoni and Lamma in [31] analyse the possibility of having different levels of details for different regions of the space. Their automatic approach, in fact, drive a search to the best cluster structure "cutting" the dendrogram with a non-horizontal border. In this way, it has been added to hierarchical clustering a further step that can provide an effective cluster partition, optimizing a chosen evaluation index. By using the non-horizontal cutting, this technique performs an index-based exploration of the clustering tree, automatically extracting a clustering partition. It is possible to explore the clustering tree in several ways and two different methods, based on the selection of the node to open, have been proposed; the first one starts by choosing the node that brings to the clustering with the best global index (Go-to-best search) while the second one by choosing to open the node with the worst specific index (Expand-worst search). The iterative exploration of the tree is going to be stopped stop when the obtained clustering solution does not improve the selected index (greedy approach). The Expand-worst search, however, seems to provide the most significant results.

Sander et al. in [32], investigated the relation between dendrograms and reachability plots, introducing methods to convert them into each other since they essentially contain the same information. A reachability plot is made by peaks and valleys; it is a representation of regions having relatively low reachability values, the valleys, separated by higher reachability values, the peaks. The reachability values that separate clusters are local maxima in the reachability plot; however, not all local maxima points are separating clusters and not all regions enclosed by local maxima are prominent clusters. To ignore regions that are too small in the reachability plot, a minimum cluster size is generally assumed. In the first step of the algorithm, all points whose reachability value is a local maximum are collected; then, the local maxima points are sorted in descending order of their reachability value and processed one by one to construct the final cluster tree. The procedure always removes the next largest local maximum from the maxima list until the list is empty, and

determines whether a split point justifies new nodes in the cluster tree and, if it does, where those nodes have to be added.

In both these examples it is possible to explore some clusters in depth, in order to understand if a better representation is obtainable dividing a cluster in more parts; however, all the decisions are based on the initial agglomeration done and, what is more, some starting merging may never be investigated.
However, in case of decontamination of phase I, a main stable cluster composed by the majority of elements is expected and, for this reason, should be correctly identified from the beginning of the cluster analysis. Other small clusters, composed by contaminating elements, can also be present but the fact that they belong to one or more clusters does not have any importance for the final purpose of decontamination.
For this reason, hierarchical clustering could be a useful methodology to complete this task.

Today, high-dimensional data are very common to analyse; hence, there is also the need to reduce the dimensionality before applying any of these clustering algorithms.
In case of profile monitoring, this can be done simply finding a model to describe data and working on these coefficients instead of working on the original raw data.
However, other dimensional reduction techniques can be find in literature.
Some are feature selection techniques, like the ones proposed in [33], [34] and [35]. In these cases, some features to describe data are already known, and just some of them should be selected in order to efficiently perform analysis.
An additional step is made in [36] by Yu and Liu, who showed that feature relevance alone is insufficient. In fact, in theory, more features should provide more discriminating power, but in practice, with a limited amount of training data, excessive features may confuse and slow down the analysis. Hence, feature redundancy is defined and an algorithm is proposed to perform explicit redundancy analysis in feature selection to take advantage from the decoupling of relevance and redundancy analysis.
Furthermore, in [37] and [38] it has been showed that clustering results can strictly be linked to the features selected to describe data; hence, in order to obtain better results it could be useful to perform these tasks simultaneously.

Another alternative, when no features to describe data are known, is using some methods that can both identify features and find a way to select just some of them. PCA, MCA or ICA, are some of the methods that can be used to do so. First they rotate and translate original data obtaining a new reference system; then, some of these new directions are selected considering a specific criteria. In

case of PCA, for example, the criteria generally used is trying to capture a certain amount of the total variation observed in the data selected just few components, the principal ones.
However, other criteria can be used; in [39] Zhou and Jin analysed PCs also taking into consideration their level of Gaussianity, in [40] and [41] PCs are analysed together with MCs in order to delete components already described through their linear combination, in [42] and [43] ICs are presented and analysed in case of deviations from the underlying hypothesis of non-normal data.

In case of a contaminated dataset, trying to determine few directions where the maximum variation can be observed can be misleading. In fact, some in control elements can be well described by some particular directions, while contaminating ones can be well described by others. Therefore, putting together these information can lead to the selection of directions that do not really permit a complete and effective description of these data.

For these reasons, the model-based approach used in [20] seems to be a suitable choice also for describing the data of the case study proposed in this work; in fact, no hypothesis on any feature will be done, avoiding any clustering complications, and no distortion will be introduced observing data through wrong directions.
After this general overview, some examples will be provided to see how generally clustering is made on profile-data.
The profiles that will take into consideration are time series, which can be considered quite similar to ECG data.

First of all, a general review of the most recent clustering techniques applicable to these kind of data has been done by Rani and Sikka in [44]. Clustering of time series can be organized into three groups depending upon whether they work directly on raw data (either in frequency or time domain), indirectly with the features extracted from the raw data or with model built from raw data. It could be shape-level, if it is performed on many individual time-series, or structure-level, if it works on single long-length time-series. Gisbrecht in [45], made a comparison between these two ways of proceed and pointed out that, surprisingly, clustering a whole data series can likely yield meaningless results.
Hence, making a parallel with the dataset that will be used in this work, the fact of working on separate heartbeats instead of working on the entire ECG series, appears to be a right choice.
Going back to [44], time-series clustering can be divided in:

– Temporal-Proximity-Based Clustering if it works directly on raw data;

- Representation-Based Clustering if it works indirectly with the features extracted from the raw data;
- Model-Based if it works with model built from raw data.

The first group of methods can vary just for what is going to substitute the distance/similarity measure generally used with static data.
Examples of the second method can be [46] and [47], where the problem of feature selection is explored. In particular, in the first case the general problem of feature extraction is tackled by means of the extraction of discriminative patterns from temporal signals, assuming that classification could be done by combining in a more or less complex way such patterns and showing that good results can be obtained, though sacrificing interpretability.
Finally, examples of the third method can be [48] and [49], where autoregressive model based approaches are used.

Often time series are analysed also considering their dynamicity.
For these reason, also topic related to data streams ([50], [51], [52], [53]) and evolving clusters ([54]) are reasons of interest, since they can provide a better understanding of the underlying process.

However, this is not strictly related to the task of decontamination, where the final goal is just the identification of anomalies, regardless to their classification. This topic can acquire more relevance in the following phase II analysis, where it is important to deeply understand the process.

3.3. A cluster based approach

The study authored by Chen, Birch and Woodall [20] is the reference methodology considered in this thesis.

The method is composed by different steps that allow a continuous improvement of the decontamination task over a selected database.
It is called "cluster based approach" because it starts with cluster analysis to obtain a stable cluster of element that can be considered referred to in control situation and, then, updates this cluster through the cyclic usage of a control chart.
A synthetic scheme representing these steps can be seen in the following picture.

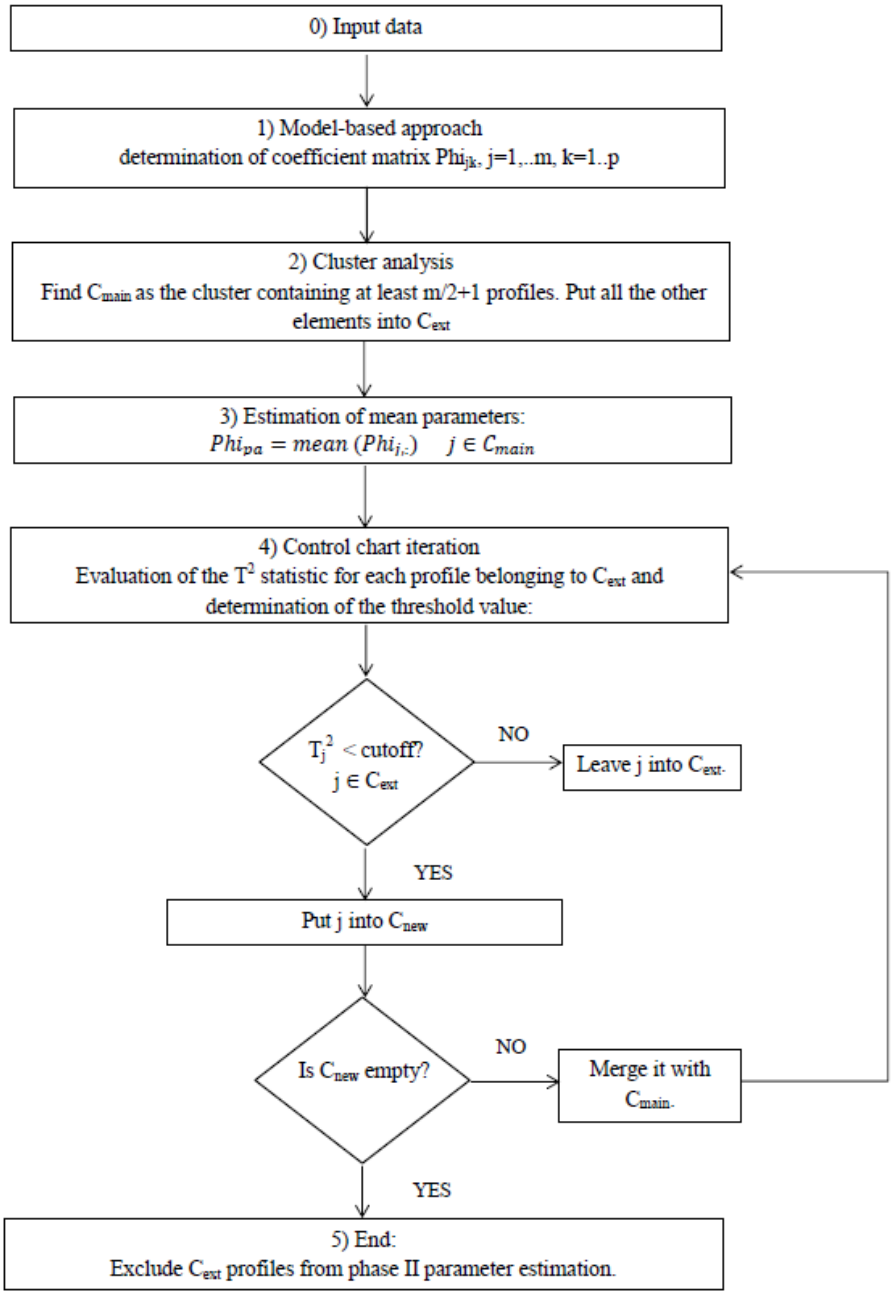

Figure 3.4: Cluster-based approach [20].

0) Input data
First of all, the analysed data are *m* profiles with equal length *n*; hence, this method can be classified as a profile monitoring methodology. In the dataset under analysis, just one kind of anomaly is present, and all the profiles that show this behaviour are positioned in succession.
Moreover, all the profiles are aligned along the x-axis.

1) Model selection
The preliminary step is the definition of a model to describe the process under analysis; having previous knowledge about it can help in suggesting which kind of model can be more suitable for this scope. In this way, it could be possible to switch from working directly on profiles to work on model coefficients.
Sometimes, however, some relations between variables are not known a-priori; hence, more general regression models can be adopted. In particular, in their work, the choice went over 1st order polynomial splines with a fixed number of equi-spaced nodes.

The general form of a polynomial spline is:

$$f(x_{i,j}) = b_0 + \sum_{l=1}^{G} b_l {x_{i,j}}^G + \sum_{k=1}^{N_k} u_{p,k} {(x_{i,j} - K_k)_+}^G$$
$$\sum_{k=1}^{N_k} {u_k}^2 \leq c \quad c \in [0, \infty) \tag{3.1}$$

Where:
- i = 1..n, with n that indicates the number of points that compose each profile;
- j = 1..m, with m that indicates the number of profiles under analysis;
- G is the polynomial degree;
- N_k is the number of nodes used;
- $(x_{i,j} - K_k)_+^G$ is set equal to 0 if $x_{i,j} \leq K_k$;
- c is a smoothing constant.

These way of modelling profiles make possible to describe both a global shape, through a complete p-degree polynomial, and a local behaviour, using a p-degree monomial for each defined segment.
Hence, if some abnormal profiles show a different behaviour in a particular part of the domain, this could be locally captured by the relative portion of spline.

There exist different algorithm for the selection and positioning of the optimal number of nodes and for the determination of splines function's degree.

Once the choice about which model to use is done, it is possible to use it to describe each profile through its coefficients so that, from now onwards, the focus will be no more on the original data *y* but on the set of estimated coefficients *Phi*. Therefore, each profile will be described by $p<n$ coefficients, where $p = N_k + G + 1$.

2) Cluster analysis

The first thing that needs to be done is to evaluate how similar or dissimilar these sets of coefficients are. Hence, an appropriate distance measure should be selected.

Their choice went on the squared Mahalanobis distance between each couple of profiles, which is a multi-dimensional generalization of the idea of measuring how many standard deviations away an element (profile) is from another.

This distance keep into consideration both the fact that the variances in each direction can be different and the existence of covariance between variables.

Hence, the similarity matrix *S* can be obtained as follow:

$$S(f,g) = (Phi_{f,:} - Phi_{g,:})V^{-1}(Phi_{f,:} - Phi_{g,:})^T \qquad f,g = 1..m \tag{3.2}$$

V is the estimated variance of the dataset. Variance estimation is strongly influenced by the situation under analysis; different estimators have strengths and weaknesses and should be selected accordingly to the kind of variation expected. In their work, the anomalies that contaminate the dataset consist of a set of subsequent profiles at the end of all the in control profiles; therefore, a successive difference variance estimator is an appropriate choice to evaluate the dispersion.

$$V = \frac{1}{2(m-1)} \sum_{j=1}^{m-1} (Phi_{j+1,:} - Phi_{j,:})^T (Phi_{j+1,:} - Phi_{j,:}) \tag{3.3}$$

With these information, it is possible to perform a cluster analysis; complete linkage hierarchical clustering is the choice made by the authors to complete this task.

In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. This kind of linkage tends to show preference for compact clusters with small diameters over long, twisted clusters; for this reason it appears to be quite sensitive to outliers, which are supposed to be different and, for this reason distant enough to be enclosed into separate clusters.

The aim of this analysis is to find a main cluster (C_{main}) which contains more than the half of the profiles under analysis (50%+1); this is a reasonable hypothesis, considering the fact that the datasets used to perform a phase I analysis should be composed by data collected from a process under control, which is not expected to show many abnormal behaviours.
Moreover, this choice permits to avoid any a priori decision about the number of cluster to search which, as already explained before, is often a parameter whose determination is not particularly easy.
All the other profiles excluded from C_{main} will be put into another cluster, C_{ext}.

3) Estimation of the mean profile
Once the main cluster has been individuated, the mean profile (y_{pa}) and its coefficients (Phi_{pa}) can be estimated; they are necessary to study all the other profiles which have not been put into the main cluster.

4) Control chart iterations
For every profile excluded from the main cluster, the T^2 statistic is computed and confronted with a cutoff value.

$$T_j^2 = (Phi_{j,:} - Phi_{pa})V^{-1}(Phi_{j,:} - Phi_{pa})^T \qquad j \in C_{ext} \tag{3.4}$$

$$cutoff = \chi^2_{1-\frac{\alpha}{m}}(p-1) \tag{3.5}$$

This is based on the underlying hypothesis that estimated coefficients follow a Normal distribution, which allows the use of a chi-squared cutoff;

In each iteration, all the C_{ext} profiles which have a T^2 value lower than the cutoff, will be added to a temporary cluster $C_{new.}$ After analysing them all, those profiles put into C_{new} will be added to C_{main} and the mean profile and its coefficients will be updated.
This operation is repeated until no more profiles can be added to the main cluster.

5) End
In the end, all the profiles belonging to the last C_{ext} cluster represent out of control (OOC) elements that should be excluded from phase II parameters estimation.

Chapter 4

The proposed approach

Starting from the method described in Chapter 3.3, some variations have been introduced, in order to adapt the methodology to the present problem. In Fig. 4.1, a synthetic scheme of the method is shown (it is possible to notice these variations as they are highlighted with red boxes).

The description of each step of this modified cluster-based methodology is provided below.

-1) Filtering, segmentation and alignment
As described in previous chapters, ECG data are generally available as a series of heartbeats, which need to be filtered, segmented and registered before computing the analysis.
Hence, in order to perform the algorithm, a pre-processing phase is needed in order to filter the whole data series, segment it into single beats and register these beats in order to avoid any misalignment.

Filtering
The signal filtering phase will be done in parallel to the recognition of each beat, as suggested in literature, using the Pan-Tompkins algorithm [10] already presented in Chapter 2.
After this phase, the whole data series will appear filtered and the position of each R peak will be available.

Segmentation
Exploiting the information about the position of each R peak, profiles will be divided considering the length of each RR interval.
The first point of each profile will be the one positioned in the half way between the R peak under analysis and the one that precedes it; similarly to that, the last point of each profile will be the one positioned in the half way between the R peak under analysis and the one that follows it.
This fact can be easily explained with Fig. 4.2.

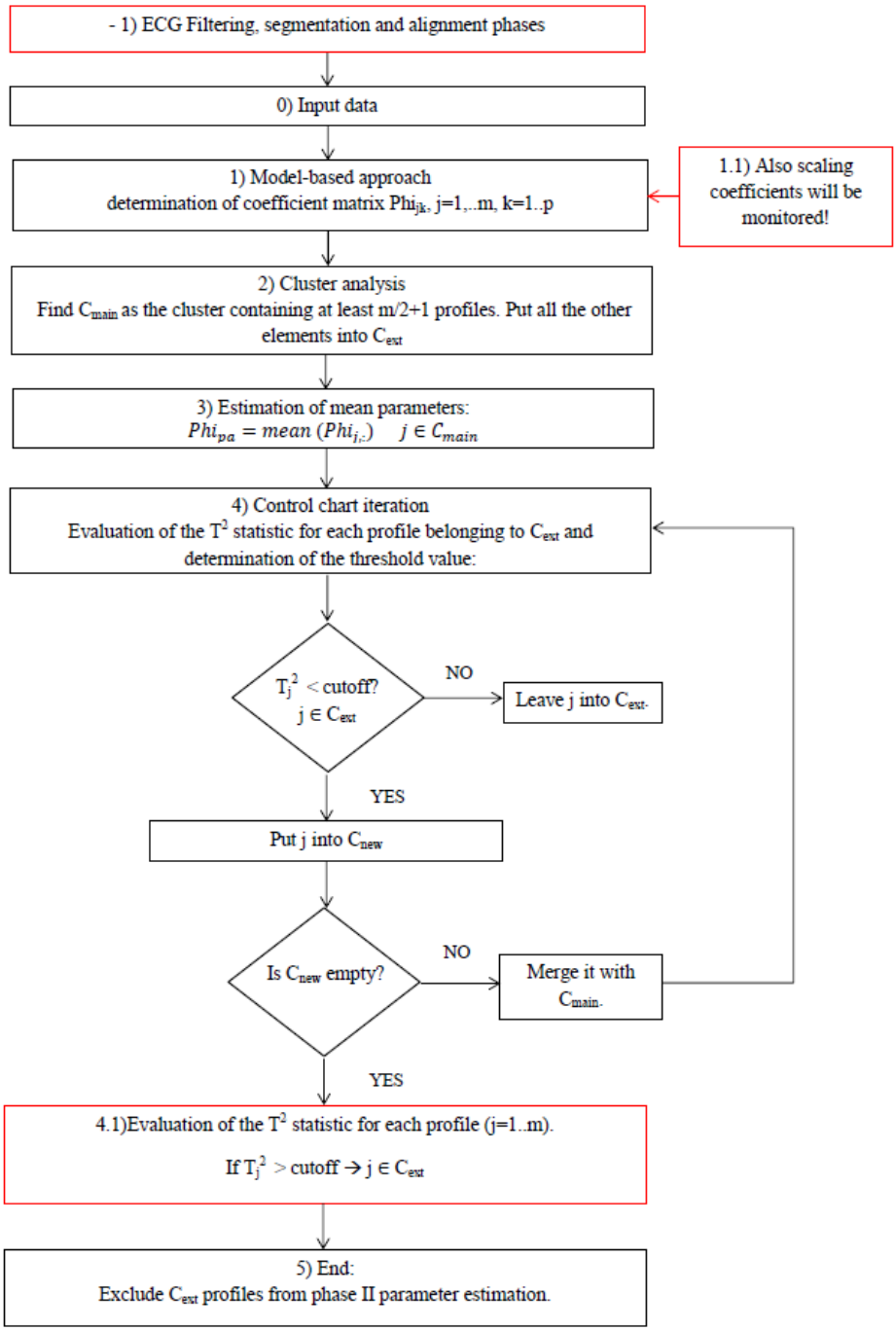

Figure 4.1 : The modified cluster-based approach

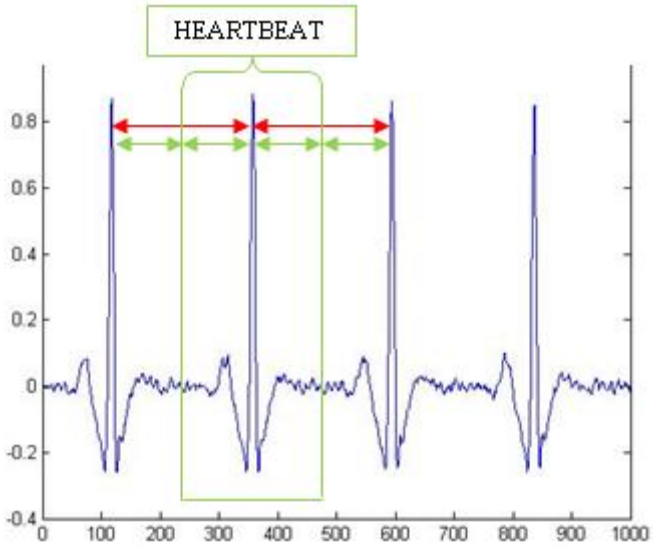

Figure 4.2 : Heartbeat segmentation

In this way, considering that RR intervals do not have a fixed duration, some profiles will be longer than others. Hence, they will need to be re-scaled in order to obtain equal length profiles. Considering that profiles will also need to be aligned, these two tasks will be made in parallel.

Alignment
Each profile will be centered on its R peak, and two separate scaling phases will be done on the segments that precede and follow this peak.
The reference length to consider in this scaling phase is the maximum length obtained for the segment under analysis; hence, the two segments of each profile will be warped in order to reach those target lengths described before.

Using two separated scaling factors permits to take trace of what had happened before and after the beat under analysis.
In fact, if this beat was anticipated by a premature beat and followed by a delayed one, resulting in a heartbeat of average length, using two different scaling coefficients would allow to take into consideration that the two segments have been caused by totally different situations, and thus two different and apt coefficients will be applied. Using just one scaling factor, on the contrary, would not help in keeping track of these behaviours, suggesting just an average warping coefficient that will not represent any of these two anomalies.

Moreover, in this way, profiles of the same length can be aligned considering the R-peak as a landmark point.

In Fig. 4.3, some profiles are plotted in order to see the difference before and after the scaling phase.

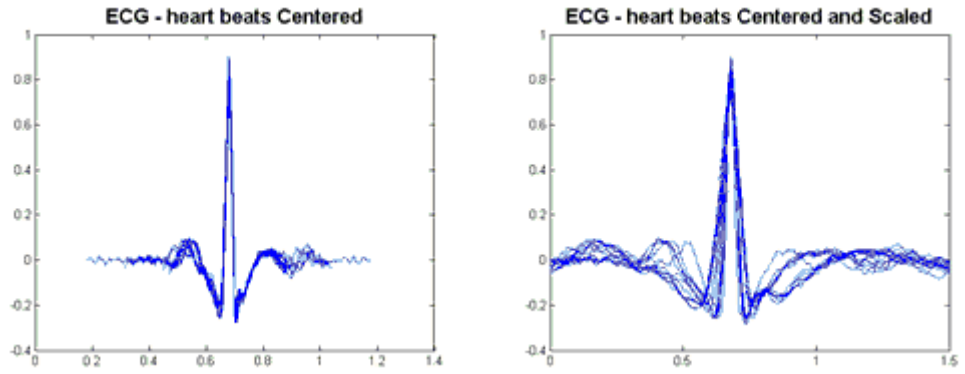

Figure 4.3: Heartbeats.
(a) On the left, segmented. (b) On the right, segmented, rescaled and aligned

All the heartbeats are going to be segmented and visualized in a typical profile monitor view with no particular signs of misalignment.
Obviously, some distortion have been introduced warping the profiles in order to obtain more points to describe each segment.
For this reason, in order not to lose information about this scaling phase, the two warping coefficients will be stored and used to describe each profile together with model coefficients, following the idea proposed by Grasso et al. in [14]. In this way, it should be possible to keep track of each level of distortion introduced in this phase.
In theory, severe distortions are linked to anomalies; in fact, a profile with a particularly smooth segment is caused by an RR interval shorter than the average, that forced the scaling phase to strongly warp it in order to obtain a segment of adequate length.
Hence, average length segments will be warped in the same way and, all the shorter or longer segments will show a similar level of distortion.
Therefore, the original heartbeats behaviour will not be influenced and compromised by this scaling phase.

0) Input data
After the pre-processing phase, data are in the form of aligned profiles and, for this reason, can be used as input for the algorithm.
Considering ECG signals, however, it is quite improbable to experience just one kind of anomaly like in the case presented in [20]. Moreover, not all the anomalies are necessarily in sequence into the dataset. For this reason, in Chapter 5 it will be shown the robustness of the algorithm in case of anomalies that occur occasionally throughout the dataset under analysis.

1) Model selection

Considering the fact that no profile monitoring techniques have ever been used to detect any anomaly between heartbeats, and that no well-known model are available to describe heartbeats, the same model used by Chen, Birch and Woodall to approximate profiles has been adopted also in this work.

In order to prove the adequacy of the model, residuals have been analysed to assure they do not contain any significant information not captured by the selected model.

First of all, considering the number of points that compose each profile, the number of nodes has been chosen maintaining a proportion of 10 nodes every 100 points. Then, the resulting positioning of all the equi-spaced nodes is presented in Fig. 4.4; the patient's heartbeats are represented in blue, and black vertical lines represent the position of each node.

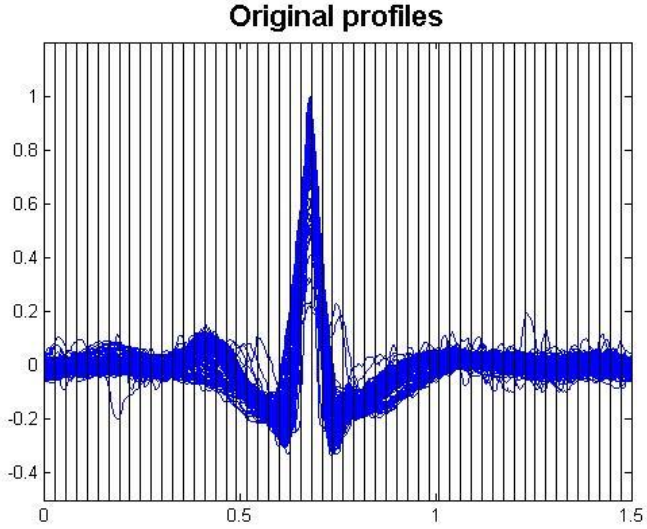

Figure 4.4: Equi-spaced node positioning

Using a first grade polynomial spline, the resulting approximation and the residuals of this model are presented in Fig. 4.5.

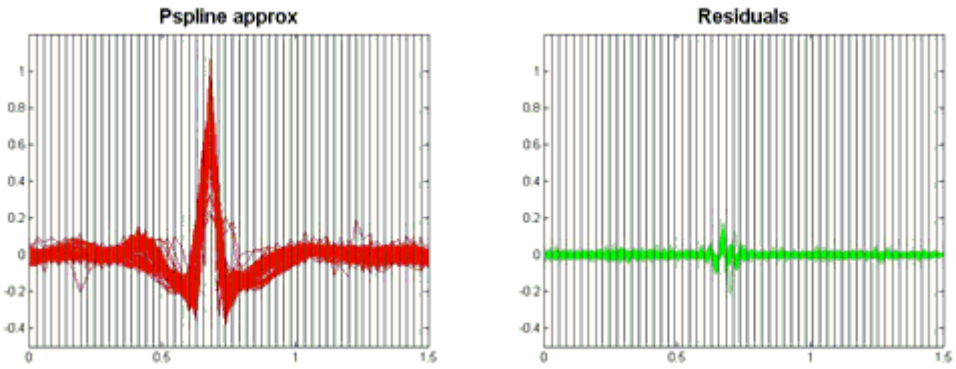

Figure 4.5: Approximation results.
(a) On the left, the approximated profiles. (b) On the right, model residuals.

Residuals show, in correspondence to the QRS complex, that some details have not been captured by the model; for this reason, an alternative way of positioning these nodes has been proposed.

Established that the QRS complex is the most significant part of an heartbeat, instead of ignoring this information, a higher number of nodes will bw positioned in this zone; the remaining ones will be distributed equi-spatially along the left over part of the domain.
The QRS complex position, that has been evaluated on the mean dataset profile identifying the maximum peak, R, and the two local minima before and after it (P and Q).
This nodes re-positioning is represented in Fig. 4.6:

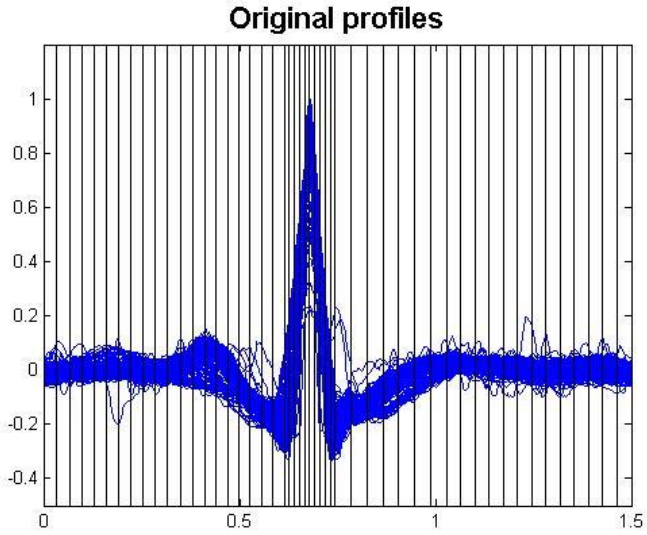

Figure 4.6: Alternative node positioning

The residuals obtained subtracting the approximation obtained from the original profiles shows that the model is able to capture all the significant content of these profiles and, for this reason, it can be considered an appropriate model (Fig. 4.7).

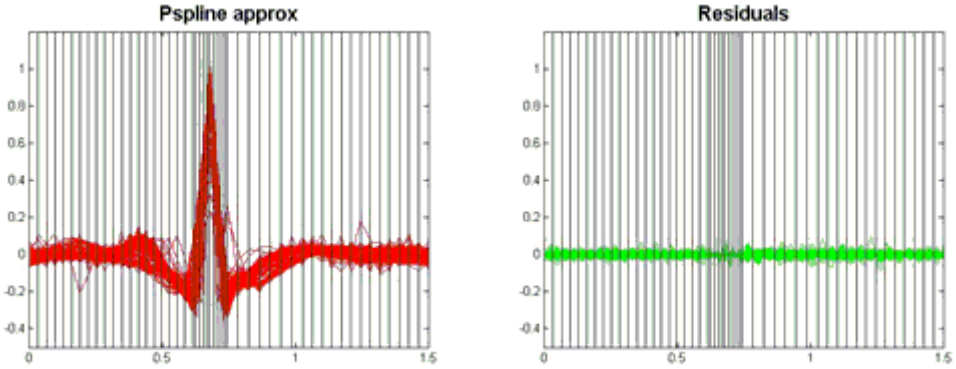

Figure 4.7: Approximation results.
(a) On the left, the approximated profiles. (b) On the right, model residuals.

Moreover, in order to exclude any loss of information derived from a bad approximation model, another control chart, the Q chart, will be added to monitor the model residuals.
An easy scheme of the Q control chart can be seen in the following picture.

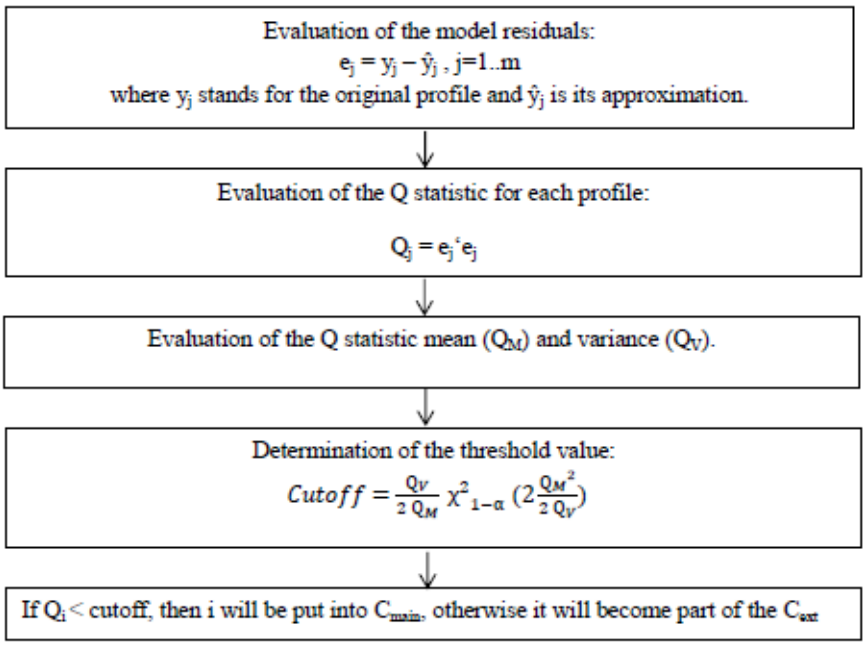

Figure 4.8: Q chart scheme.

Furthermore, not only model coefficients will be used to perform the analysis.
In fact, due to the scaling phase, also scaling coefficients should be monitored in order to exclude any loss of information.
This will be done in two different ways:

- in scenario A, model coefficients and scaling factors are considered together and analysed in one T^2 control chart. Residuals are monitored with a Q control chart;
- in scenario B, model coefficients and scaling factors are considered separately and analysed with two different T^2 control chart. Residuals are monitored with a Q control chart.

The global false alarm rate α is set at 5% for each of the two proposed scenarios. Hence, when working with more than one chart, this value should be split in order to maintain a global false alarm rate equal to the desired target, accordingly to the Bonferroni inequality.

Both T^2 and Q chart will be used with theoretical thresholds.
In particular, for what the T^2 concerns, this rely on the underlying hypothesis of normality of the monitored coefficients. If this hypothesis is not verified, results can be misleading.
For this reason, an alternative chart which does not require the normality assumption will be tested too in both scenarios. This alternative control chart is the K chart[3], which is based on a variant of the Support Vector Machine paradigm; the inputs of this chart are the same of the T^2 chart ones.

2) Cluster analysis
The cluster analysis will be performed in the same way as described in Chapter 3.3; the only difference will be in the coefficient matrix *Phi*, accordingly to what has been explained in the previous step of the algorithm. Moreover, in Chapter 5 it will be shown the robustness of this method to the variation of some choices that need to be made to perform this analysis.

3) Estimation of the mean profile
Once the main cluster has been individuated, the mean profile (y_{pa}) and its coefficients (Phi_{pa}) can be estimated. When scaling coefficients are considered, also mean scaling factors should be computed.
4) Control chart iterations
In each iteration, like the original method, all the C_{ext} profiles which have a T^2 value lower than the cutoff, will be added to a temporary cluster $C_{new.}$ After

[3] See Appendix and [55] for further information about this chart.

analysing them all, those profiles put into C_{new} will be added to C_{main} and the mean profile and its coefficients will be updated.
This operation is repeated until no more profiles can be added to the main cluster.

4.1) Test of all the profiles
In the original method, those elements positioned into the main cluster by the hierarchical clustering phase are never tested with the control chart.
As will be shown in Chapter 5, sometimes just one method is able to capture an anomaly, while the other fails to recognize it.
Moreover, clustering is generally able to recognize strong deviations from a regular situation; hence, if an element has been positioned into the main cluster, this should not be far cry from the in control situation. The detection of few anomalies with a little deviation from the in control situation, on the other side, is exactly the case where control charts outperform clustering.
Hence, this phase adds just more precision to the whole method, considering that no strong deviations (that could have distorted the mean profile coefficients estimation, downgrading the whole system performances) are expected in these profiles.

5) End
In the end, all the profiles belonging to the last C_{ext} cluster represent out of control (OOC) elements that should be excluded from phase II parameters estimation.

Considering the two scenarios presented before, the results of each control chart system should be obtained putting together the results obtained by every single chart of the system under analysis so that:
- a profile will be considered abnormal if it has been identified in that way at least by one chart;
- a profile will be considered referred to an in control situation if it has been considered in that way by all charts.

The method performances will be measured using these overall results.

Chapter 5

Performance analysis

5.1. Application on simulated profiles

A preliminary work has been done on simulated profiles in order to verify the sensitivity of the monitoring performances on some design choices.
The simulated profiles which have been considered as in control elements can be seen in Fig. 5.1 and are described by:

$$y_j(t) = \sum_{i=1}^{5} \beta_{i,j} e^{\gamma_{i,j}(t+\omega_{i,j})} + \varepsilon_j(t) \text{ , } 0 < t < 1 \qquad j = 1..m \tag{5.1}$$

Where:
- $\beta = N(\mu_\beta; \boldsymbol{\Sigma}_\beta^2)$, μ_β=[0.88,-0.5, 0.6,-0.6,-0.5], $\boldsymbol{\Sigma}_\beta^2 = diag[(8.8,5,6,6,5)e^{-2}]$;
- $\gamma = N(\mu_\gamma; \boldsymbol{\Sigma}_\gamma^2)$, μ_γ=[-20,-50, -100,-150,-200], $\boldsymbol{\Sigma}_\gamma^2 = diag[(5.5,8.5,11,9,15)e^{-1}]$;
- ω = [-0.5,-0.45,-0.3,0.7,-0.45];
- $\varepsilon = N(0,1)$.

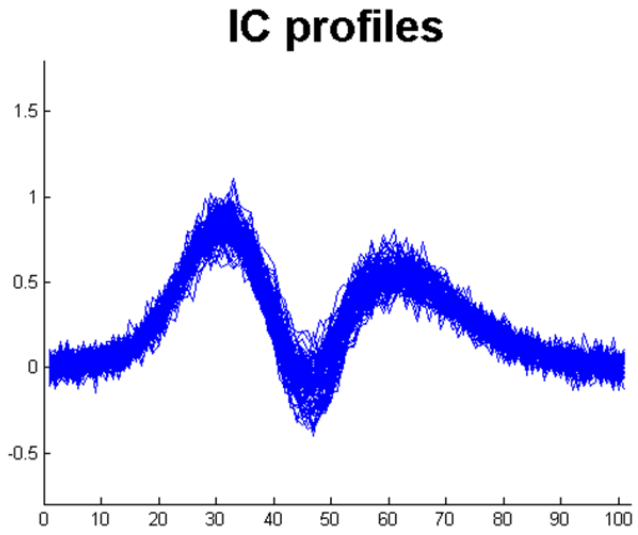

Figure 5.1: Simulated IC profiles

Then, different kinds of out of control profiles have been simulated too; they all start from the common IC equation and some variations have been introduced in

order to mimic different situations that can occur. Every OOC has been tested with different level of severity, to cover a reasonable range of distortion.
The first kind of abnormal profile, which from now onwards will be called OOC I, has been obtained introducing a shape variation all over the domain.
In particular, each β coefficient will vary accordingly to its standard deviation (Δ=[1,2,3,4]), as can be seen in:

$$y_j(t) = \sum_{i=1}^{5} (\beta_{i,j} + \Delta\sigma_i) e^{\gamma_{i,j}(t+\omega_{i,j})} + \varepsilon_j(t) \tag{5.2}$$

The resulting profiles are represented in red in Fig. 5.2.

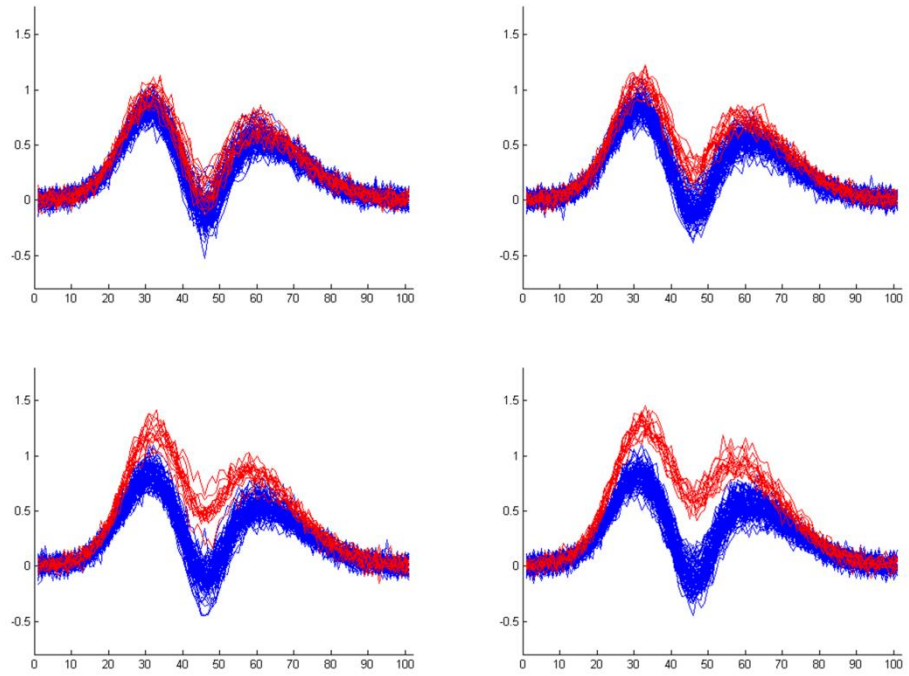

Figure 5.2: Simulated dataset. In blue, IC profiles; in red OOC I profiles from left to right, 4 different level of shift [Δ=1,2,3,4]

The second kind of abnormal profile, which from now onwards will be called OOC II, has been obtained introducing an external source of error to the original signal (Δ=[0.1,0.15,0.2,0.25]), as can be seen in:

$$y_j(t) = \sum_{i=1}^{5} \beta_{i,j} e^{\gamma_{i,j}(t+\omega_{i,j})} + \Delta \sin(2\pi t) + \varepsilon_j(t) \tag{5.3}$$

The resulting profiles are represented in red in Fig. 5.3.

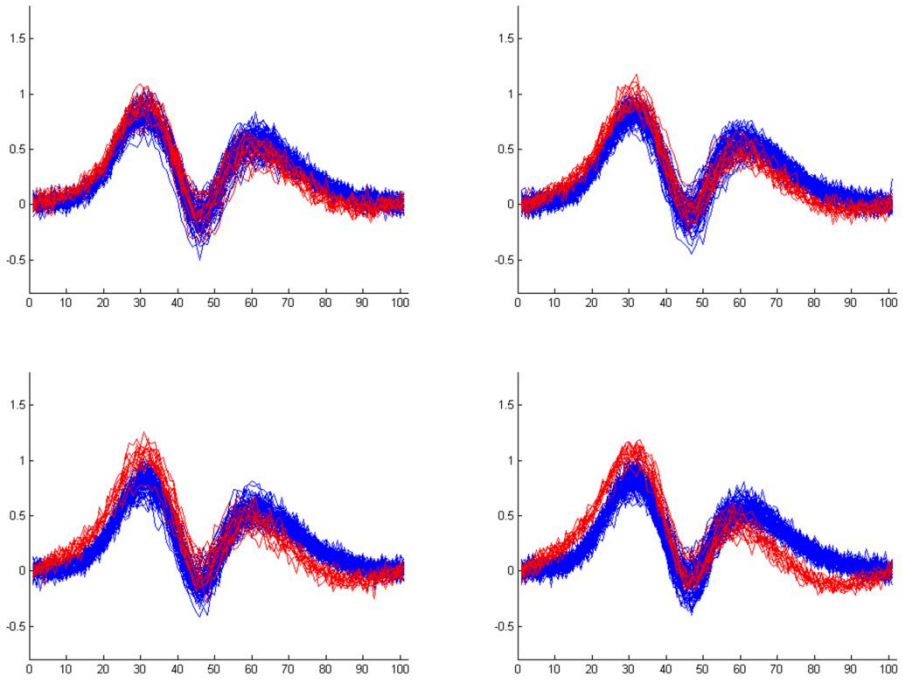

Figure 5.3: Simulated dataset. In blue, IC profiles; in red OOC II profiles from left to right, 4 different level of shift [Δ=0.1,0.15,0.2,0.25]

Finally, the third kind of abnormal profile, which from now onwards will be called OOC III, has been obtained introducing a shape variation on a limited part of the domain. In particular, this deviation has been obtained varying just one of the β parameters proportionally to its standard deviation (Δ=[0.25,0.35,0.45,0.55]), as can be seen in :

$$y_j(t) = \sum_{i=1}^{4} \beta_{i,j} e^{\gamma_{i,j}(t+\omega_{i,j})} + \left(\beta_{5,j} + \Delta\sigma_i\right) e^{\gamma_{i,j}(t+\omega_{i,j})} + \varepsilon_j(t) \tag{5.4}$$

The resulting profiles are represented in red in Fig. 5.4.

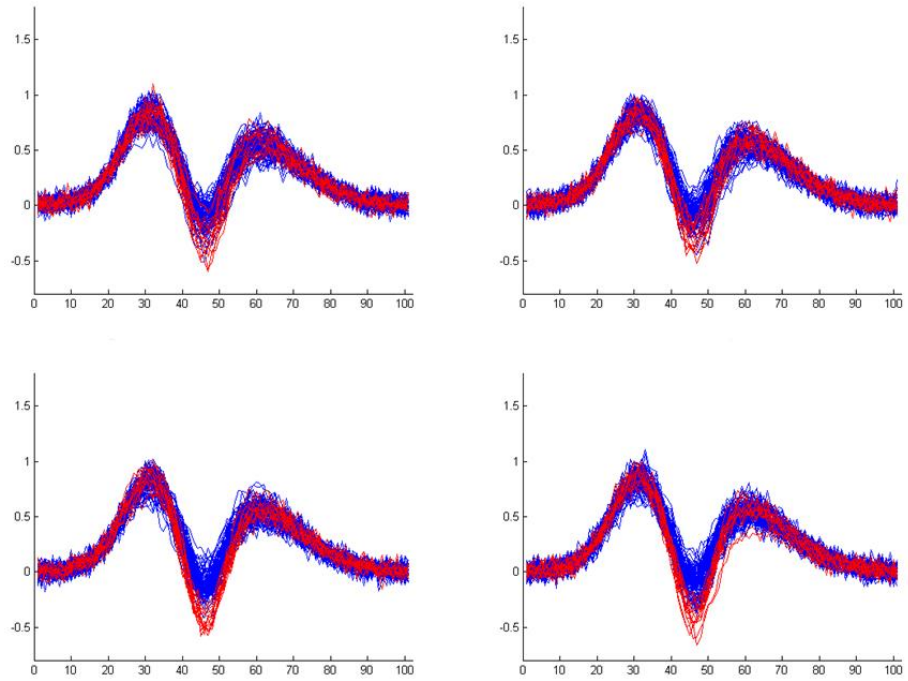

Figure 5.4: Simulated dataset. In blue, IC profiles; in red OOC III profiles from left to right, 4 different level of shift [Δ=0.25,0.35,0.45,0.55]

In order to evaluate the cluster-based method's performances, a set of indicators is needed. In particular, two indexes that monitor the I type error and the II type error will be used.
The I type error will be controlled through the FP rate, which is the percentage of IC profiles classified as OOC ones; in order to evaluate the different methods "coeteris paribus", this performance has been set to a desired value, 5%, through an empirical cutoff.
The II type error will be controlled through the FN rate, which is the percentage of OOC profiles which are classified as IC ones. Having an high value of this index means that the method does not correctly discriminate between IC and OOC profiles, with resulting bad results in the decontamination task.

Each dataset is composed by 100 profiles and the percentage of contaminating profiles is 20%.

Moreover, in order to compute some confidence intervals [58] that can help in the evaluation of these performance indicators, analysis will be replicated on each dataset for 100 times.
The baseline method, as proposed in [20], consists of using:

- 1st degree polynomial splines to fit the profiles;
- Mahalanobis distance to evaluate similarities between profiles;
- complete linkage function to perform the cluster analysis.

The sensitivity of the baseline method with respect to the design choices regarding the spline model, the distance metric and the linkage function has been investigated by means of a simulation study.
The following alternative choices have been considered:

- cubic B-spline [57] to fit the profiles;
- Chebychev distance to evaluate similarities between profiles (This way of evaluate distances focuses its attention only on the highest difference in modulus between coefficients);
- Ward linkage function [58] to perform the cluster analysis.

Setting the α value at the same level for both the alternatives, it is possible to compare results just in terms of FN rate[4].

In tables 5.1, 5.2 and 5.3 are shown the results for each of the three analysed scenarios; each column should be compared to the first one, which contains the results obtained performing the cluster-based algorithm with the original choices proposed by the authors in [20].

OOC I	Baseline method	Bspline	Chebychev	Ward linkage
Shift 1	**0,831 [0,811;0,852]**	0,829 [0,807;0,850]	0,839 [0,810;0,869]	0,827 [0,809;0,844]
Shift 2	**0,290 [0,255;0,325]**	0,290 [0,247;0,332]	0,294 [0,256;0,333]	0,288 [0,243;0,333]
Shift 3	**0,015 [0,009;0,021]**	0,014 [0,009;0,019]	0,015 [0,009;0,021]	0,015 [0,008;0,022]
Shift 4	**0 [0;0]**	0 [0;0]	0 [0;0]	0 [0;0]

Table 5.1: FN confidence intervals for OOC I

[4] FP rates have been evaluated too. All the intervals obtained for each alternative contain 0.05 and, for this reason, the desired α level is always respected. Hence, methods can vary just in terms of FN rate.

OOC II	Baseline method	Bspline	Chebychev	Ward linkage
Shift 1	**0,359 [0,305;0,414]**	0,358 [0,299;0,417]	0,360 [0,305;0,416]	0,340 [0,305;0,375]
Shift 2	**0,049 [0,042;0,056]**	0,047 [0,039;0,056]	0,049 [0,040;0,058]	0,047 [0,040;0,055]
Shift 3	**0,007 [0,002;0,013]**	0,007 [0,003;0,010]	0,007 [0,002;0,012]	0,007 [0,001;0,012]
Shift 4	**0 [0;0]**	0 [0;0]	0 [0;0]	0 [0;0]

Table 5.1: FN confidence intervals for OOC II

OOC III	Baseline method	Bspline	Chebychev	Ward linkage
Shift 1	**0,430 [0,381;0,479]**	0,428 [0,376;0,480]	0,440 [0,376;0,505]	0,427 [0,378;0,476]
Shift 2	**0,233 [0,195;0,272]**	0,217 [0,193;0,241]	0,238 [0,195;0,281]	0,216 [0,199;0,233]
Shift 3	**0,143 [0,125;0,160]**	0,144 [0,128;0,160]	0,143 [0,123;0,163]	0,139 [0,119;0,159]
Shift 4	**0,091 [0,076;0,107]**	0,092 [0,079;0,105]	0,090 [0,076;0,105]	0,090 [0,076;0,103]

Table 5.2: FN confidence intervals for OOC III

Considering that for each variation tested the intervals obtained are statistically indifferent to those obtained performing the algorithm with the original choices, as they overlap, it is possible to state that none of these variations produce a significant impact on the method's performances. Therefore, the method is robust to these implementation choice variations. However, while considering alternatives, it is important to take into consideration the fact that they should not be selected randomly, but should be always evaluated taking into consideration their ability of catching the kind of anomalies which are expected to contaminate the dataset.

5.2. Application on the case study

The modified version of the cluster-based method, presented in Chapter 4, will now be applied to ECG profiles.

As already explained in Chapter 4, a fundamental step necessary to obtain data in form of profiles is the segmentation of the whole ECG into heartbeats, followed by a scaling and alignment phase.
In order to take trace of the warping introduced in these phases, the applied coefficients will be stored and analysed too, as suggested in [14] by Grasso et al. A representation of the dataset in form of profiles can be seen in Fig. 5.5, while a representation of the scaling profiles can be seen in Fig. 5.6, where the x-axis represent the scaling coefficients applied in the segment that precedes the R peak and the y-axis represent the scaling coefficients applied in the segment that follows it.

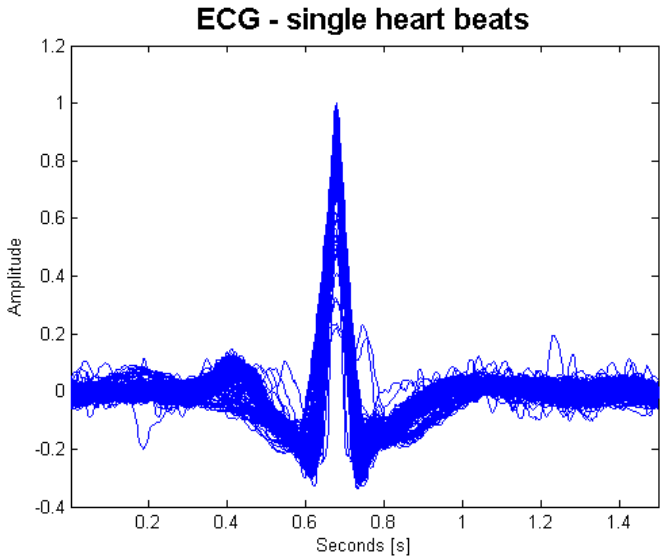

Figure 5.5: Profile monitoring view of the segmented, scaled and aligned ECG dataset.

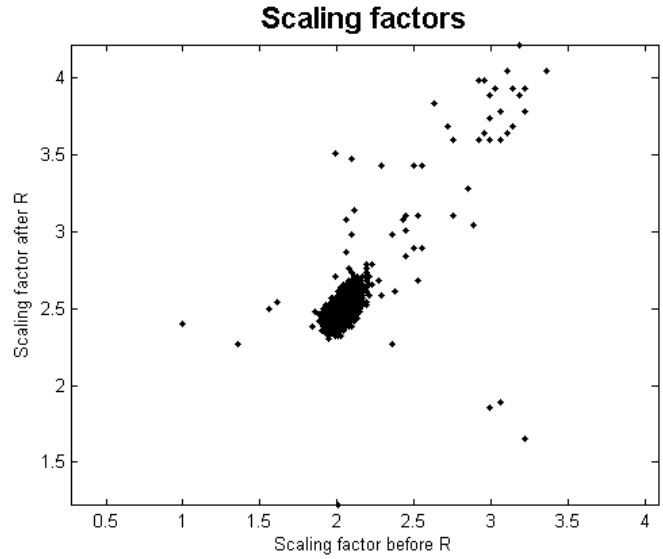

Figure 5.6: Scaling coefficient used in the scaling-alignment phase.

Remind that there are two scaling factors, i.e. (i) a scaling factors to warp the profile portion that precedes the R peak and (ii) a scaling factors to warp the profile portion that follows the R peak and. Generally speaking, these two scaling factors are not equal.
It is interesting to notice that the majority of profiles has been scaled using very similar scaling factors, while others have one or both the scaling factors higher than usual.

Considering the knowledge about clinical OOC profiles, in Fig. 5.7 it is possible to notice that not all the "strange" scaling factors belong to conditions of clinical anomalies. Hence, also other clinically regular profiles have been scaled in a peculiar way.

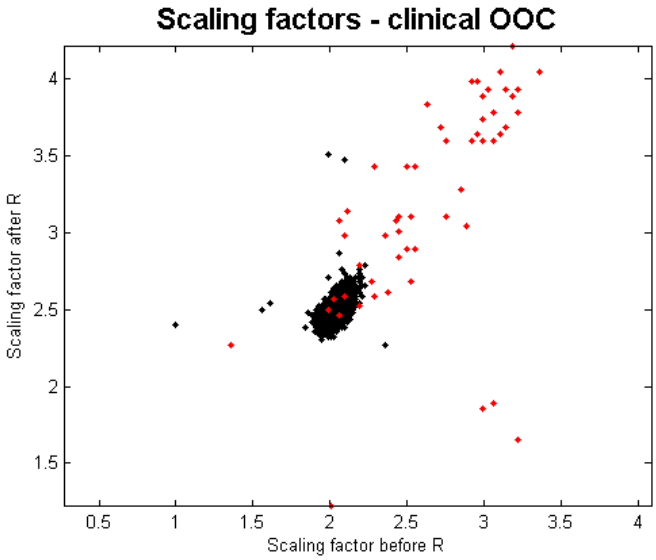

Figure 5.7: Scaling coefficient used in the scaling-alignment phase. In red, clinical OOC.

Looking at the shape of heartbeats, in Fig. 5.8, clinical abnormal profiles (represented in red) appears different from the majority of all the other profiles, clinically considered regular (in blue). However, also some of these regular profiles show some strange behaviours.
As already explained in Chapter 2, in fact, there exist also some profiles that are particularly different from the in-control pattern (i.e. the "signature profile") but they are not labelled as "clinical anomalies" within the database. Due to their anomalous pattern, they should be considered as contaminant profiles too.

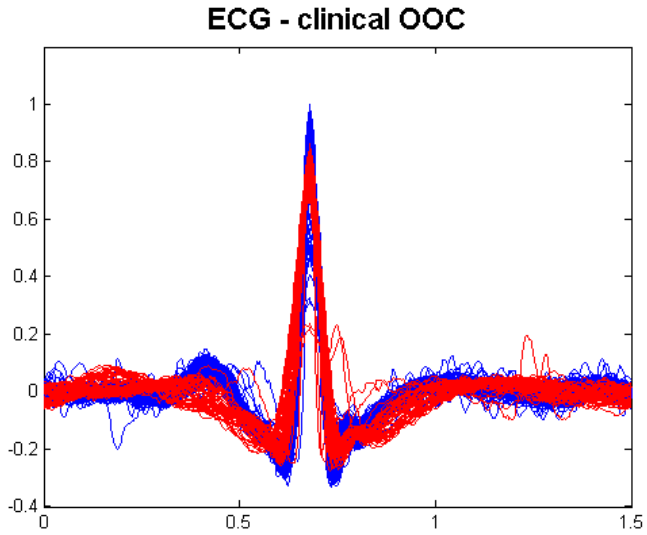

Figure 5.86: Profile monitoring view of the segmented, scaled and aligned ECG dataset. In red, clinical OOC profiles.

After the pre-processing phase, a dataset composed by 2750 profiles has been obtained. With the information available from the ECG annotation file, it has been possible to identify 54 profiles containing some clinical anomalies. However, also 45 profiles not labelled as "clinical anomalies" can be considered contaminants too, because of their departure from the "signature profile". These information will be used in the performance measurement phase, in order to evaluate whether the decontamination task has been capable of recognizing these abnormal profiles or not.

In fact, performances will be measured in terms of:

- number of profiles classified as IC;
- number of profiles classified as OOC;
- number of clinical anomalies considered IC by the algorithm;
- clinical FP (false positive) and FN (false negative) rates, obtained considering as OOC profiles just clinical contaminants (54 profiles);
- global FP (false positive) and FN (false negative) rates, obtained considering as OOC profiles both clinical and non-clinical contaminants (54 + 45 profiles);

After this introduction, it is possible to analyse the two selected scenarios in depth. Two different control charting schemes are considered, one based on the Hotelling's T^2 chart (which assumes data to be normally distributed), and one based on the K-chart (see the Appendix section), which is applicable regardless of the actual data distribution. Two implementation scenarios are considered

too: (i) one based on merging the spline coefficients and the scaling factors into a single vector and monitoring them by means of a single control chart, and (ii) one based on using two separate control charts (of the same type) for monitoring the spline coefficients and the scaling factors, respectively.

Scenario A: model coefficients and scaling coefficients monitored with the same control chart

	Classified as OOC	Classified as IC	Clinical anomalies non detected by the algorithm	Clinical FP	Clinical FN	Global FP	Global FN
T^2+Q	243	2507	0	0,07	0	0,05	0
K+Q	114	2636	24	0,03	0,44	0,03	0,55

Table 5.3 : Scenario A performances

Table 5.4 shows that all the profiles clinically described as abnormal are identified using the T^2 control charting system, while about the half of them is not recognized using the K control charting one, despite their abnormal behaviour can be seen both in terms of shape and in terms of scaling coefficients, as Fig. 5.9 shows.

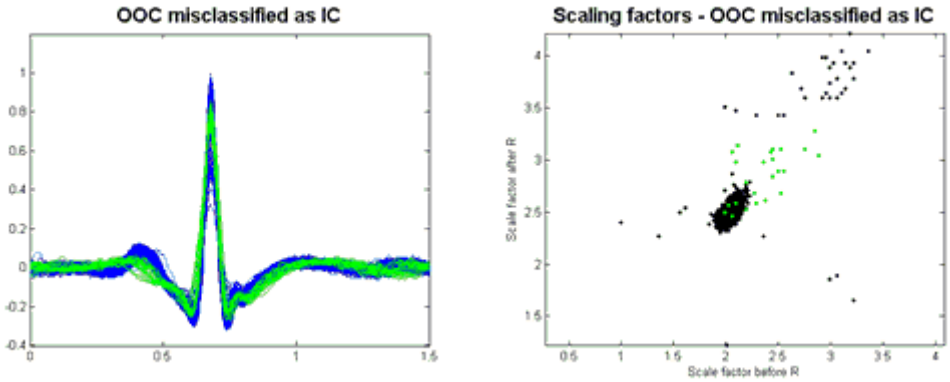

Figure 5.9: Clinical OOC elements (in green) non recognized by the K chart control system. (a) On the left, their shape. (b) On the right, their scaling coefficients.

As said before, in the dataset there are also non-clinical contaminants that should be considered abnormal profiles too due to their difference from the "signature profile"; for this reason, also a global behaviour of these two charting systems should be evaluated.
For what concern the T^2 one, also all those non-clinical contaminants can be identified successfully.

Two examples of these profiles are represented in Fig. 5.10, where all the profiles identified as IC are represented in blue and one of those non-clinical contaminant is represented in red.

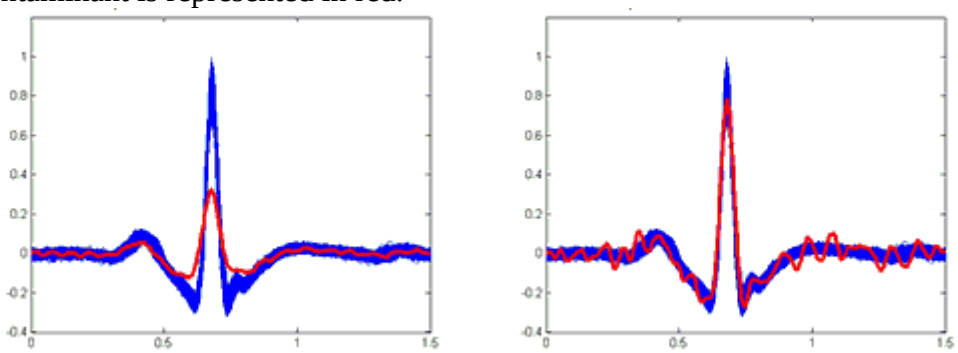

Figure 5.10: Non clinical contaminants correctly identified as OOC.

For what concern the K control chart system, on the other hand, just few of those non-clinical anomalies, almost 15, are identified as OOC.

Hence, what can be said is that K chart and Q chart together are not able to complete the discrimination task on ECG profiles in this scenario, while T^2 chart and Q chart together are able to do so.
Moreover, the global false positive rate obtained by this control charting system is equal to the value set to perform the analysis (α=5%), confirming the adequacy of theoretical control limits based on normality assumptions.

Scenario B: model coefficients and scaling coefficients monitored with two separated control charts

	Classified OOC	Classified IC	Clinical anomalies non detected by the algorithm	Clinical FP	Clinical FN	Global FP	Global FN
T^2+Q	223	2527	0	0,06	0	0,05	0
K+Q	614	2136	4	0,21	0,07	0,21	0,29

Table 5.4: Scenario B performances.

All the profiles clinically described as abnormal are identified using the double T^2 control charts system, while few of them are not recognized using the double K control charts one, as can be seen in Table 5.5.

Considering also non-clinical anomalies, the T^2 control charts system is able of recognizing them all; two examples of these correctly identified non-clinical contaminants are represented in Fig. 5.11.

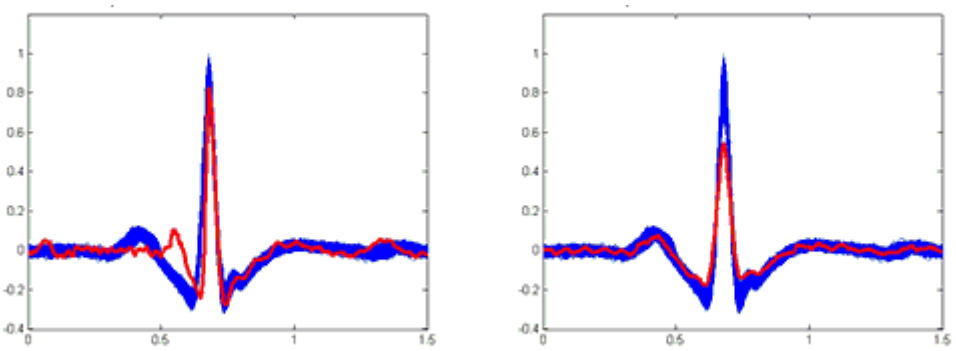

Figure 5.71: Non clinical abnormal profiles which show a strange behaviour.

For what concern the K control charting system, on the other hand, just few of those non-clinical abnormal profiles, almost 20, have been correctly identified.
However, many regular profiles have been indicated as OOC, despite their shape does not show any particular deviation from the signature profile and their scaling factors are very similar to the average ones, as can be seen in Fig. 5.12.

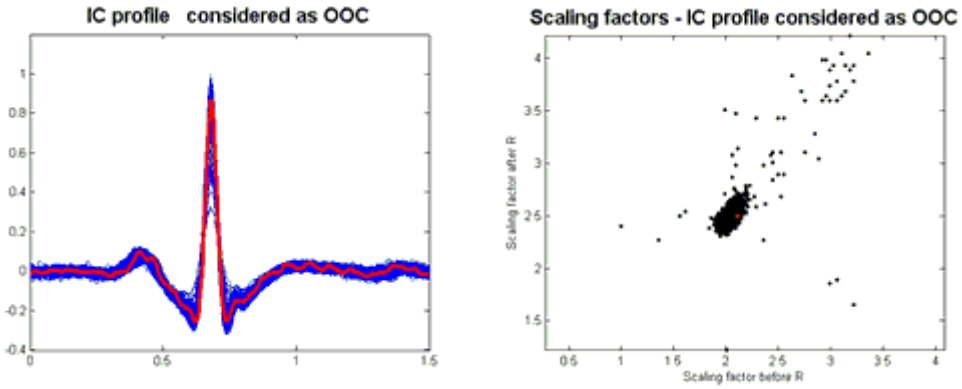

Figure 5.12: IC profile (in red) classified as OOC by the K charts control system. (a) On the left, its shape. (b) On the right, its scaling coefficients.

Therefore, also in this case K charts and Q chart together are not able to complete the discrimination task on ECG profiles, while T^2 charts and Q chart together are able to do so.
Again, the global false positive rate of the T^2 control charting system is equal to the value set to perform the analysis (α=5%), confirming the adequacy of theoretical control limits based on normality assumptions.

Before concluding this section, it is important to make a comment about the residual control chart too.
It has been already said in Chapter 4 that, if the model used to approximate the profile is apt, residuals should be Normal and independent.
Hence, the theoretical cutoff used to determine the Q chart threshold, which is based on this assumption, should be similar to the empirical cutoff used to the determine the empirical Q chart threshold, which is determined using the real data distribution; moreover, the two chart should lead to the same conclusions.

This is exactly what happens in these cases; hence, another confirmation of the model adequacy has been obtained.
Cutoff values can be seen in table 5.6 and, as example, the two Q charts for scenario B can be seen in Fig. 5.13.

	Theoretical	Empirical
Scenario A (α_Q=0,5 α_{FAM})	0.053	0.045
Scenario B (α_Q=0,33 α_{FAM})	0.056	0.050

Table 5.5: Q-chart thresholds.

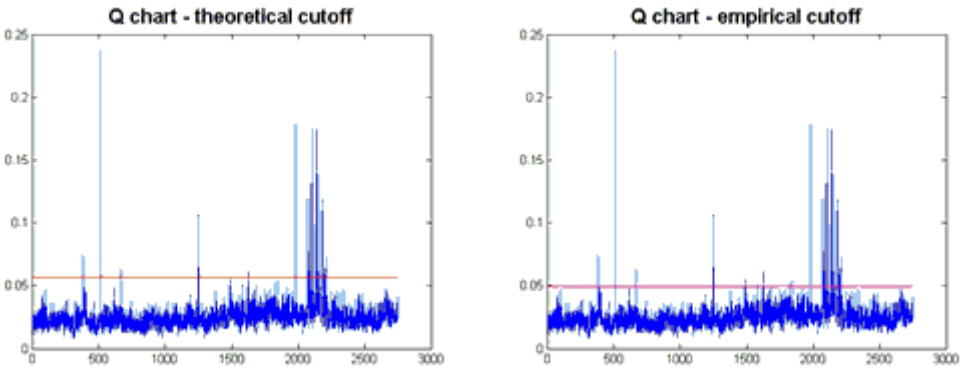

Figure 5.83: Q charts for scenario B (α_Q=0,33 α_{FAM}).
(a) On the left, using the theoretical cutoff. (b) On the right, using the empirical cutoff.

Chapter 6

Conclusions and future developments

Mainstream methods for ECG monitoring involve the identification of salient features of interest and the computation of synthetic indexes that are compared with predefined thresholds. Those thresholds rely on gender and age group of the patients, but they are applied regardless of the specific pattern of each single patient. This yields non optimal monitoring results in most cases, in terms of high false alarm rates (Type I error) and missed detection of actual anomalies (Type II error). This thesis proposes a novel idea for ECG signal monitoring. It consists of determining the "signature" of the single individual under analysis, allowing a custom monitoring system that search for deviations from a specific "in control" condition instead of searching for deviations from standard conditions. The study is focused on the "design phase" of control charts, i.e., the phase during which in-control signal should be collected in order to estimate the control limits to be applied for future signal monitoring. In this phase, it is very important to understand if the patient conditions were actually in-control during the data collection period. Possible anomalous observations (hereafter denoted by "dataset contaminations") must be detected and removed, because they may have a detrimental impact on the estimation of control limits.

In order to deal with a contamination problem that can deeply affect phase I parameter estimations, the cluster-based profile monitoring approach proposed by Chen, Birch and Woodal in [20] has been used and improved. This method has been studied in order to obtain a deep knowledge of its functioning. It has been shown that some implementation choices can vary, accordingly to the user necessity, without any statistically significant impact of the performances (under the assumption that the choices are coherent with the data under analysis). Then, it has been demonstrated that not only consecutive anomalies can be detected, but also few occasional ones.
Moreover, its strength has been explained through the synergic use of two traditional methods that alone are proven to be not capable of completing the decontamination task.

After this study, a real case has been analysed with an innovative approach. In fact, instead of using traditional methods to examine an ECG, a profile monitoring techniques based on a modified cluster-based approach has been used.

First of all, a preliminary phase to filter, identify, segment and align heartbeats has been done, in order to prepare data to be used as profiles. After using a traditional method to detect heartbeats, a particular way of segmenting and aligning them has been proposed.

Then, the cluster-based algorithm has been performed, adding also a control chart to monitor model residuals to avoid any information loss due to the fact of using a model-based approach to describe profiles. Not only model coefficients have been used to perform the analysis, but also those scaling factors used to obtain profiles of equal length and centered on their R-peak are considered. In this way, it has been possible to take trace of any kind of distortion introduced by this warping phase.

Having two classes of coefficients, two different scenarios have been tested in order to evaluate if considering all these coefficients together is better or worse that considering them separately.

A comparison between a parametric control chart, i.e., the Hotelling's T^2 control chart, and a nonparametric one, i.e., the so-called K-chart, was made. The two charts represents alternative charting schemes to implement the cluster-based approach for Phase I decontamination.
In both cases, residual are monitored with a Q control chart; however, in the first case a theoretical threshold is used, while in the second one the choice went on an empirical cutoff. However, if the model used is correct, no differences should be expected in these two different way of constructing the Q control chart because residuals should be Gaussian and identically distributed, which is the underlying hypothesis used to build the theoretical cutoff.

Both these monitoring system have drawbacks; the first one could bring wrong results in case of strong deviations from the assumptions, while the second one could bring wrong results if the data used to determine the empirical distribution does not really represent the real distribution of an in control situation.

The results presented in Chapter 5.2 show that using a T^2 control chart system allows the identification of abnormalities in both the scenarios proposed, confirming the ability of this cluster-based methodology in excluding contaminated profiles from phase II parameter estimation.
Considering also that the global false positive rate obtained is coherent to the one set to build the charts, it can be concluded that the underlying hypothesis about the coefficients distribution is not violated.
What is more, the fact of monitoring model coefficients and scaling coefficients together or separately does not particularly affect the results obtained.

On the other hand, the K chart is less suitable to the task of decontamination.
In fact, in scenario A, the global false positive rate is lower than what was set and false negative rate is particularly high; hence, considering that abnormalities are generally linked to higher statistic values, this fact means that the control limit has been significantly inflated by those abnormalities, with the results that the chart is not able to complete the task for what has been selected.
In scenario B, instead, false positive rate is higher than what was set and false negative rate is lower than before; hence, in this case, abnormalities do not have the same influence as before, and the resulting control limit is lower, allowing the chart to complete the task for what has been selected also if not in an efficient way.

Therefore, using empirical control limits while trying to decontaminate a phase I dataset is not suitable for these kind of situation.

Future researches can be made to develop a way to identify control limits strictly related to each patient heartbeats, in order to prevent situations where hypothesis violations can affect phase I results, with severe implications for the following analysis.
Another area of interest can be related to merging together relevant information obtained from different sources, which can be both sensors (and in this case sensor fusion has to be performed) or medical indications, in order to obtain a more reliable approach to identify various kind of potentially dangerous situations.

References

1. POLLACK, Martha E. Intelligent technology for an aging population: The use of AI to assist elders with cognitive impairment. *AI magazine*, 2005, 26.2: 9.

2. HAREVEN, Tamara K. Historical perspectives on aging and family relations. *Handbook of aging and the social sciences*, 2001, 5: 141-159.

3. MYNATT, Elizabeth D.; ROGERS, Wendy A. Developing technology to support the functional independence of older adults. *Ageing International*, 2001, 27.1: 24-41.

4. ABOWD, Gregory D., et al. The aware home: A living laboratory for technologies for successful aging. In: *Proceedings of the AAAI-02 Workshop "Automation as Caregiver*. 2002. p. 1-7.

5. WOOD, Anthony, et al. Context-aware wireless sensor networks for assisted living and residential monitoring. *Network, IEEE*, 2008, 22.4: 26-33.

6. ALEMDAR, Hande; ERSOY, Cem. Wireless sensor networks for healthcare: A survey. *Computer Networks*, 2010, 54.15: 2688-2710.

7. CLIFFORD, Gari D., et al. *Advanced methods and tools for ECG data analysis*. London: Artech house, 2006

8. WOODALL, William H. Current research on profile monitoring. *Production*, 2007, 17.3: 420-425.

9. WILLIAMS, James D.; WOODALL, William H.; BIRCH, Jeffrey B. Statistical monitoring of nonlinear product and process quality profiles. *Quality and Reliability Engineering International*, 2007, 23.8: 925-941.

10. PAN, Jiapu; TOMPKINS, Willis J. A real-time QRS detection algorithm. *Biomedical Engineering, IEEE Transactions on*, 1985, 3: 230-236.

11. SAYADI, O.; SHAMSOLLAHI, M. B. A model-based Bayesian framework for ECG beat segmentation. *Physiological Measurement*, 2009, 30.3: 335.

12. MADEIRO, Joao PV, et al. A new approach to QRS segmentation based on wavelet bases and adaptive threshold technique. *Medical engineering & physics*, 2007, 29.1: 26-37.

13. RAMSAY, J. O.; LI, Xiaochun. Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1998, 60.2: 351-363.

14. GRASSO, Marco, et al. Functional Data Analysis and Classification for Profile Monitoring and Fault Diagnosis in Waterjet Machining Processes. 2013.

15. VULLINGS, H. J. L. M.; VERHAEGEN, M. H. G.; VERBRUGGEN, H. Automated ECG segmentation with dynamic time warping. In: *Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE*. IEEE, 1998. p. 163-166.

16. BERNDT, Donald J.; CLIFFORD, James. Using Dynamic Time Warping to Find Patterns in Time Series. In: *KDD workshop*. 1994. p. 359-370.

17. ZIFAN, Ali, et al. Automated ECG segmentation using piecewise derivative dynamic time warping. *International Journal of Biological and Medical Sciences*, 2006, 1.3.

18. KEOGH, Eamonn, et al. Locally adaptive dimensionality reduction for indexing large time series databases. *ACM SIGMOD Record*, 2001, 30.2: 151-162.

19. SHORTEN, Gavin P.; BURKE, Martin J. Pre-Processing for Value Based Dynamic Time Warping of the ECG Signal. In: *Proc. 22nd. Irish Sig. & Sys. Conf*. 2011. p. 201-206.

20. CHEN, Yajuan; BIRCH, Jeffrey B.; WOODALL, William H. A Phase I Cluster-Based Method for Analysing Nonparametric Profiles. *Quality and Reliability Engineering International*, 2014.

21. ESTER, Martin, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Kdd*. 1996. p. 226-231.

22. WANG, Wei, et al. STING: A statistical information grid approach to spatial data mining. In: *VLDB*. 1997. p. 186-195.

23. SHEIKHOLESLAMI, Gholamhosein; CHATTERJEE, Surojit; ZHANG, Aidong. Wavecluster: A multi-resolution clustering approach for very large spatial databases. In: *VLDB*. 1998. p. 428-439.

24. AGRAWAL, Rakesh, et al. Automatic subspace clustering of high dimensional data. *Data Mining and Knowledge Discovery*, 2005, 11.1: 5-33.

25. DEMPSTER, Arthur P.; LAIRD, Nan M.; RUBIN, Donald B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1977, 1-38.

26. SUGAR, Catherine A.; JAMES, Gareth M. Finding the number of cluster in a dataset. Journal of the American Statistical Association, 2003, 98.463

27. TIBSHIRANI, Robert; WALTHER, Guenther; HASTIE, Trevor. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2001, 63.2: 411-423.

28. PEDERSEN, Ted; KULKARNI, Anagha. Automatic cluster stopping with criterion functions and the gap statistic. In: Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations. Association for Computational Linguistics, 2006. p. 276-279.

29. MUHR, Markus; GRANITZER, Michael. Automatic cluster number selection using a split and merge k-means approach. In: Database and Expert Systems Application, 2009. DEXA'09. 20th International Workshop on. IEEE, 2009. p. 363-367.

30. PELLEG, Dan, et al. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In: ICML. 2000. p. 727-734.

31. FERRARETTI, Denis; GAMBERONI, Giacomo; LAMMA, Evelina. Automatic cluster selection using index driven search strategy. In: AI* IA 2009: Emergent Perspectives in Artificial Intelligence. Springer Berlin Heidelberg, 2009. p. 172-181.

32. SANDER, Jörg, et al. Automatic extraction of clusters from hierarchical clustering representations. In: Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2003. p. 75-87.

33. GUYON, Isabelle; ELISSEEFF, André. An introduction to variable and feature selection. The Journal of Machine Learning Research, 2003, 3: 1157-1182.

34. MURPHEY, Yi Lu; GUO, Hong. Automatic feature selection-a hybrid statistical approach. In: Pattern Recognition, 2000. Proceedings. 15th International Conference on. IEEE, 2000. p. 382-385.

35. MITRA, Pabitra; MURTHY, C. A.; PAL, Sankar K.. . Unsupervised feature selection using feature similarity. IEEE transactions on pattern analysis and machine intelligence, 2002, 24.3: 301-312.

36. YU, Lei; LIU, Huan. Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research, 2004, 5: 1205-1224.

37. CAI, Deng; ZHANG, Chiyuan; HE, Xiaofei. Unsupervised feature selection for multi-cluster data. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010. p. 333-342.

38. DY, Jennifer G.; BRODLEY, Carla E. Feature selection for unsupervised learning. The Journal of Machine Learning Research, 2004, 5: 845-889.

39. ZHOU, Shiyu; JIN, Jionghua. Automatic feature selection for unsupervised clustering of cycle-based signals in manufacturing processes. IIE Transactions, 2005, 37.6: 569-584.

40. HAWKINS, Douglas M.; FATTI, L. Paul. Exploring multivariate data using the minor principal components. The Statistician, 1984, 325-338.

41. CHEN, Tianping; AMARI, Shun Ichi; LIN, Qin. A unified algorithm for principal and minor components extraction. Neural Networks, 1998, 11.3: 385-390.

42. HYVÄRINEN, Aapo. Survey on independent component analysis. Neural computing surveys, 1999, 2.4: 94-128.

43. HYVÄRINEN, Aapo; OJA, Erkki. Independent component analysis: algorithms and applications. Neural networks, 2000, 13.4: 411-430.

44. RANI, Sangeeta; SIKKA, Geeta. Recent Techniques of Clustering of Time Series Data: A Survey. International Journal of Computer Applications, 2012, 52.

45. GISBRECHT, Andrej. Time series clustering. ICOLE 2007, Lessach, Austria, 48.

46. GEURTS, Pierre. Pattern extraction for time series classification. In: Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg, 2001. p. 115-127.

47. KÖHLER, Andreas, et al. Unsupervised feature selection for pattern search in seismic time series. In: Journal of Machine Learning Research. In: Workshop and Conference Proceedings: New Challenges for Feature Selection in Data Mining and Knowledge Discovery. 2008. p. 106-121.

48. XIONG, Yimin; YEUNG, Dit-Yan. Mixtures of ARMA models for model-based time series clustering. In: Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on. IEEE, 2002. p. 717-720.

49. CORDUAS, Marcella; PICCOLO, Domenico. Time series clustering and classification by the autoregressive metric. Computational Statistics & Data Analysis, 2008, 52.4: 1860-1872.

50. GUHA, Sudipto, et al. Clustering data streams: Theory and practice. Knowledge and Data Engineering, IEEE Transactions on, 2003, 15.3: 515-528.

51. KRANEN, Philipp, et al. Self-adaptive anytime stream clustering. In: Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on. IEEE, 2009. p. 249-258.

52. GUHA, Sudipto, et al. Clustering data streams. In: Foundations of computer science, 2000. proceedings. 41st annual symposium on. IEEE, 2000. p. 359-366.

53. AGGARWAL, Charu C., et al. A framework for clustering evolving data streams. In: Proceedings of the 29th international conference on Very large data bases-Volume 29. VLDB Endowment, 2003. p. 81-92.

54. CHAKRABARTI, Deepayan; KUMAR, Ravi; TOMKINS, Andrew. Evolutionary clustering. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006. p. 554-560.

55. GRASSO, Marco, et al. A Comparison Study of Distribution-Free Multivariate SPC Methods for Multimode Data. *Quality and Reliability Engineering International*, 2015, 31.1: 75-96.

56. ALEXOPOULOS, Christos; SEILA, Andrew F. Implementing the batch means method in simulation experiments. In: *Proceedings of the 28th conference on Winter simulation*. IEEE Computer Society, 1996. p. 214-221

57. EILERS, Paul HC; MARX, Brian D. Flexible smoothing with B-splines and penalties. *Statistical science*, 1996, 89-102.

58. WARD JR, Joe H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 1963, 58.301: 236-244

59. DAVIES, David L.; BOULDIN, Donald W. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1979, 2: 224-227.

Appendix

K-chart

K-chart [55] is a multivariate control chart whose control statistic consists of the kernel distance of any observation z ∈ R^p from the center o ∈ R^p, of that region. The control limit is estimated to guarantee a target type I error with the available dataset. A kernel distance, denoted by kd(z), replaces the traditional Euclidean and statistical distance notions to adapt the control region boundary to the actual spread of the data; hence, the K-chart is aimed at monitoring the stability over time of the kernel distance kd(z) of any new observation z ∈ R^p from the center o ∈ R^p.

Support Vector Data Description (SVDD) will be used to obtain a minimal volume control region that adapts to the actual spread of the data.
The estimation of such a region, centered in o ∈ R^p and with radius R, requires the solution of a data-driven optimization problem:

$$\begin{aligned} &\min \left(R^2 + C\sum_{j=1}^{M}\xi_j\right) \\ &s.t. \quad \left(x_j - o\right)^T\left(x_j - o\right) \le R^2 + \xi_j\,, \qquad j = 1..M \end{aligned} \tag{A.1}$$

where ξj, j=1,…,M, are slack variables, and C is a penalty coefficient used to weight the trade-off between the volume of the region and the percentage of enclosed data (C>0).

By introducing the Lagrangian function,

$$L\left(R, o, \xi_j, \alpha_j, \gamma_j\right) = \quad R^2 + C\sum_{j=1}^{M}\xi_j - \sum_{j=1}^{M}\alpha_j\left(R^2 + \xi_j - \left(x_j - o\right)^T\left(x_j - o\right)\right) - \sum_{j=1}^{M}\gamma_j\xi_j \tag{A.2}$$

and by setting the partial derivatives with respect to R, o, and ξj, j=1,…,M, to zero, the problem (A.1) can be simplified as follows:

$$\max \left(\sum_{j=1}^{M} \alpha_j {x_j}^T x_j - \sum_{j,k=1}^{M} \alpha_j \alpha_k {x_j}^T x_k\right)$$
$$s.t \sum_{j=1}^{M} \alpha_j = 1 \quad and \quad 0 \leq \alpha_j \leq C, \quad j = 1..M \tag{A.3}$$

The points whose Lagrangian coefficients are larger than zero are known as support vectors, and it can be demonstrated that the shape of the region is determined by those points only.

By introducing the kernel trick, it is possible to replace the inner product $a^T b$, by a kernel function K(a × b) that allows the generation of a more flexible and data-adaptive control region. Hence, the kernel distance will be calculated as:

$$kd(z) = K(z \times z) - 2\sum_{j=1}^{M} \alpha_j K(x_j \times z) + \sum_{j,k=1}^{M} \alpha_j \alpha_k K(x_j \times x_k) \tag{A.4}$$

It has been showed that there are different possible approaches to the design of the K-chart because there are three major parameters to set: the kernel width parameter denoted by S, the penalty coefficient C, and the control limit denoted by h.
By comparing different design solutions, best performances might be achieved by reducing the number of parameters to two (i.e., S and h). In fact, by assuming C>1, the constraint $0 \leq \alpha j \leq C$ is replaced by $\alpha j \geq 0$, and problem (A.3) can be solved by introducing the kernel function K(x. × x.).
In this case, no penalty is applied, and hence, the kernel-based boundary is estimated by enclosing all the training data. The false alarm rate is controlled by setting a proper value for the control limit h. Thus, only the S and h parameters remain to be determined.

The following procedure can be used to automatically select those two parameters and to design the K-chart.
With respect to the kernel function, the most common choices include the Gaussian radial basis (GRB) and the polynomial and sigmoidal functions.
It has demonstrated that the GRB function is more appropriate than other kernel functions in classification problems. Thus, the GRB function is used as the default choice.

If a, b ∈ R^p, the GRB function with kernel width parameter S ∈ R^+ is:

$$K(a \times b) = exp\left\{-\frac{\|a-b\|^2}{S^2}\right\} \tag{A.5}$$

In most cases, the selection of the kernel width parameter involves trial and error. When in-process monitoring is considered, an automated data-driven procedure is required. To this aim, it has been proposed a method derived from multiclass SVM problems in which the classification errors can be used as a standard to select S.
In a one-class-classification problem, a similar approach might be applied by generating artificial outliers, drawing them from a block-shaped or a hyperspherical uniform distribution that encloses the training data in R^p.
Given $f_{o+}S$, the proportion of artificial outliers that are classified as in-boundary data for a given choice of S, and #SV(S), the number of support vectors, S can be selected by minimizing:

$$\gamma(S) = (1-v)\frac{\#SV(S)}{M} + vf_{o+}(S) \tag{A.6}$$

because #SV(S)/M is the counterpart of the type I error and $f_{o+}S$ is the counterpart of the type II error, while 0<ν<1 is a weight.

The procedure for the selection of the kernel width parameter is applied as follows:

1. Given a training set of M observations, generate a number M_o of artificial outliers;
2. Set S equal to an initial value S_0, and solve the following problem for the M+ M_o available data;

$$K_{poly}(a \times b) = (1 + a^T b)^d$$
$$K_{sigm}(a \times b) = tanh(d_1 a^T b + d_2) \tag{A.7}$$

3. Compute $f_{o+}S$ and #SV(S_0);
4. Set S equal to a new value S_0 + s, where s is a step value, and repeat steps 3 and 4 until S equals a prefixed upper limit S_U;
5. Find the value of S (known as S*) such that # SV(S*)/M is nearest to the targeted type I error;
6. Calculate the weight ν as follows:

$$v(S^*) = \left(1 + \frac{f_{o+}(S^*)}{\frac{\#SV(S^*)}{M}}\right)^{-1} \quad \text{(A.8)}$$

7. Calculate the γ(S) value in the following equation, where ν = ν(S*), for S values in the range $[S_0, S_U]$;
8. Eventually, S is determined by the minimal γ(S).

Once the optimal value of the kernel width parameter is determined, the control region can be estimated. The control limit h can be estimated as the 100(1-α)% empirical percentile of the kernel distance $kd(z_j)$, j=1, 2,…,M, where α is the targeted type I error.