4.2 Segmentation of the ECG Signal Complexes

The segmentation is one of the most important stages in the processing of ECG signals, because the analysis of the patterns that compose the ECG signal is the starting point for the compression tasks [52], filtration [62], heart rate variability studies (HRV) [63], and beats classification or grouping [64].

Due to its relevance, there is a broad-literature about segmentation of ECG signals, taking into account that those researches are focused mainly on the estimation of the fiducial mark of the R-peak, being the starting point for analysis of P and T waves.

In [65] is taken into account the non-stationarity of the ECG signal in order to apply the Wavelet Transform (WT) on the ECG signal and, in that way, to perform for a certain scale that contains the spectral information of the QRS complex the estimation by taking advantage of the fact that the R-peak is represented as a crossing zero between a maximum and a minimum on the scale. In this process, the band pass filter, the derivative and non-linear transformation stages are replaced by a specific wavelet transform (WT). Specifically, in this work is investigated a wavelet mother that represents properly the ECG signal, where the first derivative of Gaussian wavelet ranged in the scale $2^j, j = 1, 2$ is of great interest. The process is reduced to analyze the transformation through adaptive thresholding and after estimated the R-peak, the refractory period is set. Once the R- peak has been estimated, the beginning of the Q wave and the end of the S wave are found with the scale 2^1, characterizing the maximum module produced by the WT. With scale 2^3, is possible to find the P and T waves, using the maximum modules produced in that scale.

After this research, other wavelet functions were investigated that represent the QRS complex in a parsimony way, such as the Mexican Hat Wavelet on the paper [66]. In order to improve the performance of benefits, in [67] it is used a wavelet spline adjusted to the spectral features of the QRS complex. In [68], the computational cost is optimized using the multiresolution analysis with approximations and details at level $j = 2$ by using a wavelet Daubechies 4, achieving comparable results to the performance of the continuous WT.

The work presented in [69] uses a methodology which replaces the WT and applies the Hilbert transformed to the ECG signal derivation, producing crossing-zeros at R-peak location, therefore, carrying out an adaptive thresholding stage and a refractory period. The advantage of this method lies in the processing time at the Hilbert transformed calculation, which, compared to the CWT, requires fewer operations. But using

DWT requires a number of similar operations, leaving it as a parameter to choose the algorithm with the higher sensitivity and positive predictivity ($+se$ and $+P$). In a report from [70], an algorithm which consists on threshold-tuning and the determination of crossing-zeros between the threshold and the QRS complex is implemented. This makes sure that the R-peaks are contained into the set of crossing-zeros locations, providing in this way a fixed precision of the algorithm. The main part of the method is the use of the Poisson technique with the theory of root-moments, which allows moving this part of the problem in a problem of polynomial-zero estimating that lies on the circumference of the circle unit. The locator polynomial has an equal order to two times the total number of peaks from the processing block. The roots of the locator polynomial produce limits on the R-peak, allowing its detection by thresholding.

In [71], it is performed a quantitative analysis of three classic segmentation algorithms corresponding to the Hilbert transformation, transformation of the signal (quadratic function) and second derivative of the signal. The first two algorithms were measured over 99%, while the second derivative approach had less performance. Also, it is proposed to mix up the first two methods with their best features in order to have better results.

In general, the ECG transformation methodology through WT is the most used for ECG segmentation, finding a current research [72], which makes a novel detection based on a wavelet pre-filter and an adaptive-thresholding technique. The algorithm uses a bi-orthogonal wavelet filter to perform denoising on the signal. The QRS complexes are identified by calculating the first derivative of the signal and applying a set of adaptive thresholds that are not limited in a narrow range. The QRS complexes are identified in multiple ECG channels with a 5-leads configuration. The proposed algorithm is able to detect QRS complexes achieving high values of $+Se$ and $+P$, besides the possibility of real-time implementation. A current relevant work is presented in [73], which uses a quadratic spline wavelet for the segmentation, following the methodology discussed above.

There are other kinds of methodologies that use an automated process for segmentation task, avoiding as much as possible to use heuristics stages, such as thresholding-stage that somehow limits the performance of algorithms for some specific data (e.g. morphology, noise level of the signal).

In [74], it is implemented an optimization of stages for detecting R-peaks of ECG signals, using genetic algorithms (GA). Two stages are considered. The first one consists of enhancing the QRS complex with respect to P and T waves, which needs the

use of a polynomial filter that operates over a small number of selected input-samples through GA. For the second stage, a maximums detector is applied, where the threshold Y is performed in order to avoid false detection of QRS. The detector-designing requires the definition of the polynomial filter features as well as the selection of the coefficients and parameters of the maximums detector. The polynomial order and the number of input samples affect the number of operations per sample. For a maximum efficiency, only low-order polynomials and a limited number of input samples must be considered. However, the use of a small number of samples does not mean that the filter will work in a short region of the signal, but the delay $d_1, ..., d_N$ can be selected by the GA. In general, the coefficients of a polynomial filter, $a_{k_1}, ..., a_{k_n}$, and the parameters of the detector should always be under genetic optimization, except a_0, which should be to set as zero, and it is not independent of such parameters. It is also necessary to define an adaptability function, that must decrease according to the number of false detections (F_P) and missed detections (F_N) produced when an ECG set is trained. As conclusion about the GA, it can be said that GA algorithm optimizes the parameters of the maximums detector and the filter coefficients according to single-criterion: minimizes the number of detection errors. Finally, the joint-optimization of the two stages of the detector was successfully adapted, which helped to find robust parameters in order to detect the QRS with few operations per sample.

In [75], the segmentation of ECG signal is performed using the best basis algorithm, commonly used in tasks for abrupt changes detections in signals. The best-basis notion for a given signal, takes into account the parsimonious representation resultant for some basis, which is normally selected using the entropy criterion based on a metric called, Energy Concentration Measure (ECM). This method has been applied successfully in representation wavelet packet, and it has been proposed for the segmentation of time/space, using local trigonometric basis. The aim of this method is to capture the localized-coherent dominant structures and assign them a corresponding segmentation. Although the ECM criterion has gotten great results for class segmentation of specific signals, is not universal, it is to say, not only a desirable segmentation may need to reflect the morphology of the signal, but also needs to optimize its parsimony representation. Brooks [75] explains that the usual entropy criterion has errors in the segmentation of ECG signals, because of its highly dynamic and Local SNR variability. That is why the research presents a new criterion that reflects the morphological structures for an optimal segmentation. Visually, the discrimination of ECG segments, not only use amplitude but also smoothness and curvature information (at least the first

and second derivatives), therefore it is proposed a criterion that reflects not only the parsimony-representation but also smoothness. The criterion uses a function suitable-constructed about the expected signal smoothness with a measure of entropy that reflects its Parsimony- representation. The criterion is based on an entropy function $\phi_1(W_{(.,.)})$ and for the smoothness signal morphology as its variation defines a function $kW_{(.,.)}$, Which is used in an extra-cost function $\phi_2(kW_{(.,.)})$. For the search-criterion, where $\phi_2(.)$, is decreasing monotonic function. Therefore, this criterion not only penalizes the tendency of over-segmentation but also penalizes non-softness changes in the signal, when in fact weak changes are expected at specific intervals (eg. ST segment - T wave).

Another type of automated method for ECG segmentation, developed by literature, corresponds to the Hidden Markov models (HMM). Among relevant researches, there is [76], which proposes a new algorithm for segmenting ECG waves (P, QRS, T) that are based on semi-supervised learning, using the maximization algorithm of hope (EM), in order to estimate the maximum verisimilitude, which is used for probabilistic models labeled in a subjectively way, it is to say, the data labels are not assumed as perfect. The first stage of development is the modelling of the signal, which is made using the Undecimated Wavelet Transform (UWT), which has certain advantages in time-scale representation of the signal. When the representation of signal is gotten, some probabilistic models are used to analyze the statistics features of the UWT coefficients in the task of automatic signal segmentation. Two methods are proposed for the probabilistic model, HMM and HSMM, where the last one (Semi-Markov's Hidden Chains), shows better results than HMM, by the fact that auto-transition coefficients are set to zero and an explicit probability function is specified for the duration of each state. In this way, the distributions of the individual state duration rules the amount of time, which the model uses in a given state, and the transition matrix rules the probability of the next state once the time is up. Finally, getting the representation of the signal model and its architecture, it allows to proceed to the learning parameters for an optimal segmentation, using semi-supervised learning.

Generally, some researches using several techniques for ECG signal segmentation that have been exposed in this review. Taking into account that most of them uses the R-peak estimation, those based on WT and HMM employ methodologies to segment the remaining waves of the ECG signal with good results in termos of Se and P. It should be noted that most of the works are oriented to WT methodologies. Table 4.1 shows some relevant results that are published in the literature in terms of performance and

Table 4.1: Comparison among *QRS* detectors of the literature

Database	QRS detector	# labels	Se %	P %
MIT/BIH	Martinez et al [73]	109428	99.8	99.86
	Aristotle [15]	109428	98.3	99.91
	Li et al [77]	104182	99.89	99.94
	Alfonso et al [78]	90909	99.59	99.56
	Bahoura et al [79]	109809	99.83	99.88
	Lee et al [80]	109481	99.69	99.88
	Hamilton et al [81]	109267	99.69	99.77
	Pan et al [82]	109809	99.75	99.54
	Poli et al [83]	109963	99.6	99.5
	Madeiro et al [66]	109494	98.47	98.96
	Moraes et al [84]	N/R	99.22	99.73
	Hamilton et al [81]	N/R	99.8	99.8
QTDB	Martinez et al [73]	86892	99.92	99.88
	Aristotle [15]	86892	97.2	99.46
	Madeiro et al [66]	86995	99.56	99.74
CSE	Gritzali [85]	19292	99.99	99.67

fiducial estimation marks, where the number of inputs is compared, i. e., the number of assessed beats, the sensitivity results and prediction. The abbreviation N/R in some rows corresponds to (not reported)

4.3 Feature Extraction and Selection of ECG Signals

4.3.1 Feature extraction of ECG signals

The stage of extraction and feature selection is used in several focuses for an ECG signal analysis. The most investigated belongs to the classification and support tasks in the diagnosis of cardiac pathologies. To carry out this task, it is selected from a dataset the minimum relevant information that allows to identify a set of classes. Usually, the dataset corresponds to the samples of the signal, representative coefficients of the signal or time series formed from the *RR* distances (HRV: Heart Rate Variability). Another focus belongs to filtering tasks, compression and clustering of ECG signals, using a great variety of applications and system analyzers of biological signals. Either way, there are several methods for feature extraction: Heuristic methods, methods

that use statistical information in the signal (PCA [86], NLPCA [87]), methods for representation through basis (Hermite [88], WT [77], time-frequency distributions [89].), polynomial approximation methods or curves [64], nonlinear sample of the signal [64], among others. Working with signal features or signal transformed, has certain advantages over working directly with the samples, especially, the fact that the signal can be affected by biological or artificial perturbations, as it was discussed in previous sections. On the other hand, it is necessary to do the effective-selection stage of features when the features are even in a large space of analysis in order to reduce a number of parameters and obtain classification rates equally high.

Taking into account that specialist doctors use rules based on ECG signal features such as amplitudes, slopes, distances, morphology, in order to do the diagnosis and classification tasks, some cases work with calculated parameters to perform automated tasks of characterization or diagnosis. In [90], it intends to make the ECG characterization from its segmentation. The parameters obtained are both the beginning and the ending of the P and T waves, and the QRS complex. The main objective on the paper is to perform the extraction in real time and provide information to the specialist for further analysis. In [91], it shows the research on efficient feature extraction of ECG signal to improve the performance of automatic detection and classification of cardiac arrhythmias. The features that form the input vector for a neural network, are divided into two groups: *i)* morphological features and extracted statistics from the signal (RS interval, QRS area, R-R interval, R amplitude, ST segment area, T wave amplitude, signal energy, QRS energy, auto-correlation coefficient, maximum amplitude of the signal histogram), *ii)* A compressed form of the ECG signal pre-aligned with a 4:1 compression ratio (13 features). The intention is to classify the 23 features in four pathologies, obtaining a classification error of 0.95%.

In [92], heuristics parameters are used for classification purposes. The selected parameters are the amplitude and duration of P, T waves, the QRS complex, ST segment, P-R, Q-T intervals. In order to extract these parameters, is used the WT and a smoothing function Θ, using a cubic and square spline.

Several studies have been published, which instead of using diagnostic features, use obtained features by applying transformations or representation of the signal. Alike or superior results have been published in performance of computational cost, and performance in the classification.

In [93], the feature extraction is performed based on generalized-methods that are applied mainly to time series. For this task, approximations by segments of the signal,

using six defined functions (constant, straight, triangular, trapezoidal, exponential and sinusoidal) are accomplished. Each of the functions is configured through parameters to fit in the signal using the Mean Square Error criterion. Functions were tested independently as well as itself combinations on the signal. To measure the effectiveness of the technique, two signal-classes were classified (abnormal and normal), using a decision tree-classifier (CART) and the percentage of classification performance, they were compared with techniques such as identity, Fourier and WT transforms. Superior results were obtained, but not exceed 90% of classification rate.

There is a parametric model, called the Hermite model, which has been used extensively in feature extraction of ECG signals for both compression and classification tasks [94], [95], [88].

The Hermite bases do not depend on signal statistics, but are fixed, except for one width parameter (σ). The Hermite base-functions have the property that an arbitrary bounded signal in time can be represented by a single sum of these functions. The error in the approximation can decrease by increasing the number of base-functions used in the expansion. In [94], it was determined for example, that on average 98.6% of the energy of the QRS complex can be represented using three Hermite coefficients.

In [88], Hermite bases are chosen to extract used-parameters for the detection of Acute Myocardial Infarction (AMI), using Artificial Neural Networks (ANN) in 12-leads ECG signals. To make the process, the ECG signal was segmented and the QRS complex and T wave were represented through the first 11 bases, obtaining in this way 11 coefficients used as features. To measure the effectiveness of the method, the set of features was put into a Bayesian classifier and the performance was compared with a specific heuristic features. A similar performance of 84.3% was obtained for heuristic features, compared with 83.4% for Hermite features. Although performance is slightly smaller, the processing time is better and additionally, this type of representation is invertible.

In [95], a system of detection, classification and identification online of QRS complexes is developed, in which one of the methods for feature extraction is the Hermite model. The calculation of the parameter is automated from Levemberg-Marquardt algorithm. It also employs from Hermite coefficients, some heuristic features, obtaining classification rates around 93% of 75988 normal and abnormal heartbeats from the MIT/BIH database. Although the process is relatively fast, the calculation of the parameter should be conducted for each QRS complex detected, which delays the process for real-time requirements.

Another technique used for the extraction of ECG parameters is the Karhunen-Loéve Transform (KLT), which is orthogonal-linear and optimal in the sense of Mean Square Error (MSE), it means, concentrating the signal information in a minimum number of parameters. It has additional properties such as minimum entropy representation and uncorrelated coefficients. Performing an analysis of the expansion of a time serie in an orthonormal base vector, can reach the problem of eigenvalues, so that the eigenvectors corresponding to the highest eigenvalues, represent more proportion of projected energy. Therefore, as feature extractor of order N, the KLT, uses the first N eigenvectors in descending order. The KLT has been used extensively in analysis of ECG signals, as in ST segment analysis and T wave [96], and compression of ECG [52].

In [95], the feature extraction method used is the KLT, compared with the Hermite model discussed above. The KLT eigenvectors are obtained from the segmented QRS complexes and are taken the first 6 bases of representation. It should highlight the fact that for getting the bases, a training set of signals was used, corresponding to the MIT/BIH database. As a result, a good estimation of the two methods is obtained (Hermite and KLT), for the classification of QRS complexes, and finally proposing to use the KLT as features extractor, because its bases are calculated off-line and thus, has less computational cost than the Hermite model.

In [96], KLT is proposed to be used to represent ST-T complexes from recordings of patients with myocardial ischemia induced by PTCA (Percutaneous Transluminal Coronary Angioplasty). The research compares the system performance between features obtained with KLT and heuristic methods from ST-T complexes. Once the KLT bases are gotten with a training set of approximately 200.000 complexes, a sensitivity study of the obtained-parameters with both techniques is conducted. The work concludes that the KLT features have lower sensitivity than the heuristic features, resulting in better classification rates.

In [87], the technique used is the NLPCA , which improves the performance of PCA, because it does not depend on second-order moments of signal and is implemented with a multilayer neural-network. It has been noticed a superior performance than PCA in problems where the relationships between variables are nonlinear. NLPCA is used to classify ECG signal segments into two classes: normal and abnormal (ST +, ST-, or artifacts). During the training stage of the algorithm, it was only used normal patterns, and just for purposes of classification only two non-linear features were used for each segment ST. The distribution of these features was modeled using a radial basis function (RBFN). The results of the tests using the European database ST-T showed

that only two non-linear components and a training set of 1000 normal samples of each record, produce a classification rate of approximately 80% for normal beats and greater than 90% for ischemic beats.

Some time-frequency distribution techniques have been used for feature extraction of ECG signals. In [89], an investigation from the performance of feature selection techniques to extracted parameters in time-frequency distributions is produced. The signal is processed to obtain 25 parameters after applying a Wigner-Ville distribution. The methods for effective selecting features were PCA, self-organizing maps (SOM) and decision trees (*CART*). Four signal classes were identified: normal sinus rhythm, ventricular fibrillation, ventricular tachycardia, and other rhythms. It was found that all methods have optimal results in the task of sorting and greatly simplify the computation time. In [97], it is proposed to use an instantly controlled distribution of Descending Radially Gaussian Kernel, for diagnosis tasks of ischemia without take into account the stage of angina.

As it was previously discussed, due to the non-stationary of ECG signal, its analysis using the WT, has given good results in the most of signal analysis tasks, from preprocessing to classification. Therefore, for feature extraction are several articles in the literature.

In [98], WT is used by a Mexican Hat mother wavelet, which has a fit to the ECG signal according to [99]. Dyadic scales were used 2^j, $j = 1, 2, 3$, in order to get the parameters. The transformations are the input to a Hidden Markov Model (HMM) model for purposes of characterization of the signal at its main components.

In [100], the performance of some methodologies is investigated for diagnosis of 10 different types of arrhythmias. One of them corresponds to WT-NN, which extracts features with WT and these features are entered in a classical multilayer perceptron neural-network with backpropagation training. This serves as a reference for the proposed technique on *FCM-PCA-NN*, where features are extracted with a clustering method (Fuzzy C-means), the features are selected with PCA and finally they are classified by the neural network. For this process, the WT corresponds to the calculation of coefficients using a Daubechies mother wavelet of order 2, working with Mallat algorithm (DWT), where the coefficients of approximation and detail are used as features. The importance of wavelet coefficients is, that allow a compact representation and show the energy distribution of the signal in time and frequency. As a result, even though there is a classification performance in order of 99% with the proposed methodology, the technique *FCM-PCA-NN* has lower computational cost, being adequate to the task

of classification.

Some of the relevant researches in extraction tasks of ECG signal have been exposed. It can be concluded, that the techniques of representation as the Hermite parametric model, techniques based on statistical signal such as PCA or time-frequency representation techniques like WT, are powerful tools to extract relevant information of the signal, whether it is for segmentation, classification or compression purposes.

4.3.2 Feature selection for classification

Coming with the rapid growth of high dimensional data collected in many areas such as text categorization and gene selection there is an increasing demand for the feature selection in classificatory analysis [101, 102]. To describe the domain of applications as good as possible, real-world data sets are often characterized by many irrelevant and/or redundant features due to the lack of prior knowledge about specific problems [103]. If these features are not properly excluded, they may significantly hamper the model accuracy and the learning speed. Because the primary task of classificatory analysis is to extract knowledge (e.g., in the form of classification rules) from the training data the presence of a large number of irrelevant or redundant features can make it difficult to extract the core regularities of the data. Conversely, if the learned rules are based on a small number of relevant features, they are more concise and hence easier to understand and use [102, 104]. Therefore, it is very important to reduce the dimensionality of the raw input feature space in classificatory analysis to ensure the practical feasibility of the classifier. Feature selection is to select a subset of original features that is good enough regarding its ability to describe the training data set and to predict for future cases. Broadly, methods for feature selection fall into three categories: the filter approach, the wrapper approach and the embedded method. In the first category, the filter approach is first utilized to select the subsets of features before the actual model learning algorithm is applied. The best subset of features is selected in one pass by evaluating some predefined criteria independent of the actual generalization performance of the learning machine. So a faster speed can usually be obtained. The filter approach is argued to be computational less expensive and more general. However, it might fail to select the right subset of features if the used criterion deviates from the one used for training the learning machine. Another drawback involved in the filter approach is that may also fail to find a feature subset that would jointly maximize the criterion, since most filters estimate the significance of each feature just by means of

evaluating one feature a time [105]. Thus, the performance of the learning models is degraded. Methods from the second category, on the other hand, utilize the learning machine as a fitness function and search for the best subset of features in the space of all features subsets. This formulation of the problem allows the use of the standard optimization techniques with the learning machine of interest as a black box to score subsets of features according to their predictive power. Therefore, the wrapper approach generally outperforms the filter approach in the aspect of the final predictive accuracy of a learning machine. The wrapper methodology is greatly popularized by Kohavi and John [106], and offers a simple but powerful way to address the problem of feature selection, despite the fact that involves some more computational complexity and requires more execution time than that of the filter methodology. Besides wrappers and filters, the embedded methods are another category of feature selection algorithms, which perform feature selection in the process of training and are usually specific to given learning machines [102]. Some examples of the embedded methods are decision tree learners, such as tree decision, or the recursive feature elimination (RFE) approach, which is a recently proposed feature selection algorithm derive based on support vector machine (SVM) theory and has been shown good performance on the problems of gene selection for microarray data [107, 108]. The embedded methods are argued to be more efficient because they avoid retraining a predictor from the scratch for every subset of features investigated. However, they are much intricate and limited to a specific learning machine. Recently, research on feature selection in mainly focused on two aspects: criteria and search strategies. As we known, an optimal subset is always optimal relative to a certain criterion. In general, different criteria may not lead to the same optimal feature subset. Typically, a criterion tries to measure the discriminating ability of a feature or a subset to distinguish the different class labels. M. Dash called these criteria the evaluations functions and grouped them into five categories [103]: distance, information (or uncertainty), dependence, consistency and classifier error. The distance measure, e.g., the Euclidean distance measure, is a very traditional discrimination or divergence measure. The dependence measure, also called the correlation measure, is mainly utilized to find the correlation between two features or a feature and a class. The consistency measure relies heavily on the training data set and is discussed for feature selection in [109]. These three measures are all sensitive to the concrete values of the training data; hence they are easily affected by noise or outlier data. In contrast, the information measures, such as the entropy or mutual information, investigate the amount of information or uncertainty of a feature for the

classification. The data classification process is aimed at reducing the amount of uncertainty or gaining information about the classification. In Shannon's information theory [110], information is defined as something that removes or reduced uncertainty. For a classification task, the more information we get, the higher the accuracy of a classification model becomes, because the predicted classes of new instances are more likely to correspond to their true classes. A model that does not increase the amount of information is useless and its prediction accuracy is not expected to be better than just a random guess [111]. Thus, the Information measure is different from the above three measures by its metric-free nature: it depends only on the probability distribution of a random variable rather than on its concrete values. The Information measures have been widely used in feature selection [112–115], including many famous learning algorithms such as tree decision and C4.5.

Searching for the best m features out of n available for the classification task is known to be a NP-hard problem and the number of local minima can be quite large [116]. Exhaustive evaluation of possible feature subsets is usually unfeasible in practice due to the large amount of computational effort required. A wide range of heuristic search strategies have been used including forward selection [112], backward elimination [117], hill-climbing [118], branch and bound algorithms [119], and the stochastic algorithms like simulated annealing [120] and genetic algorithms (GAs) [121]. Kudo and Sklansky [122] made a comparison among many of the feature selection algorithms and explicitly recommended that Gas should be used for large-scale problems with more than 50 candidate variables. They also described a practical implementation of GAs for feature selection. The advantages of GAs for feature selection are often summarized as follows: First, compared with those deterministic algorithms, they are more capable of avoiding getting stuck in local optima often encountered in feature selection problems. Second, they may be classified into a kind of anytime algorithms [123], which can generate currently best subsets constantly and keep improving the quality of selected features as time goes on. However, the limitations of a simple GA algorithm have been uncovered in many applications, such as premature convergence, poor ability of fine-tuning near local optimum points. A practical and effective way to overcome these limitations is to incorporate domain-specific knowledge into the GA. In fact, some hybrids GAs have been deployed in diverse applications and successful performance has been obtained [124].

4.4 Classification of Cardiac Arrhythmias

Next, some works related to classification and diagnosis for ECG signals will be reviewed, taking into account two different approaches: supervised and unsupervised one.

4.4.1 Supervised classification of cardiac arrhythmias

Among supervised classification methods that have been applied to biosignals processing, it can be distinguished the statistical, syntactic and artificial intelligence methods. There exist a particular interest in Neural Networks in the ECG processing field, where some works have been proposed, for example in [125] and [8], approaches based on the well known multilayer perceptron, self-organizing networks, fuzzy or neuro-fuzzy based-systems and hybrid systems are proposed.

In [126], it is presented a comparison of different wavelet subband features for the classification of ECG beats using probabilistic Neural Network PNN to discriminate six ECG beat types. The effects of two wavelet decomposition structures, the two–stage two–band and the two–stage full binary decomposition structures in the recognition of ECG beat types are studied. The ECG beat signals are first decomposed into components in different subbands using discrete wavelet transformation. Three statistical features of each subband-decomposed signal as well as the AC power and instantaneous RR interval of the original signal are exploited to characterize the ECG signals. A PNN then follows to classify the feature vectors. The results show that features extracted from the decomposed signals based on the two–stage two–band structure outperform the two–stage full binary structure. A promising accuracy of 99.65%, with equally well recognition rates of over 99% throughout all type of ECG beats, has been achieved using the optimal feature set. Only 11 features are needed to attain such performance. The results demonstrate the effectiveness and efficiency of the proposed method for the computer–aided diagnosis of heart diseases based on ECG signals. However, the ECG beat types analyzed in this work do not fulfil the requirement of the AAMI (Association for the Advanced of Medical Instrumentation), which proposes standards to assess the performance of algorithms that analyze disorders of rhythm [127]. Some works that fulfil this requirements have been found: two of them use supervised classification, [128] and [129]. The third one uses non-supervised classification [8].

The work developed in [128] presents a specific classifier of heartbeats for a par-

ticular patient (known as local classifier) which is combined with a global classifier designed from a ECG training data base. Classifiers were combined by employing a mixture of experts (*MOE*). Local classifier requires that a cardiologist take notes about signal segments from a specific patient in order to implement the *MOE* method. Experiments show that global classifier achieve 62.2% of effectiveness, while *MOE*-based classifier achieve 94%.

In [129], a methodology for ECG heartbeats detection using preprocessing and supervised classification techniques is presented. This methodology includes several stages. First, high and low frequency signal disturbances are filtered by means of digital classical filters. In this work do not apply an algorithm to detect the *R*-peak but fiducial marks in the recordings provided by data base used, in this case, MIT/BIH data base. To segment the ECG signal waves, authors used the software *puwave* that is available online and was developed by *Laguna et al*[1]. In feature extraction stage, 15 heuristic parameters were analyzed, which are obtained from: signal morphology (8), time analysis (period *R-R* (4)) and complexes duration (*P*, *QRS*, *T* (3)). Linear discriminant analysis (*LDA*)is used for classification stage. This model is done by calculating the estimated values of maximum likelihood over training data.

A feasible solution for this problem is including a sub-sample of the majority classes in the training process despite some data points are wasted. However, another work [129] includes all training samples by reducing the relative contribution of majority classes. This is done by measuring the contribution of each training sample in the likelihood function multiplied by a factor that depends on the class ω_k and the number of training samples. Two *LDA* based classifiers were implemented (each classifier processes one signal channel). In order to obtain a final decision, classifiers were combined. After making 12 combinations of features and classifier, it is performed better on sensitivity (75.9 %), prediction positive (38.5 %) and false positive rate (*FPR*) 4.7 % for the class *SVEB* (Supraventricular Ectopic Beats). These results are slightly higher than reported in literature.

4.4.2 Non-supervised classification of cardiac arrhythmias

In [8], taking into consideration the recommendations given by AAMI to develop algorithms for heartbeats processing, a clustering procedure to group prior known-class heartbeats is presented . This method uses the hermite model to represent each com-

[1]"ecgpuwave": dirección internet: http://www.physionet.org/physiotools/software-index.html

plex and self-organized maps (SOM) for clustering. It was found that, by using a R-peak estimation as is described in *Nygards et al*, *QRS* complexes can be extracted properly (99.7% of accuracy).

Complexes are represented with polynomials of order $n = 1, \ldots, 5$, which present a good trade-off between classification performance and computational cost. Method is evaluated by employing a 5×5 out matrix in three different e independent ways.

When comparing the results of clustering with other methods supervised learning for classification of 25 clusters according to dominant beats, the unsupervised method exceeds supervised one. It should be noted that this method provides topological information, which can be completely used by the cardiologists for diagnosis.

Summarized below are studies that use an unsupervised scheme and, despite they do not take into account the AAMI standards, have achieved significantly good results in the analysis of arrhythmias.

In [6] it is presented a comparative analysis of procedures for arrhythmias classification by using two methods. Reference method corresponds to neuronal networks with architecture type mulilayer perceptron (MLP) and backpropagation training. Proposed method is a neuronal network (NN) combined with fuzzy clustering (FC), and is called FCNN. ECG Signals are taken from MIT/BIH data base and they are used to train the classifier for 10 different types of arrhythmias. Results proved that FCNN method can generalize and learn better than classical architecture MLP. Also, FCNN method works faster. The main advantage of proposed methods consist of decreasing the number of segments per group into the training data with fuzzy clustering via c-means algorithm.

In [130], a clustering methodology for ECG signals heartbeats from MIT/BIH data base is proposed. First stage correspond to the estimation of R-peak location through the algorithm proposed in [131] that uses first derivative of ECG signal, non-linear transformations and adaptive thresholding to estimate the fiducial mark.

Subsequently, heartbeats are extracted taking into account the following criterion: 20% of distance between current R-peak and the preceding one (RR) to determine the start of heartbeat, and 80% of distance between current R-peak and the following one to determine the end. In order to obtain a correspondence among heartbeats for later comparison, it is done an amplitude and time normalization process over signals. Time normalization is carried out by means a method called DTW (Dynamic Time Warping) with local and global constraints to improve the computational cost, in that way results are similar those obtained without using constraints. This technique is based on subsampling and uniform interpolation procedures applied on times series to

be compared.

The following stage corresponds to heartbeat feature extraction, in which four techniques are taking into consideration: signal samples, trace segmentation (non-linear signal sampling), polygonal approximations and WT. Next, for classification stage is first performed a heartbeat labeling in order to measure clustering algorithms performance by means of supervised measures. Heartbeats features are compared by means of similarity measures corresponding to L_1 and L_2 *Minkowski* norms that have been widely used.

First, a pre-clustering stage is applied to decrease the amount of heartbeats for analysis. This stage takes importance because a Holter recording can store hundreds of thousands of heartbeats. To carry out this task a dissimilarity measure (DTW) among beats with a relatively low threshold is applied. Given this, heartbeats that present certain likeness with compared ones are discarded.

Finally, for clustering stage, two algorithms widely used and recommended by literature are implemented. First method is called *Max–Min* algorithm that is a non-parametric partitional clustering based on a criterion to decrease computational cost. Second one corresponds to k–means which is partitional and non-parametric but based on re-calculation of center.

In this case it is necessary a complete center initialization, then, if a random initialization is chosen, a resultant partition can be obtained. Given that re-calculation of center is not possible in a *Euclidean* space because of heartbeats length variability, another criterion is applied that corresponds to the median. In such a way, k-means algorithm is modified to be k-medians.

After several combinations and tests between feature extraction and clustering algorithms, it was concluded that the combination that presented better performance in clustering process consist of trace segmentation as feature extraction and k-medians as clustering algorithm.

The work developed in [132] applies the results obtained in [130] on classification of *VE* (Ventricular extrasystole), following a similar scheme. First, R-peak for each heartbeat is identified. The heartbeats are extracted and some intrinsic features of *VE* are including in the feature extraction process. The estimated features consist of RR interval and a polarity index that measures the relative position of the average value of beat amplitude between maximum and minimum values. These two features with samples obtained by applying trace segmentation form the feature set to be evaluated by the classification stage. For centroids initialization, some restrictions that takes

into account the *VE* beat morphology are applied in order to reduce computational cost. Finally, it is implemented the k-means clustering algorithm to group the beats. Experiments showed results over the 90 % of sensitivity being comparables with the best works of literature.

In [133] the performance of methods for segmentation and clustering to be applied on signals from MIT/BIH and ST-T/VALE for European society data bases is studied. The proposed segmentation algorithm is sensitive to noise and works in the time domain and is founded on the concept of *curve length*, where time parameters are extracted and used as a decision point to finally determine the *QRS* complex.

Despite this algorithm presents good results regarding other classical segmentation algorithms, it can be affected by morphologies where the R-peak is short with respect the rest of complexes. Once estimated the fiducial mark, heartbeats are extracted by determining the representative points of the remaining complexes.

The next stage correspond to heartbeat feature extraction by using the three first components of PCA to characterize each complex. This is because of the accumulated variance of three first components represents more than 90%. These features are grouping by means of a clustering algorithm, as is described in [134], that consist of a modification of k-means method, called kh-means. The biggest problem of k-means lies in the sensitivity to initial partition selection, converging to a local minimum of the objective function when the centers are not properly chosen. The algorithm *kh-means* solves this problem by replacing the minimum distance of a sample at the centers by the harmonic mean of the distance between samples and centers. The algorithm shows better performance than the k-means, mainly when replacing the harmonic mean by strategy *winner-takes-all*, that commonly uses the k-means. [133] shows that the computational cost of the algorithm is less than classical clustering algorithms. Although, in this work is not specified the number of conditions to be analyzed, it presents overall results, where 97% of beats from databases was correctly clustered.

The study presented in [135] is a comparative analysis of similarity measures applied on clustering of complexes obtained from MIT-BIH database. There are four similarity measures: Manhattan (L_1), Euclidean (L_2), correlation coefficient and the Gray relation degree. The clustering algorithm discussed corresponds to method called the two-step unsupervised method, which reported better performance than hierarchal clustering algorithms. To avoid dependence on the initial partition, five random initial partitioning were done for each method, selecting the average of five iterations. It was performed a threshold setting for each iteration in order to obtain the best result.

The classification performance was used to measure similarity measures performance. As a result, the gray measure introduced in the worst case with a performance of 97.55% compared to 3 % below the percentage in other measures. For the other four iterations, performance exceeds 99 %. Finally, it is concluded that in the proposed clustering algorithm, gray measure provides better results.

In general, in this review were presented several techniques for both heartbeat and *QRS* complexes classification for diagnosis of pathologies. Supervised and unsupervised techniques were reviewed. Within the review are the NN, LD, SVM, with different configurations for supervised technics. Algorithms such as SOM, K-means, K-medians, KH-means, max-min, who have presented remarkable results in the analysis of data bases such as MIT/BIH and ST-T, correspond to non-supervised techniques that will be used in this work.

Part II

Theoretical Background

Chapter 5

Preprocessing and Feature Estimation

The ECG signal is often contaminated by relatively strong disturbances, which can modify the ECG signal shape or which can manifest with similar morphologies as the ECG itself. It becomes difficult to the specialist to diagnose diseases if the artifacts are present in the signal. Likewise, disturbances can decrease the performance of preprocessing algorithms such as waves segmentation or feature estimation. In addition, ECG variability makes necessary the use of procedures carefully selected to characterize and estimate signal complexes. In this chapter the methods to accomplish that features are discussed.

5.1 Wavelet Transform

In this work, the Wavelet Transform (WT) is broadly used, which is a fundamental tool in preprocessing and characterization procedures. Therefore, in this section, the definition, types and multiresolution analysis of WT are briefly analyzed.

Wavelet analysis provides information that is localized in frequency and in time, which makes it highly suitable for analysis of non-stationary signals and in this context, applications in biosignal analysis, such as, signal denoising, wave detection, data compression, feature extraction, among others. The analysis is carried out using finite basis functions termed wavelets. These basis are actually a family of functions which are derived from a single generating function called the mother wavelet by translation and dilation operations. Dilation, also known as scaling, compresses or stretches the

mother wavelet and translation shifts it along the time axis [36], [136], [58].

The WT is classified into Continuous (CWT) and Discrete (DWT) wavelet transform. The former transform is defined by,

$$X_w(a,b) = \int_{-\infty}^{\infty} x(t)\psi^*_{a,b}(t)dt, \tag{5.1}$$

where $x(t)$ represents the analyzed signal, while a and b represent the scaling factor, i.e. the dilation/compression coefficient, and translation along the time axis (shifting coefficient), respectively. The superscript $(*)$ denotes the complex conjugation. The function $\psi_{a,b}(\cdot)$ is obtained by scaling the wavelet at time b and scale a as the next expression:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{5.2}$$

where $\psi(\cdot)$ is a continuous function in both the time domain and the frequency domain and represents the mother wavelet. The main purpose of the mother wavelet is to provide a source function to generate the daughter wavelets which are simply the translated and scaled versions of the mother wavelet. To recover the original signal $x(t)$, inverse continuous wavelet transform can be exploited:

$$x(t) = \int_0^{\infty}\int_{-\infty}^{\infty} \frac{1}{a^2} X_w(a,b)\frac{1}{\sqrt{|(a)|}}\tilde{\psi}\left(\frac{t-b}{a}\right)\, db\, da \tag{5.3}$$

where $\tilde{\psi}(t)$, is the dual function of $\psi(t)$. The dual function should satisfy:

$$\int_0^{\infty}\int_{-\infty}^{\infty} \frac{1}{|a^3|}\psi\left(\frac{t_1-b}{a}\right)\tilde{\psi}\left(\frac{t-b}{a}\right)\, db\, da = \delta(t-t_1) \tag{5.4}$$

Sometimes, $\tilde{\psi}(t) = C_{\psi}^{-1}\psi(t)$, where,

$$C_{\psi} = \frac{1}{2}\int_{-\infty}^{+\infty} \frac{\left|\hat{\psi}(\zeta)\right|^2}{|\zeta|}d\zeta \tag{5.5}$$

is called the admissibility constant and $\hat{\psi}$ is the Fourier transform of ψ. For a successful inverse transform, the admissibility constant has to satisfy the admissibility condition, $0 < C_{\psi} < +\infty$. It is possible to show that the admissibility condition implies that $\hat{\psi}(0) = 0$ so that a wavelet must integrate to zero [137].

Continuous, in the context of the WT, implies that the scaling and translation

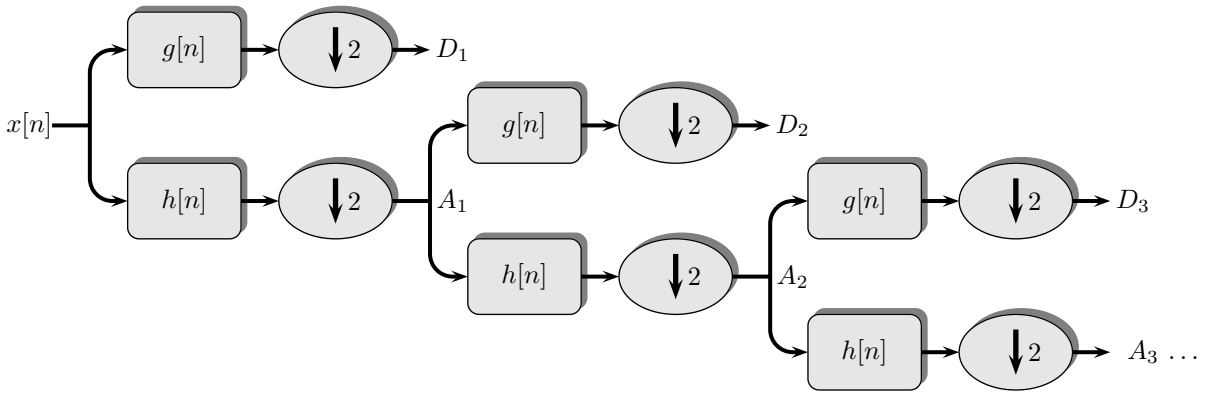


Figure 5.1: Subband decomposition of discrete wavelet transform implementation; $g[n]$ is the high-pass filter, $h[n]$ is the low-pass filter.

parameters a and b change continuously. However, calculating wavelet coefficients for every possible scale can represent a considerable effort and result in a vast amount of data. In this way, the second transform termed Discrete Wavelet Transform (DWT) is often used.

The DWT, which is based on subband coding is found to yield a fast computation of Wavelet Transform. It is easy to implement and reduces the computation time and resources required. In CWT, the signals are analyzed using a set of basis functions which relate to each other by simple scaling and translation. In the case of DWT, a time-scale representation of the digital signal is obtained using digital filtering techniques. Such process called multiresolution analysis, which decomposes a signal $x[n]$ is schematically shown in Figure 5.1. Each stage of this scheme consists of two digital filters and two downsamplers by 2. The first filter, $g[\cdot]$ is the discrete mother wavelet, high-pass in nature, and the second, $h[\cdot]$ is its mirror version, low-pass in nature. The downsampled outputs of first high-pass and low-pass filters provide the detail, D_1 and the approximation, A_1, respectively. The first approximation, A_1 is further decomposed and this process is continued as shown in Figure 5.1 [136]. All wavelet transforms can be specified in terms of a low-pass filter h, which satisfies the standard quadrature mirror filter condition:

$$H(z)H(z^{-1}) + H(z)H(-z^{-1}) = 1, \tag{5.6}$$

where $H(z)$ denotes the z-transform of the filter h. Its complementary high-pass filter can be defined as,

$$G(z) = zH(-z^{-1}) \tag{5.7}$$

A sequence of filters with increasing length (indexed by i) can be obtained:

$$\begin{aligned} H_{i+1}(z) &= H(z^{2^i})H_i(z) \\ G_{i+1}(z) &= G(z^{2^i})H_i(z), \end{aligned} \tag{5.8}$$

where $i = 0, \ldots, I-1$, with the initial condition $H_0(z) = 1$. It is expressed as a two-scale relation in time domain

$$\begin{aligned} h_{i+1}(k) &= [h]_{\uparrow 2^i} * h_i(k), \\ g_{i+1}(k) &= [g]_{\uparrow 2^i} * h_i(k), \end{aligned} \tag{5.9}$$

where the subscript $[\cdot]_{\uparrow m}$ indicates the up-sampling by a factor of m and k is the equally sampled discrete time. The normalized wavelet and scale basis functions $\varphi_{i,l}(k)$, $\psi_{i,l}(k)$ can be defined as:

$$\begin{aligned} \varphi_{i,l}(k) &= 2^{i/2} h_i(k - 2^i l), \\ \psi_{i,l}(k) &= 2^{i/2} g_i(k - 2^i l), \end{aligned} \tag{5.10}$$

where the factor $2^{i/2}$ is an inner product normalization, i and l are the scale parameter and the translation parameter, respectively. The DWT decomposition can be described as

$$a_{(i)}(l) = x(k) * \varphi_{i,l}(k), \qquad d_{(i)}(l) = x(k) * \psi_{i,l}(k), \tag{5.11}$$

where $a_{(i)}(l)$ and $d_{(i)}(l)$ are the approximation coefficients and the detail coefficients at resolution i, respectively [136].

The concept of being able to decompose a signal totally and then perfectly reconstruct the signal again is practical, but it is not particularly useful by itself. In order to make use of this tool it is necessary to manipulate the wavelet coefficients to identify characteristics of the signal that were not apparent from the original time domain signal.

5.2 ECG Filtering

According to the review of preprocessing methods in Section 4.1, two approaches that have achieved good results in power line removal, baseline wander and *EMG* noise reduction, were selected to be applied in ECG filtering stage. First one corresponds to adaptive filtering and the second one corresponds to WT-based filtering.

5.2.1 Adaptive filtering

Elimination of sinusoidal interferences from a ECG signal is a typical procedure in biosignals filtering due to magnetic induction and displacement current from power line. Based on adaptive noise canceling approach the ASIC (Adaptive Sinusoidal Interference Canceler) [54] has been proposed for eliminating a single pure sinusoid assuming that its frequency is given. The idea is to use a synthetic tone which is a function of the explicit amplitude and phase measurements provided by an LMS-style algorithm to remove the interfering signal. Basically the algorithm is a follow-up of [54] that generalizes the ASIC so that it can be applied to situations when there are multiple interfering sinusoids such as harmonic noise cancellation in power line communications.

The ASIC generalized takes into account the presence of sinusoidal interferences with frequencies which are not exactly known, in this way, the received signal, $r(kT_s)$, is expressed as:

$$r(kT_s) = s(kT_s) + \sum_{i=1}^{M} a_i \cos((\omega_i + \Delta\omega_i)kT_s + \phi_i) \tag{5.12}$$

where $s(kT_s)$ is the stationary source signal, T_s is the sampling period, and $(\omega_i+\Delta\omega_i)$, a_i and ϕ_i, $i = 1, \ldots, M$, represent the frequencies, amplitudes and phases of the interfering sinusoids, respectively.

It is assumed that M and ω_i are known while $\Delta\omega_i$, a_i and ϕ_i are unknown constants with $\Delta\omega_i \ll 1$ for $i = 1, \ldots, M$. The task is to estimate $s(kT_s)$ from the corrupted signal $r(kT_s)$. For simplicity, the sampling time T_s is dropped in the following analysis.

By extending the ASIC to multiple sinusoidal interference cancellation, it is constructed the recovered signal, $\hat{s}(k)$, which has the form:

$$\hat{s}(k) = r(k) - \sum_{i=1}^{M} \hat{a}_i(k) \cos(\omega_i k + \hat{\phi}_i(k)) \tag{5.13}$$

where $\hat{a}_i(k)$ and $\hat{\phi}_i(k)$, $i = 1, \ldots, M$, are the amplitude and phase parameters, respectively. It is observed from (5.13) that $s(k)$ can be extracted perfectly when $\hat{a}_i(k) = a_i$ and $\hat{\phi}_i(k) = \phi_i + \Delta\omega_i k$. These desired values of $\hat{a}_i(k)$ can be acquired by minimizing

the mean square value of $\hat{s}(k)$, i.e. $\mathbf{E}\{\hat{s}^2(k)\}$, which is derived as,

$$\mathbf{E}\left\{\hat{s}^2(k)\right\} = \sigma_s^2 + \frac{1}{2}\sum_{i=1}^{M}(\hat{a}_i^2(k) + a_i^2) - \sum_{i=1}^{M}\hat{a}_i(k)a_i\cos(\hat{\phi}_i(k) - \phi_i - \Delta\omega_i k) \tag{5.14}$$

where σ_s^2 denotes the power of $s(k)$. In the generalized ASIC, $\hat{a}_i(k)$ and $\hat{\phi}_i(k)$ are adapted on a sample-by-sample basis to minimize $\mathbf{E}\{\hat{s}^2(k)\}$ according to the LMS-style algorithm, as is described in [138], obtaining the following recursive equations:

$$\begin{aligned}\hat{a}_i(k+1) &= \hat{a}_i(k) + \mu_{a_i}\hat{s}(k)\cos(\omega_i k + \hat{\phi}(k)) \\ \hat{\phi}_i(k+1) &= \hat{\phi}_i(k) + \mu_{\phi_i}\hat{s}(k)\sin(\omega_i k + \hat{\phi}(k))\end{aligned} \tag{5.15}$$

The quantities μ_{a_i} and μ_{ϕ_i}, $i = 1, \ldots, M$, are positive scalars that control convergence rate and ensure system stability of the algorithm. In [138], a procedure to the convergence of the parameters $\hat{a}_i$ and $\hat{\phi}_i$ is performed. In addition a discussion of the SNR improvement ratio regarding the received signal is performed.

5.2.2 WT-based filtering

According to the expression (5.11) that corresponds to approximation coefficients $a_{(i)}(l)$ and detail coefficients $d_{(i)}(l)$ of the DWT, it can be noted that the decomposition coefficients contain information about the frequency content and amplitude of the signal and noise. In this way is possible to analyze the decomposition in order to remove types of noise such as *EMG* or baseline wander by employing the detail and approximation coefficients, respectively, at specific scales. Due to the sources of noise are located at different frequency bands of the signal, this analysis is feasible.

High-frequency noise filtering

In the first case, it is considered the following model of a discrete noisy signal:

$$y(k) = f(k) + \sigma e(k), \quad k = 1, \ldots, N \tag{5.16}$$

where $y(k)$ represents noisy signal and $f(k)$ is unknown deterministic signal. It is assumed that e is Gaussian white noise with zero mean and unit variance, i.e. $N(\mu, \sigma^2) = N(0, 1)$.

The method for filtering out the white noise is a well-know method proposed by Donoho [55].

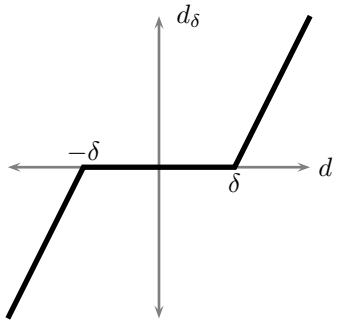


Figure 5.2: Soft-threshold function

The WT locates the most important spatial and frequency features of a regular signal in a limited number of wavelet coefficients. Moreover, the orthogonal transform of stationary white noise results in stationary white noise. This means that the expected noise energy is the same in all detail coefficients. If this energy is not too large, noise has a relatively small influence on the important large signal coefficients. These observations suggest that small coefficients should be replaced by zero, because they are dominated by noise and carry only a small amount of information.

The procedure that involves the operations over the coefficients is the thresholding. In this work, the Donoho's *soft-thresholding* or *shrinking* function is studied, as is shown in Figure 5.2.

Therefore the wavelet coefficients $d_i(l)$ between $-\delta$ and δ are set to zero, while the others are shrunk in absolute value. The threshold δ proposed by [55] is:

$$\sigma = \sqrt{2\log(N)\tilde{\sigma}} \tag{5.17}$$

where $\tilde{\sigma}$ is estimation of the noise variance σ^2 given by [55]:

$$\tilde{\sigma} = \frac{\text{median}|d_i(l)|}{0.6745}$$

Low-frequency noise filtering

Baseline wandering can make inspection of ECG signals difficult, because some features can be masked by this kind of noise. Moreover, in automatic inspection systems, other processing tasks such as wave detection, signal classification, among others, can be affected. It is, therefore, of importance to reduce as much as possible its effect. The

baseline wander removal, corresponds to the second case of WT-based filtering [139], taking into account the DWT approximation coefficients $a_i(l)$ of the expression (5.11).

In this case the model 5.16 is modified, by adding an interference $s(k)$:

$$y(k) = f(k) + \sigma e(k) + s(k), \quad k = 1, \ldots, N, \tag{5.18}$$

that represents the baseline wander. The goal of this method is to obtain an estimate of the interference, $\tilde{s}(k)$, which is subtracted from $y(k)$ in order to achieve a signal without low-frequency variations.

A way to accomplish that is to reconstruct only the approximation coefficients, by making zero the detail coefficients. This procedure is known as extreme thresholding [55]. Nevertheless, to obtain good results, the level of such approximation must be defined. Namely, the degree of accuracy of the approximation. Otherwise, there will be an over-fitting effect in the baseline approximation due to an overly low level or to the contrary, with a poor approximation due to an overly high level.

The best level depends on the amplitude and main spectrum distribution of the baseline interference. In [139], a method to automatically ascertain the best level is presented, which makes the process unsupervised. The method is based on measures of the resulting signal variance and on spectrum energy dispersion of the approximation.

In order to eliminate or reduce the baseline wandering, the approximation found must have a narrow spectrum, as such interferences are usually almost pure sinusoids. Besides, the variance of the resulting signal should be as low as possible, since the approximation must not have high frequency components such as peaks following R waves, and so, the final signal must be quite flat. Once the level is established, the wavelet approximation is calculated, and then, it is subtracted from the signal. Consequently, the baseline wander of this signal is greatly reduced. The whole process is carried out without user intervention, which represents an advantage compared with other more traditional methods.

5.3 QRS Complex Detection

In ECG signal processing, a remarkable stage to the identification of cardiac pathologies or HRV analysis corresponds to the detection of main waves and complexes of the ECG signal such as P wave, T wave and QRS complex.

In Chapter 4.3.2 some methods to estimate the fiducial points of ECG signal for its

posterior wave delineation are described. However, due to the proposed methodology requires only the QRS complex and HRV estimation, a specific method to estimate the R fiducial point is carried out, hence the QRS complex is calculated with a symmetric window around the R detected peak. Regarding the HRV, it is calculated between two consecutive R peaks.

Next, some requirements for a QRS detector and a specific procedure to estimate the R peak are presented.

5.3.1 Requirements for a general QRS detector algorithm

A QRS detector must be able to detect a large number of different QRS morphologies in order to be clinically useful and able to follow sudden or gradual changes of the prevailing QRS morphology. Furthemore, the detector must not lock onto certain types of rhythm, but treat the next possible event as if it could occur at almost any time after the most recently detected beat [1].

A QRS detector can, in general terms, be described by the block diagram presented in Figure 5.3 [140]. Within such detector structure, the purpose of the preprocessor is to enhance the QRS complexes while suppressing noise and artifacts; the preprocessor is usually implemented as a linear filter followed by a nonlinear transformation. The output of the preprocessor is then fed to a decision rule for R-peak detection. The purpose of each preprocessing block is summarized below.

1. *Linear filter.* It is designed to have bandpass characteristics such that the essential spectral content of the QRS complex is preserved, while unwanted ECG components such as the P and T waves are suppressed. The center frequency of the filter varies from 10 to 25 Hz and the bandwidth from 5 to 10 Hz. In contrast to other types of ECG filtering, waveform distortion is not a critical issue in QRS detection. The focus is instead on improving the SNR to achieve good detector performance [1].

2. *Nonlinear transformation.* Mainly, the transformation enhances the QRS complex in relation to the background noise as well as transforming each QRS complex into a single positive peak better suited for threshold detection. The transformation may consist of a memoryless operation, such as rectification or squaring of the bandpass filtered signal, or a more complex transformation with memory. Not all preprocessors employ nonlinear transformations, but the filtered signal is instead fed directly to the decision rule [1].

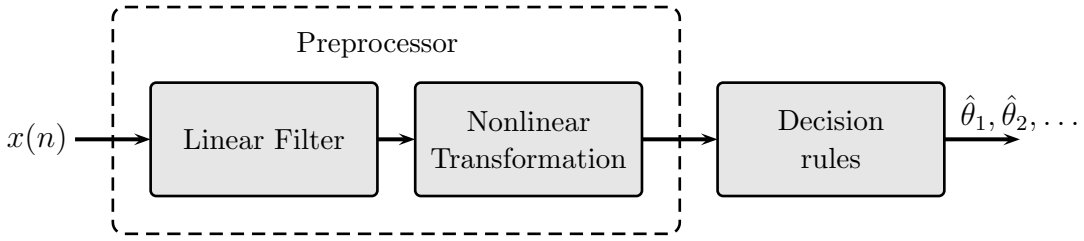


Figure 5.3: Block diagram of a commonly used QRS detector structure. The input is the ECG signal, and the output $\hat{\theta}_1, \hat{\theta}_2, \ldots$, is a series of occurrence times of the detected QRS complexes [1].

3. *Decision rules.* After the output of the preprocessor, a test is performed in order to determine whether a QRS complex is true or false. The decision rule can be implemented as a simple amplitud threshold procedure, but may also include additional test, for example, adaptive thresholds, to assure better immunity against different kind of noise [1] and different heartbeat morphologies.

Several detector-critical types of noise and artifacts exist depending on the ECG application of interest. The noise may be highly transient in nature or to be of a more persistent nature, as exemplified by the presence of powerline interference. In the case of an ECG recording with episodes containing excessive noise, it may be necessary to exclude such episodes from further analysis [1]. On the other hand, some applications such as Holter recordings require to analyze any type of lead, changing drastically the morphology between patients, hence T waves or Q waves can be higher than the R peaks. In this way, it is necessary to enhance the nonlinear transformation and threshold stages in order to avoid the increasing of negative (FN) and positive false (FP), where, a false negative (FN) occurs when the algorithm fails to detect a true beat quoted in the corresponding annotation file of the recording and a false positive (FP) represents a false beat detection.

A detector procedure that satisfies the requirements above described is as follows.

5.3.2 Hybrid algorithm

Basically, the R-peak detection algorithm uses a band-pass filter by means of the method described in [1], a high-pass filter based on a quadratic spline, described in [77], an adaptation of the nonlinear transformation developed in [2] applied over phonocardiographic signals for its segmentation and a stage of adaptive thresholding to process QRS complexes with low amplitude [95]. In Figure 5.4, it is shown the procedure of

R-peak detection.

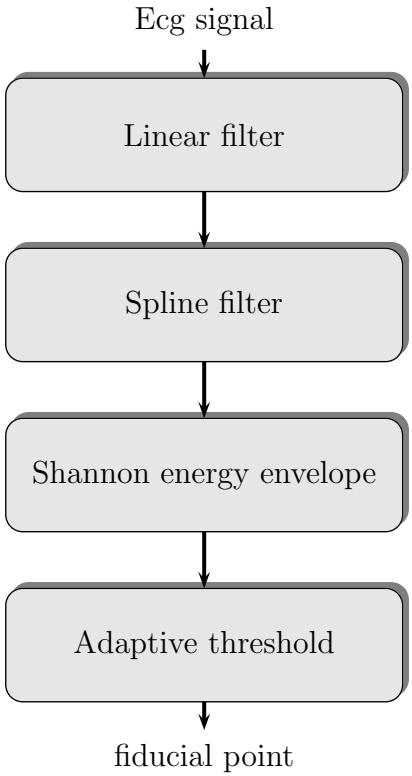


Figure 5.4: Block diagram of the hybrid algorithm

Linear filter

In this stage is applied a differentiation process in order to emphasize segments with rapid transients, such as the R peaks [141]. In discrete-time, differentiation can be approximated by a filter $H(z)$ that produces the difference between successive samples,

$$H(z) = 1 - z^{-1} \tag{5.19}$$

Such differencing filter may perhaps be an acceptable choice when analyzing resting ECGs; however, it accentuates high-frequency noise and is, therefore, inappropriate in situations with moderate or low SNRs [1].

In this case, appropriate results to combine differentiation with lowpass filtering in such way noise activity above a certain cut-off frequency $\omega_c = 2\pi f_c$ is attenuated [142].

The frequency response on the ideal *lowpass differentiator* is given by

$$H(\mathrm{e}^{\mathrm{j}\omega}) = \begin{cases} j\omega, & |\omega| \leq \omega_c; \\ 0, & \omega_c < |\omega| < \pi, \end{cases} \tag{5.20}$$

And the corresponding impulse response is

$$\begin{aligned} h(n) &= \frac{1}{2\pi} \int_{-\omega_c}^{\omega_c} j\omega \mathrm{e}^{\mathrm{j}\omega \mathrm{n}} \mathrm{d}\omega \\ &= \begin{cases} 0, & n = 0; \\ \dfrac{1}{\pi n} \left(\omega_c \cos(\omega_c n) - \dfrac{1}{n} \sin(\omega_c n) \right), & n \neq 0. \end{cases} \end{aligned} \tag{5.21}$$

Before the filter is used in practice, its infinite impulse response must be truncated using windowing or, better, by determining the coefficients of a FIR filter so that the error between its magnitude function and $H(\mathrm{e}^{\mathrm{j}\omega})$ in (5.20) is minimized in the MSE sense [142].

The large variability in signal and noise properties of the ECG implies that the requirements on frequency response have to be rather loose, and, as a result, simple structured filters can be applied. One family of such filters is defined by [143].

$$H(\mathrm{z}) = \left(1 - \mathrm{z}^{-\mathrm{L}_1}\right) \left(1 + \mathrm{z}^{-1}\right)^{\mathrm{L}_2}, \tag{5.22}$$

Where L_1, L_2 are two integer-valued parameters. The corresponding frequency response is given by

$$H(e^{j\omega}) = j2^{L_1+1} e^{j\omega(L_1+L_2)/2} \sin\left(\frac{\omega \mathrm{L}_1}{2}\right) \cos^{\mathrm{L}_2}\left(\frac{\omega}{2}\right). \tag{5.23}$$

The first part, $\left(1 - \mathrm{z}^{-\mathrm{L}_1}\right)$, forms the difference between the input signal and the delayed input, whereas the second part, $(1 + \mathrm{z}^{-1})^{L_2}$, is a lowpass filter whose bandwidth decreases as L_2 increases. Filters belonging to the family in (5.22) can be implemented without multipliers, thus only requiring addition and subtraction. Consequently, these filters are attractive for systems which analyze long-term ECG recordings [1]. The filter $(L_1, L_2) = (5, 4)$ may be a suitable choice for a higher sampling rate of 250 *Hz*, resulting in a filter with a center frequency of 20 *Hz* [144]. Figure 5.5 shows some combinations of the parameters. Other values of both parameters are discussed in [1].

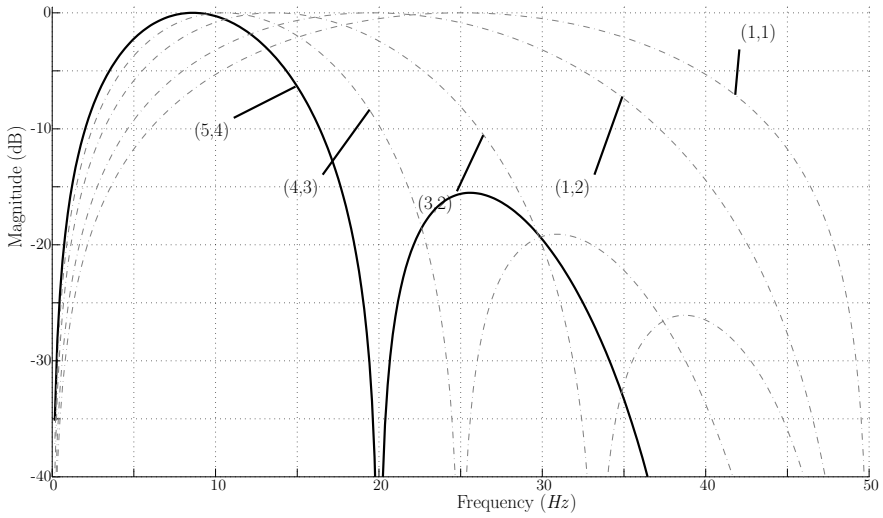


Figure 5.5: The magnitude function of filter in (5.22), defined by the two integer parameters L_1 and L_2, displayed for the combinations (1,1),(1,2),(3,2),(4,3) and (5,4). The last one is used throughout this work. Each magnitude function has been normalized so that its maximum gain corresponds to 0dB. The sampling rate is assumed to be 100 *Hz*.

Nonlinear transformation

This stage has two procedures. At first, a filter to enhance the QRS presence is introduced. Secondly, a nonlinear transformation that attenuates the effect of low value noise and makes the low amplitude waves easier to be found.

Spline Filter:

The filter is based on a quadratic spline wavelet with compact support and one vanishing moment. The basic wavelet ($\psi(x)$) is a first derivative of a smooth function, which Discrete Fourier transform is,

$$\breve{\Psi}(\omega) = i\omega \left(\frac{\sin(\frac{\omega}{4})}{\frac{\omega}{4}} \right)^4 . \tag{5.24}$$

By considering a dilation of a basic wavelet $\psi(x)$ by the scale factor s and taking s as powers of two, $s = 2^j, \quad j \in \mathbb{Z}$, a specific version of the wavelet transform (WT) of the signal $f(x)$ results, termed dyadic WT. The dyadic WT of a digital

signal $f(n)$ can be calculated with Mallat algorithm [59] as follows:

$$V_{2^j} f(n) = \sum_{k\in\mathbb{Z}} h_k V_{2^{j-1}} f(n - 2^{j-1}k) \tag{5.25}$$

$$W_{2^j} f(n) = \sum_{k\in\mathbb{Z}} g_k V_{2^{j-1}} f(n - 2^{j-1}k), \tag{5.26}$$

where, V_{2^j}, is a smoothing operator and $V_{2^0} f(n) = d_n$, being d_n the digital signal to be analyzed, which is the output of linear filter described in the previous section. $w_{2^j} f(n)$ is the WT of digital signal. h and g are coefficients of a lowpass filter $H(\omega)$ and a highpass filter $G(\omega)$, respectively; that means

$$H(\omega) = \sum_{k\in\mathbb{Z}} h_k \mathrm{e}^{-ik\omega}, \quad G(\omega) = \sum_{k\in\mathbb{Z}} g_k \mathrm{e}^{-ik\omega} \tag{5.27}$$

By calculating the filters (5.27) using the quadratic spline wavelet, the following expressions are obtained,

$$H(\omega) = \exp^{i\omega/2} \left(\cos\frac{\omega}{2}\right)^3 \tag{5.28}$$

$$G(\omega) = 4i\exp^{i\omega/2} \left(\sin\frac{\omega}{2}\right). \tag{5.29}$$

The discrete Fourier transform of WT using (5.28) and (5.29) is

$$\check{\Psi}(\omega) = \begin{cases} G(\omega)\check{f}(\omega)\check{\phi}(\omega) & j = 1 \\ G(2\omega)H(\omega)\check{f}(\omega)\check{\phi}(\omega) & j = 2 \\ G(2^{j-1}\omega)H(2^{j-1}\omega)\ldots H(\omega)\check{f}(\omega)\check{\phi}(\omega) & j > 2 \end{cases} \tag{5.30}$$

where ϕ is a smooth function, and $\check{f}(\omega)\check{\phi}(\omega)$ is the discrete Fourier transform of the input signal. From (5.30), the WT of $f(n)$ at scale 2^j is equal to filtered signal of d_n that passed through a digital bandpass filter.

By defining $Q^j(\omega)$ as the transform function of the equivalent filter, it is possible

rewrite (5.30) only in terms of $G(\omega)$ and $H(\omega)$, as follows:

$$Q^j(\omega) = \begin{cases} G(\omega) & j = 1 \\ G(2\omega)H(\omega) & j = 2 \\ G(2^{j-1}\omega)H(2^{j-1}\omega)\dots H(\omega) & j > 2 \end{cases} \tag{5.31}$$

From (5.28), (5.29) and (5.31), the following expression is deduced,

$$Q^j(\omega) = \frac{2}{8^{j-1}} \sum_{k=1-2^{j-1}}^{2^j+2^{j-1}-2} q_k^j \mathrm{e}^{ik\omega} \tag{5.32}$$

where $q^j_{1-2^{j-1}+k} = -q^j_{2^j+2^{j-1}-2-k} \neq 0$, with $k \in [1-2^{j-1}, 2^j+2^{j-1}-2]$.

The filter $Q^j(\omega)$ corresponds to FIR digital filter with generalized linear phase. The filter is antisymmetric and the delay time of this central point is $\frac{2^j-1}{2}$.

The equivalent filters of the WTs using the first derivative of a smooth function, have bandwidths approximating those of the quadratic spline wavelet, so the results of ECG detection with these wavelets are almost the same as those with quadratic spline wavelet, therefore more time is required to calculate their WT's.

The normalized average Shannon energy:

The nonlinear stage is based on the envelope of the previously filtered signal, calculated using the normalized average Shannon energy, which attenuates the effect of low value noise enhancing the QRS complexes with low amplitude.

Figure 5.6 shows different methods to calculate the envelope of the normalized signal. Because of the symmetry of the results, as we can see from the following definitions, only the positive part is shown here. The figure is drawn based on the following definitions, where $\boldsymbol{x}_n$ is the normalized signal regarding its amplitude, which has the real value from -1 to 1, i.e. $\boldsymbol{x}_n = \boldsymbol{x}/\max(|\boldsymbol{x}|)$.

- Shannon energy: $\boldsymbol{E}_{se} = -\boldsymbol{x}_n^2 \cdot \log(\boldsymbol{x}_n)^2$
- Shannon entropy: $\boldsymbol{E}_{st} = -|\boldsymbol{x}| \cdot \log|\boldsymbol{x}_n|$
- Absolute value: $\boldsymbol{E}_{\boldsymbol{ab}} = |\boldsymbol{x}_n|$
- Energy: $\boldsymbol{E}_{sq} = \boldsymbol{x}_n^2$

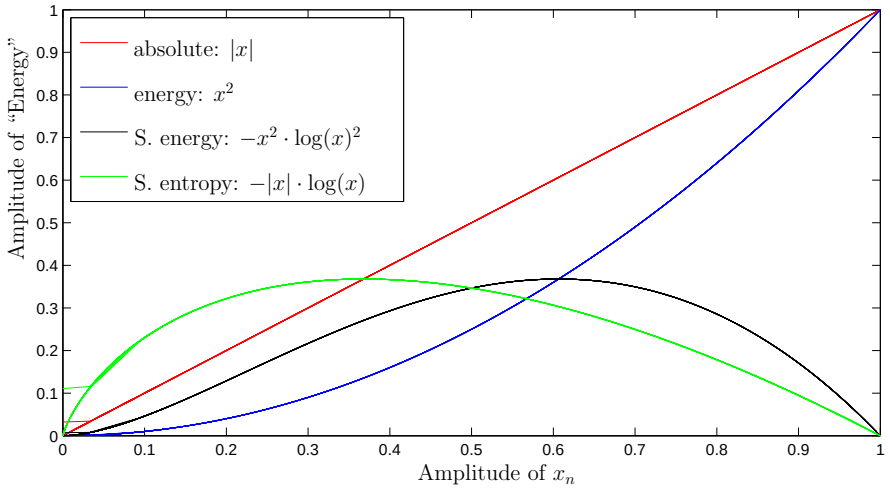


Figure 5.6: The comparison of different envelope methods [2].

The Figure 5.6 indicates that the energy (square) will reduce the low amplitude complexes under the high amplitude ones by enlarging the high/low amplitude ratio. The Shannon entropy accentuates the effect of low value noise that makes the envelope too noisy to read. The absolute value gives the same weight to all the signal. The Shannon energy emphasizes the medium intensity signal and attenuates the effect of low intensity signal much more than that of high intensity signal. So, the last one is better than the absolute value in shortening the difference of the envelope intensity between the low amplitude complexes and the high amplitude complexes. This shortening makes the finding of low amplitude complexes easier.

In this line, the Shannon energy E_{se} is calculated over the normalized signal x_n defined previously, in continuous segments of length l_s with l_{ov}-s. segment overlapping. The average Shannon energy is calculated as:

$$\boldsymbol{E}_{se} = -\frac{1}{l_s}\sum_{i=1}^{l_s} x(i)_n^2 \cdot \log(x(i)_n^2) \tag{5.33}$$

where, l_s is signal length. Then the normalized average Shannon energy versus time axis is computed. The normalized average Shannon energy is computed as

follows,

$$P(t) = \frac{E_{se}(t) - \mu(E_{se}(t))}{\sigma(E_{se}(t))} \tag{5.34}$$

Decision rules

The envelope resulting from the nonlinear transformation is used in order to determine the location of the R peaks. This phase is based on the works [65] and [95]. It is established the parameter A_j^{m+1}, which represents the maximum value of the envelope at a specific interval. The computation of a detection threshold for the next heartbeats, is accomplished with such parameter, under the following conditions:

If the maximum value of the envelope $max(E_{sn}) > 2A_j^m$, then:

$$A_j^{m+1} = A_j^m \tag{5.35}$$

otherwise:

$$A_j^{m+1} = \left(\frac{7}{8}\right) A_j^m + \left(\frac{1}{8}\right) |max(E_{se})| \tag{5.36}$$

The initial threshold is taken as a fraction of the maximum value calculated, i.e. $Th = bA_j, 0 \leq b \leq 1$.

In this way, the highest envelope peaks regarding peaks with lower amplitude, do not affect in the posterior analysis.

While R-peaks are detected, a statistical measure is applied, which assesses the distance between the last n R-peaks, denoted as R_{AV1}, and a statistical index for last m R intervals in the range of $0.8R_{AV1} < R_{peak} < 1.3R_{AV1}$, named R_{AV2}. When the distance computed for the last R-peak exceeds the time threshold ($1.6R_{AV2}$), a backward search is developing where the amplitude threshold is reduced by half: $Th = Th/2$. Thereby, it is posible to identify a R-peak at 20 samples before found maximum value $max(E_{se})$ in order to compensate the filter delay. Additionally, in order to avoid false detections because of artifacts present in signal, a refractory period set to be $200ms$.

5.4 Feature Estimation Methods

In the feature extraction stage, numerous different methods can be used so that several diverse features can be extracted from the same raw data. In this section, the Wavelet Transform (WT) which can be applied as feature extractor and the Hermite parametric model, are described.

5.4.1 WT-based characterization

The WT provides very general techniques which can be applied to many tasks in signal processing. Wavelets are ideally suited for the analysis of sudden short-duration signal changes. One very important application is the ability to compute and manipulate data in compressed parameters which are often called features [136]. Thus, the time-varying biomedical signal, consisting of many data points, can be compressed into a few parameters by the usage of the WT. These parameters characterize the behavior of the time-varying biomedical signal. This feature of using a smaller number of parameters to represent the time-varying biomedical signal is particularly important for recognition and diagnostic purposes [136].

In the present study, feature extraction from the ECG signals is performed by usage of the DWT as is discussed in the Section 5.4.1. The computed wavelet coefficients can be used as the representing features of the ECG signals. These features can be used as inputs of classification models such as supervised or non-supervised approaches.

5.4.2 Hermite based characterization for QRS complex using Hermite parametric model

This section describes a methodology to reconstruct and characterize the QRS complex using the Hermite parametric model. Complexes are extracted using the R-peak location and considering a fixed window length. Reconstruction is carried out by applying the optimal value of scale parameter obtained by means of the minimization of dissimilarity between original and reconstructed signal. DTW is used as dissimilarity measure. In addition, it is also described a method to determine the minimum number of coefficients that generate a highly-approximate reconstruction based on the comparison of frequency spectrum in the range of $1 - 20$ Hz. Then, the Hermite model and the proposed methodology to characterize QRS complexes, are described in detail.

Hermite parametric model

Hermite polynomial H_n of order n is a feasible solution of the following differential equation:

$$\varphi''(z) - 2z\varphi'(z) + 2n\varphi(z) = 0$$

where n is a non-negative integer. Thereby, Hermite polynomials can be defined as follows:

$$H_n(z) = (-1)^n e^{z^2} \frac{d^n}{dz^n} e^{-z^2} \tag{5.37}$$

Hermite polynomials represent an orthonormal set with respect to the weight function e^{-z^2} , i.e,

$$\frac{1}{\sqrt{2^n n! \sqrt{\pi}}} \langle e^{-z^2} H_n(z), H_m(z) \rangle = \delta_{m,n},$$

where $\delta_{m,n}$ is the delta function of Kronecker ($\delta_{m,n} = 1$ if $m = n$, otherwise $\delta_{m,n} = 0$) and $\langle \cdot, \cdot \rangle$ denotes inner product.

By letting $z = \dfrac{t}{\sigma}$, it is posible to establish a base of the form:

$$\phi_n^\sigma(t) = \frac{e^{-t^2}/2\sigma^2}{\sqrt{2^n \sigma n! \sqrt{\pi}}} H_n(t/\sigma) \tag{5.38}$$

where σ is a scale parameter (see figure 5.8). The expression (5.38) is known as Hermite parametric model.

Then, Hermite coefficients for signal $s(t)$ are given by:

$$C_n^\sigma = \frac{1}{Fs} \int_{t=-\infty}^{\infty} s(t) \phi_n^\sigma(t) dt \tag{5.39}$$

Finally, signal reconstruction can be written as:

$$s(t) = \sum_{n=0}^{\infty} C_n^\sigma \phi_n^\sigma(t) \tag{5.40}$$

Signal reconstruction

In practice, for signal reconstruction is used a recursive equation to compute the Hermite polynomials, as follows:

$$H_n(z) = 2zH_{n-1}(z) - 2(n-1)H_{n-2}(z) \tag{5.41}$$

being $H_0 = 1$ and $H_1 = 2z$.

The elements of Hermite base are ranged in the interval $(-t_0, t_0)$ where the value of t_0 is chosen according to the nature of signals. The number of elements and the signal length must be adjusted by applying a time vector of the form:

$$t = -t_0 : \tfrac{2t_0}{L_{QRS}-1} : t_0,$$

where L_{QRS} is the QRS length. In figure 5.7(c) examples of elements of Hermite base are shown.

Scale parameter σ is added to the considered window which can be adjusted to the QRS width, as can be seen in figure 5.8.

For easiness in implementation, Hermite coefficients can be computed through the discrete form of (5.39), assuming that elements out of the interval $(-t_0, t_0)$ are zero:

$$C_n^\sigma = \frac{1}{Fs}\sum_{i=-t_0}^{t_0} s(i).\phi_n^\sigma(i) = \frac{1}{Fs}\langle \boldsymbol{s}, \boldsymbol{\phi}_n^\sigma \rangle \tag{5.42}$$

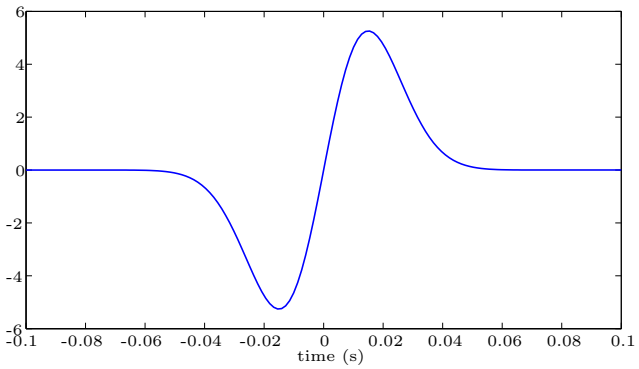
Given this, reconstruction can be accomplished with

$$s(t) = \sum_{n=0}^{N-1} C_n^\sigma \phi_n^\sigma(t) + \xi(t) = \hat{s}_N^\sigma(t) + \xi(t) \tag{5.43}$$

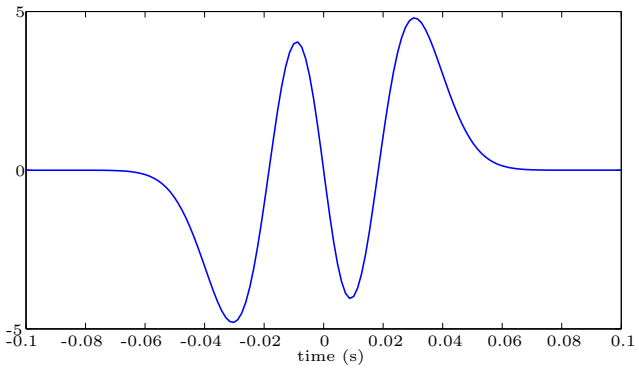
where $\hat{s}_N^\sigma(t)$ is the truncated reconstructed signal using the first N elements and $\xi(t)$ is a truncating factor. Discrete signal $\hat{\boldsymbol{s}}_N^\sigma$ must be centered and normalized with respect amplitude.

Comparison among frequency spectra

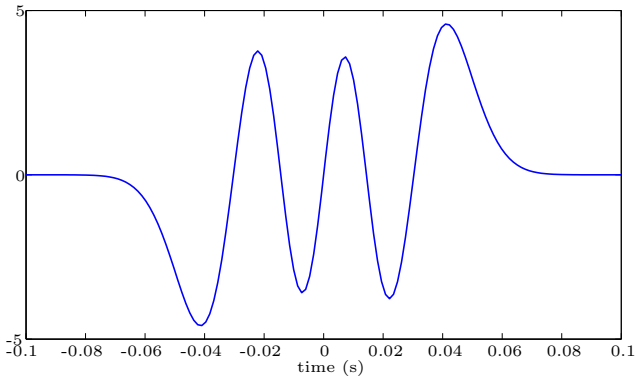
Here, the change of the reconstructed signal in comparison with the original signal is analyzed, considering different values for N ($N \in [3, 20]$). To that aim, the spectrum of original signal is compared with the reconstructed signal in the range of $1 - 20$ Hz



(a) $N = 1$, $\sigma = 0.015$, $t_0 = 200ms$



(b) $N = 3$, $\sigma = 0.015$, $t_0 = 200ms$



(c) $N = 5$, $\sigma = 0.015$, $t_0 = 200ms$

Figure 5.7: Hermite base using different values of N with σ and t_0, constants

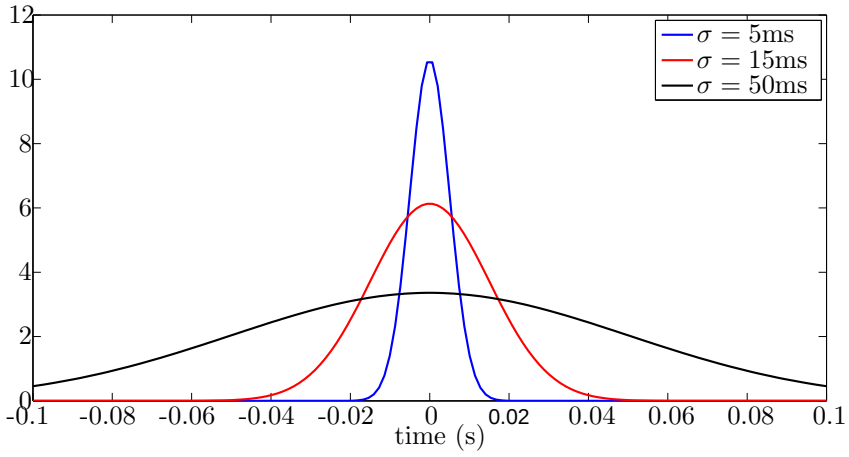


Figure 5.8: First element of Hermite base ($N = 0$) calculated for different values of σ

in order to determine the proper value of N, i.e., the minimum number of elements (N_{min}) that generate a reconstruction with the least loss of spectral information in comparison with the reconstruction results. Then, an advisable value of σ is chosen.

Power spectral density is estimated employing periodogram [145]:

$$S(e^{j\omega}) = \frac{1}{n}||\sum_{l=1}^{n} s_l e^{j\omega l}||^2 \tag{5.44}$$

The spectral difference is computed as follows:

$$diff_N = \frac{1}{F}\sum_{f=2}^{F}|S_f(\boldsymbol{s}) - S_f(\hat{\boldsymbol{s}}_N^{\sigma})| \tag{5.45}$$

where $F = 20\ Hz$ y $N \in (3, 20)$.

Optimal value of σ

The optimal value of scale parameter (σ_{opt}) is obtained by applying a dissimilarity measure between the spectrum of original signal and its reconstruction. In this case, the basic method of DTW is recommended and is implemented as described in [64], without using global constraints.

Finally, by considering that a value of σ less than 5 ms or major than 100 ms is not required for QRS complex reconstruction, the optimization problem to obtain σ_{opt} can

be written as:

$$\begin{aligned} &\min_{\sigma} \ \mathrm{dtw}(\boldsymbol{s}_i, \hat{\boldsymbol{s}}^{\sigma}_{N_{min}}) \\ &\text{s.t.} \ \ \sigma \in (5, 100)\text{ms} \end{aligned} \tag{5.46}$$

Resultant feature set using Hermite model

This methodology provides the following feature set:

- QRS energy:
  Because of morphology of ventricular arrhythmias, energy is a proper feature to be considered:

$$E(\boldsymbol{s}) = \sum_{i=1}^{L_{QRS}} s_i^2$$

- σ_{opt}
- C_{σ}^{n} with $n = 6$.
- Difference between $\boldsymbol{s}_i$ and QRS complex template $\boldsymbol{s}_{temp}$ applying (5.45), where:

$$\boldsymbol{s}_{temp} = \mu(\boldsymbol{s}_i) \ \forall i,$$

With this methodology, the establishment of a minimum number of elements in the signal reconstruction process, allows to reduce the search space of optimal scale parameter σ_{opt} for Hermite model by minimizing the dissimilarity of spectra between the reconstructed and original signals.

Chapter 6

Analysis of Relevance

6.1 Introduction

In pattern recognition context, to collect descriptive patterns of samples or observations, in many problems there is no prior information about the number or relevance for classification of such patterns. Then, characterization process can yield high-dimensional data matices and therefore can be a problem for the following classification stage in terms of both performance due to redundant information that could be considered and processing time because of initial data dimension. These issues are studied by feature selection methods. In general, feature selection aims to reduce the dimensionality of pattern for classificatory analysis by selecting the most informative rather than irrelevant and/or redundant features [146].

In the context of this work, it must be remarked that during Holter monitoring there is a huge amount of information stored, then classification of heartbeats usually becomes very time–consuming and hence any automated processing of the ECG assisting this process would be of benefit, particularly, a feature selection procedure might be considered. In this connection, and based on multivariate representation of input data, a direct approach is the use of linear decomposition methods to decrease the dimensionality of the feature space, resulting from heartbeat characterization. Among linear decomposition methods, PCA and its variations have shown to be a good alternative for this aim [147]. Moreover, the non–parametric nature, feasibility of implementation and versatility are some advantages of PCA. Nonetheless, Holter monitoring of cardiac arrhythmias is an application where the conventional PCA might be not recommended

because it gives the same importance to all observations, being sensitive to the presence of outliers and noise in the data. In fact, because of strong asymmetry among class observations, it have been remarked that the heartbeat features must be properly selected to provide convenient separability among heartbeat types [9, 148]. To that end, in this chapter, a weighted version of PCA (termed WPCA) is studied, where introduced weights are given on dependence on variable–wise relevance criteria, making possible to assess the relative importance of each feature (variable) immersed on the original data representation by using a kind of weighting factor. This work takes advantage of the following two linear projection methods to estimate the feature–wise weighting factor. Namely, MSE (Section 6.2) and M–inner product (Section 6.3) are studied. The last one leads in an algorithm known as $Q-\alpha$ [3]. The following sections describe the convergence of $Q-\alpha$ algorithm (Section 6.4), sparsity and positivity of weighting factor (Section B) and a variant of the algorithm without using parameter tuning (Section 6.6). At the end, in Section 6.7, the linear projection of weighted data is described.

Notation

Given a set of p-dimensional vector data, $\{\boldsymbol{x}_i\}$, being centered, i.e., $\mathbf{E}\{\boldsymbol{x}_i\} = \mathbf{0}, \forall i$, where all n training observations can be aligned in the input matrix $\boldsymbol{X} = [\boldsymbol{x}_1 \mid \cdots \mid \boldsymbol{x}_n]^\mathsf{T} \in \mathbb{R}^{n\times p}$, then the respective linear projection is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{V}$, $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$. Generally, the ortonormal projection is performed to a q–dimensional space ($q < p$), being $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ an orthogonal matrix, where the representation quality of $\boldsymbol{X}$ is measured by using a given error function ε between the original data and the truncated orthonormal projection $\widehat{\boldsymbol{V}} \in \mathbb{R}^{p\times q}$, which can be expressed as a distance measure: $\varepsilon = d(\boldsymbol{X}, \widehat{\boldsymbol{X}})$, where $\widehat{\boldsymbol{X}} = \widehat{\boldsymbol{Y}}\widehat{\boldsymbol{V}}^\mathsf{T}$, being $\widehat{\boldsymbol{X}} \in \mathbb{R}^{n\times p}$ the truncated input matrix. There exist several alternatives for calculating this distance, such as, the Minkowski distance (L_p metrics), square Euclidean distance, angle–based distance, Mahalanobis, among others, as discussed in [147]. Commonly, analysis of relevance methods aim to minimize ε.

By denoting $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{W}$ as the weighted data matrix, likewise, a set of their q most relevant eigenvalues can be estimated, the weighted relevance (weighting covariance) matrix is introduced as follows [149]:

$$\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}} = \widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}} = \boldsymbol{W}^\mathsf{T}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{W}, \ \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}} \in \mathbb{R}^{p\times p} \tag{6.1}$$

where $\boldsymbol{W} \in \mathbb{R}^{p \times p}$ is a diagonal weighting matrix.

6.2 MSE–based Approach

The main goal of conventional PCA is to find out the optimum transform for a given data set in least square terms, being the simplest eigenvector–based multivariate analysis, where the linear decomposition of matrix $\boldsymbol{X}$ by singular value decomposition takes place,

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}_{\boldsymbol{X}}\boldsymbol{V}^{\mathsf{T}} = \sum_{i=1}^{p} \mu_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}. \tag{6.2}$$

Vector $\boldsymbol{\mu} = [\mu_1, ..., \mu_p]$ is the singular value vector, matrix $\boldsymbol{\Lambda}_{\boldsymbol{X}} = \text{diag}(\boldsymbol{\mu})$ is a diagonal matrix formed by singular values, $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ corresponds to eigenvectors of $\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}$ and $\boldsymbol{V}$ holds eigenvectors of $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}$ if $\boldsymbol{W} = \text{diag}(\mathbf{1}_p)$, where $\mathbf{1}_p$ is a p-dimensional all–ones vector.

Therefore, the minimum square error (MSE) distance is achieved to assess the representation quality, which yields to the following minimization problem:

$$\min_{\widehat{\boldsymbol{V}}}\{\varepsilon\} = \mathbf{E}\left\{\min\{(\boldsymbol{X} - \widehat{\boldsymbol{Y}}\widehat{\boldsymbol{V}}^{\mathsf{T}})^{\mathsf{T}}(\boldsymbol{X} - \widehat{\boldsymbol{Y}}\widehat{\boldsymbol{V}}^{\mathsf{T}})\}\right\} \tag{6.3}$$

Let, $\boldsymbol{x}^{(l)} \in \mathbb{R}^{n \times 1}$, $l = 1, \ldots, p$, the l-th feature of the input matrix, $\boldsymbol{X}$ that can be approximated by a truncated projection into a q-dimensional ortonormal space by the following linear combination:

$$\widehat{\boldsymbol{x}}^{(l)} = \sum_{i=1}^{q} c_i^{(l)} \boldsymbol{u}_i \tag{6.4}$$

then, the MSE value between the original and the reconstructed features is estimated as,

$$\overline{e^2} = \mathbf{E}\left\{(\boldsymbol{x}^{(l)} - \widehat{\boldsymbol{x}}^{(l)})^{\mathsf{T}}(\boldsymbol{x}^{(l)} - \widehat{\boldsymbol{x}}^{(l)})\right\} = \mathbf{E}\left\{(\sum_{i=q+1}^{p} c_i^{(l)}\boldsymbol{u}_i)^{\mathsf{T}}(\sum_{i=q+1}^{p} c_i^{(l)}\boldsymbol{u}_i)\right\} \tag{6.5}$$

that can be minimized if maximizing its complement, and therefore the final expression to be maximized is:

$$\mathbf{E}\left\{(\sum_{i=1}^{q} c_i^{(l)}\boldsymbol{u}_i)^{\mathsf{T}}(\sum_{i=1}^{q} c_i^{(l)}\boldsymbol{u}_i)\right\} = \mathbf{E}\left\{\sum_{i=1}^{q} \left(c_i^{(l)}\right)^2\right\} \tag{6.6}$$

From the expression given by (6.2), the truncated data representation can be written as $\widehat{\boldsymbol{X}} = \sum_{i=1}^{q} \mu_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}$. Then, it can be deduced that $\widehat{\boldsymbol{x}}^{(l)} = \sum_{i=1}^{q} \mu_i v_i^{(l)} \boldsymbol{u}_i$ and therefore the coefficients for linear combination are obtained with

$$c_i^{(l)} = \mu_i v_i^{(l)} \tag{6.7}$$

where $v_i^{(l)}$ represents the l-th element of the i-th column vector of matrix $\boldsymbol{V}$. By replacing $c_i^{(l)}$, the expression (6.6) is rewritten as follows:

$$\mathbf{E}\left\{\sum_{i=1}^{q} \left(c_i^{(l)}\right)^2\right\} = \mathbf{E}\left\{\sum_{i=1}^{q} (\mu_i)^2 (v_i^{(l)})^2\right\} = \mathbf{E}\left\{\sum_{i=1}^{q} \lambda_i (v_i^{(l)})^2\right\} \tag{6.8}$$

where $\lambda = [\lambda_1, \ldots, \lambda_p]$ is the vector of the eigenvalues of $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}$.

Definition 6.2.1. A relevance measure based on MSE approach. *By generalizing the expression (6.8) for the p features, and taking as estimation of expectance operator the simple average, then the following relevance measure is assumed:*

$$\boldsymbol{\rho} = \frac{1}{q} \sum_{i=1}^{q} \lambda_i \boldsymbol{\nu}_i, \tag{6.9}$$

where $\boldsymbol{\nu}_i$ is a vector compounded by the square of each one of the elements of $\boldsymbol{v}_i$.

It should be remarked that vector $\boldsymbol{\rho}$ yields a relevance index, which measures the accumulated variance of eigenvalues and eigenvectors, and is used as weighting factor. Then, accordingly to the quadratic form of the generalized covariance matrix (see (6.1)), the weighting matrix can be obtained as $\boldsymbol{W} = \text{diag}(\sqrt{\boldsymbol{\rho}})$.

In the end, the commonly known criterion of variance explained is used to find q, which rejects the elements that do not significantly contribute to the accumulated variance of data set. In addition, since the first principal component holds most of explained variance, the particular case $q = 1$ is also considered throughout this work.

6.3 M-inner Product Approach

This case recalls the M-inner product as error measure between the original variable and its orthonormal projection. Let $\boldsymbol{U}_p \in \mathbb{R}^{p \times p}$ be an arbitrary orthonormal matrix, and $\widehat{\boldsymbol{x}}^{(l)} = \boldsymbol{u}_p^{(l)^{\mathsf{T}}} \boldsymbol{X}$ the linear combination to estimate the l-th feature. Then, the error

measure for each feature is given by:

$$d_{\boldsymbol{A}}(\boldsymbol{x}^{(l)}, \widehat{\boldsymbol{x}}^{(l)}) = \langle \boldsymbol{x}^{(l)}, \widehat{\boldsymbol{x}}^{(l)} \rangle_{\boldsymbol{A}} = \boldsymbol{x}^{(l)\mathsf{T}} \boldsymbol{A} \widehat{\boldsymbol{x}}^{(l)} \tag{6.10}$$

where $\langle \cdot, \cdot \rangle_{\boldsymbol{A}}$ is the M-inner product regarding to the symmetric positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, which relates to the outer product between variables, i.e.,

$$\boldsymbol{A} = \sum_{l=1}^{p} \boldsymbol{x}^{(l)} \boldsymbol{x}^{(l)\mathsf{T}} = \boldsymbol{X}\boldsymbol{X}^{\mathsf{T}} \tag{6.11}$$

The previous expression, in terms of spectral analysis based on graphs, represents the affinity matrix because shows the relation of data points (nodes) each other [3]. In this case, it corresponds to trivial affinity matrix because it only uses the inner product between observations.

Next, if definition for the l-th estimated feature $\boldsymbol{x}^{(l)}$, given by (6.4), is replaced in (6.10), the following expression holds:

$$(\boldsymbol{x}^{(l)} - \widehat{\boldsymbol{x}}^{(l)})^{\mathsf{T}} \boldsymbol{A} (\boldsymbol{x}^{(l)} - \widehat{\boldsymbol{x}}^{(l)}) = \left(\sum_{i=q+1}^{p} c_i^{(l)} \boldsymbol{u}_i \right)^{\mathsf{T}} \boldsymbol{A} \left(\sum_{i=q+1}^{p} c_i^{(l)} \boldsymbol{u}_i \right) \tag{6.12}$$

that can be minimized if maximizing its complement, i.e, $\widehat{\boldsymbol{x}}^{(l)\mathsf{T}} \boldsymbol{A} \widehat{\boldsymbol{x}}^{(l)}$. Thus, by replacing $\boldsymbol{A}$ and $\widehat{\boldsymbol{x}}^{(l)}$ (6.4), and generalizing for all variables, the following expression to be maximized yields:

$$\operatorname{tr}(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{A} \boldsymbol{X}) = \operatorname{tr}(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{X} \boldsymbol{X}^{\mathsf{T}} \boldsymbol{X}) = \sum_{i=1}^{q} \lambda_i^2 \tag{6.13}$$

where λ are the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}$.

Furthermore, since the eigenvalues of $\widehat{\boldsymbol{X}}^{\mathsf{T}} \widehat{\boldsymbol{X}}$ matrix are the first p eigenvalues of $\widehat{\boldsymbol{X}} \widehat{\boldsymbol{X}}^{\mathsf{T}}$, then, to maximize (6.13) is equivalent to maximize the expression:

$$\operatorname{tr}(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}} \boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}) = \operatorname{tr}(\boldsymbol{A}\boldsymbol{A}) \approx \sum_{i=1}^{q} \lambda_i^2 \tag{6.14}$$

Definition 6.3.1. Relevant Features Optimization. *By establishing a weighted relevance matrix as*

$$\boldsymbol{A}_{\boldsymbol{\alpha}} = \sum_{l=1}^{p} \alpha_l \boldsymbol{x}^{(l)\,T} \boldsymbol{x}^{(l)} = \boldsymbol{X}\boldsymbol{W}\boldsymbol{W}\boldsymbol{X}^{T} \tag{6.15}$$

where $\boldsymbol{W} = \operatorname{diag}(\sqrt{\boldsymbol{\alpha}})$ *and* $\boldsymbol{\alpha} \in \mathbb{R}^{p \times 1}$ *is a weighting vector, and assuming the orthonor-*

mal invariance criterion [150], the optimization problem can be rewritten as:

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{Q}} \; & tr(\boldsymbol{Q}^T\boldsymbol{A}_\alpha\boldsymbol{A}_\alpha\boldsymbol{Q}) = \sum_{i=1}^{q}\lambda_i^2 \\ s.t. \quad & \boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1, \quad \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I} \end{aligned} \tag{6.16}$$

being matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ *an arbitrary orthonormal matrix.*

Expression given by equation 6.15 represents the weighted affinity matrix. Besides, the weighting vector is adjusted to be $\sqrt{\boldsymbol{\alpha}}$ to make the optimization problem in hand to be bilinear regarding $\boldsymbol{\alpha}$, thus, $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\text{diag}(\sqrt{\boldsymbol{\alpha}})$. The weight vector α and the orthonormal matrix $\boldsymbol{Q}$ are determined at the maximal point of the optimization problem.

By assuming a new matrix $\boldsymbol{M} = \boldsymbol{X}^\mathsf{T} = [\boldsymbol{m}_1^\mathsf{T},\ldots,\boldsymbol{m}_p^\mathsf{T}]^\mathsf{T}$ it can be deduced that $\text{tr}(\boldsymbol{A}) = \sum_{l=1}^{p}\boldsymbol{x}^{(l)}\boldsymbol{x}^{(l)\mathsf{T}}$ from (6.11). Likewise, from (6.15) it can be inferred that $\text{tr}(\boldsymbol{A}_{\boldsymbol{\alpha}}) = \sum_{i=1}^{p}\alpha_i\boldsymbol{m}_i\boldsymbol{m}_i^\mathsf{T}$. Then, $\text{tr}(\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{A}_{\boldsymbol{\alpha}}) = \sum_{i=1}^{p}\sum_{j=1}^{p}\alpha_i\alpha_j(\boldsymbol{m}_i^\mathsf{T}\boldsymbol{m}_j)(\boldsymbol{m}_i^\mathsf{T}\boldsymbol{m}_j)$ and therefore $\text{tr}(\boldsymbol{Q}^T\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{Q}) = \sum_{i=1}^{p}\sum_{j=1}^{p}\alpha_i\alpha_j(\boldsymbol{m}_i^\mathsf{T}\boldsymbol{m}_j)(\boldsymbol{m}_i^\mathsf{T}\boldsymbol{Q}\boldsymbol{Q}^\mathsf{T}\boldsymbol{m}_j)$. In that way, the objective function can be rewriting as the following quadratic form:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \; & \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{G}\boldsymbol{\alpha} \\ \text{s.t.} \quad & \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{\alpha} = 1 \end{aligned} \tag{6.17}$$

where $\boldsymbol{G} \in \mathbb{R}^{p\times p}$ is an auxiliary matrix with elements $g_{ij} = (\boldsymbol{m}_i^\mathsf{T}\boldsymbol{m}_j)\boldsymbol{m}_i^\mathsf{T}\boldsymbol{Q}\boldsymbol{Q}^\mathsf{T}\boldsymbol{m}_j$, $i,j = 1,\ldots,p$. As consequence, the previous equation becomes the objective function to be used in the unsupervised $Q-\alpha$ algorithm, described below.

In the following, it is demonstrated the solution of the optimization problem.

Proposition 6.3.1. *Given the quadratic form* $\boldsymbol{\alpha}^T\boldsymbol{G}\boldsymbol{\alpha}$ *to be maximized regarding* $\boldsymbol{\alpha}$*, a feasible solution can be obtained from eigen-decomposition of matrix* $\boldsymbol{G}$*; and, in fact, the largest eigenvector is an optimal solution.*

Proof 6.3.1. *It is easy to prove that a feasible set of solutions can be obtained via eigen-decomposition. Firstly, it is evident that* $\boldsymbol{\alpha}^T\boldsymbol{G}\boldsymbol{\alpha} = \lambda$*, where* λ *is any scalar value. Then, as* $\boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1$*, the quadratic form can be written as* $\boldsymbol{G}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}$*, that corresponds to an equality to compute the eigenvectors. In addition, solving for the variable* λ*,* $\lambda = \boldsymbol{\alpha}^T\boldsymbol{G}\boldsymbol{\alpha}$*, it can be seen that the quadratic form presents the major value when* λ *is the largest eigenvalue. Therefore, the maximal value of the quadratic*

Algorithm 1 Power-embedded $Q-\alpha$ method [3]

1. Initialize: $\boldsymbol{M}=\boldsymbol{X}^{\mathsf{T}}$, chose at random $k\times n$ matrix $\boldsymbol{Q}^{(0)}$ $(\boldsymbol{Q}^{(0)\mathsf{T}}\boldsymbol{Q}^{(0)}=\boldsymbol{I}_n)$, $\boldsymbol{m}_i\leftarrow(\boldsymbol{m}_i-\mu(\boldsymbol{m}_i))/\|\boldsymbol{m}_i\|$.
2. Make $\boldsymbol{G}^{(r)}: g_{ij}=(\boldsymbol{m}_i^{\mathsf{T}}\boldsymbol{m}_j)\boldsymbol{m}_i^{\mathsf{T}}\boldsymbol{Q}^{(r-1)}\boldsymbol{Q}^{(r-1)\mathsf{T}}\boldsymbol{m}_j$
3. Compute $\boldsymbol{\alpha}^{(r)}$ as the eigenvector associated with the major eigenvalue of $\boldsymbol{G}^{(r)}$.
4. Compute matrix: $\boldsymbol{A}_{\boldsymbol{\alpha}}^{(r)}=\boldsymbol{M}^{\mathsf{T}}\mathrm{diag}(\boldsymbol{\alpha}^{(r)})\boldsymbol{M}$
5. Compute the orthonormal transformation: $\boldsymbol{Z}^{(r)}=\boldsymbol{A}_{\boldsymbol{\alpha}}^{(r)}\boldsymbol{Q}^{(r-1)}$
6. Compute QR decomposition: $[\boldsymbol{Q}^{(r)},\boldsymbol{R}]=\mathrm{qr}\,(\boldsymbol{Z}^{(r)})$
7. Make $r\leftarrow r+1$ and return to the step 2

expression occurs when $\boldsymbol{\alpha}$ is the largest eigenvector, i.e, eigenvector associated with the largest eigenvalue. □

It must be quoted that given that the matrix $\boldsymbol{G}$ is obtained from an arbitrary orthonormal transformation, it is necessary to apply an iterative method to tune the matrix $\boldsymbol{Q}$ and the weighting vector $\boldsymbol{\alpha}$. From the optimization problem, described by (6.17), it can be seen that vector $\boldsymbol{\alpha}$ points to the direction of most relevant features, while matrix $\boldsymbol{Q}$ means its rotation, and therefore the adjustment of these parameters should be mutually dependent and must be achieved on an alternating way, as shown in algorithm 1, which in step 6 introduces a QR decomposition, to refine matrix $\boldsymbol{Q}$ at each iteration. Then, the q most relevant features are those elements of $\boldsymbol{M}$ that satisfy $\sum_{i=1}^{q}\alpha_i^2\approx\sigma_e/100$, for a given percentage fraction σ_e of accumulated variance.

In the next section it is shown that the re-computation of $\boldsymbol{\alpha}$ does not alter the convergency property of the orthogonal iteration scheme, thus the overall scheme converges to a local maxima.

6.4 Convergence of Power-Embedded $Q-\alpha$ Method

An indicator of the algorithm convergence could be the change of the vector $\boldsymbol{\alpha}$, i.e, the difference between the current and preceding vector: $\|\boldsymbol{\alpha}^{(r)}-\boldsymbol{\alpha}^{(r-1)}\|<\delta$, where $\delta\geq 0$ stands for any needed accuracy amount, being $\boldsymbol{\chi}^{(r)}$ achieved value of $\boldsymbol{\chi}$ for r-th iteration.

Nevertheless, it is possible to prove the claim for the case $q=1$, i.e., the scheme optimizes over the weight vector $\boldsymbol{\alpha}$ and the largest eigenvector $\boldsymbol{q}$ of $\boldsymbol{A}_{\boldsymbol{\alpha}}$ in the Algorithm

1.

Proposition 6.4.1. (convergence of the Algorithm 1). *The Power-embedded $Q-\alpha$ method converges to a local maxima of the optimization function given by the expression (6.16).*

Proof 6.4.1. *Because the computation of $\boldsymbol{\alpha}$ is analytic, i.e. the largest eigenvector of $\boldsymbol{G}$, and because the optimization energy is bounded from the expression in (6.16), it is sufficient to show that the computation of $\boldsymbol{q}$ monotonically increases the criterion function. It is therefore sufficient to show that:*

$$\boldsymbol{q}^{(r)}\boldsymbol{A}_{\boldsymbol{\alpha}}^2\boldsymbol{q}^{(r)} \geq \boldsymbol{q}^{(r-1)}\boldsymbol{A}_{\boldsymbol{\alpha}}^2\boldsymbol{q}^{(r-1)}, \tag{6.18}$$

for all symmetric matrices $\boldsymbol{A}_{\boldsymbol{\alpha}}$. Since steps 5 and 6 of the algorithm are equivalent to the step:

$$\boldsymbol{q}^{(r)} = \frac{\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{q}^{(r-1)}}{\|\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{q}^{(r-1)}\|},$$

the right hand side can be substituted into (6.18) and to obtain the following condition:

$$\boldsymbol{q}^{\mathsf{T}}\boldsymbol{A}_{\boldsymbol{\alpha}}^2\boldsymbol{q} \leq \frac{\boldsymbol{q}^{\mathsf{T}}\boldsymbol{A}_{\boldsymbol{\alpha}}^4\boldsymbol{q}}{\boldsymbol{q}^{\mathsf{T}}\boldsymbol{A}_{\boldsymbol{\alpha}}^2\boldsymbol{q}}, \tag{6.19}$$

which needs to be shown to hold for all symmetric matrices $\boldsymbol{A}_{\boldsymbol{\alpha}}$ and unit vectors $\boldsymbol{q}$.

Let $\boldsymbol{q} = \sum_i \gamma_i \boldsymbol{v}_i$ be represented with respect to the orthonormal set of eigenvectors $\boldsymbol{v}_i$ of the matrix $\boldsymbol{A}_{\boldsymbol{\alpha}}$. Then, $\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{q} = \sum_i \gamma_i\lambda_i\boldsymbol{v}_i$, where λ_i is the i-th eigenvalue. Since $\boldsymbol{q}^{\mathsf{T}}\boldsymbol{A}_{\boldsymbol{\alpha}}^2\boldsymbol{q} \geq \boldsymbol{0}$, it is sufficient to show that: $\|\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{q}\|^4 \leq \|\boldsymbol{A}_{\boldsymbol{\alpha}}^2\boldsymbol{q}\|^2$, or equivalently:

$$\left(\sum_i \gamma_i^2\lambda_i^2\right)^2 \leq \sum_i \gamma_i^2\lambda_i^4. \tag{6.20}$$

Let $\mu_i = \lambda_i^2$ and let $f(x) = x^2$. The following expression takes place:

$$f(\sum_i \gamma_i^2\mu_i) \leq \sum_i \gamma_i^2 f(\lambda_i^2), \tag{6.21}$$

which follows from convexity of $f(x)$ and the fact that $\sum_i \gamma_i^2 = 1$. □

6.5 Sparsity and Positivity of α

The optimization criteria (6.16) are formulated as a least-squares problem and as such there does not seem to be any apparent guarantee that the weights $\alpha_1, \ldots, \alpha_p$ would come out *non-negative* (same sign condition), and in particular sparse when there exists a sparse solution (i.e., there is a relevant subset of features which induces a coherent clustering).

The positivity of the weights is a critical requirement for the $Q - \alpha$ to form a "feature weighting" scheme. In other words, if it is possible guarantee that the weights would come out non-negative then $Q - \alpha$ would provide feature weights which could be used for selection or for simply weighting the features as they are being fed into the inference engine of choice. If in addition the feature weights exhibit a "sparse" profile, i.e., the gap between the high and low values of the weights is high, then the weights could be used for selecting the relevant features as well. Thereafter, the gap between the high and low weights is defined as "sparsity gap" and will be discussed later the value of the gap in simplified domains. With the risk of abusing standard terminology, the property of having the weight vector concentrate its (high) values around a number of coordinates as a sparsity feature. Typically, in the algorithm, none of the values of the weight vector strictly vanish.

For most feature weighting schemes, the conditions of positivity and sparsity should be specifically presented into the optimization criterion one way or the other. The possible means for doing so include introduction of inequality constraints, use of L_0 or L_1 norms, adding specific terms to the optimization function to "encourage" sparse solutions or use a multiplicative scheme of iterations which preserve the sign of the variables throughout the iterations.

The formal details of the proof of sparsity and positivity of $\boldsymbol{\alpha}$ are discussed in the appendix B.

6.6 A Parameter Free Algorithm

The procedure above described for computing the weighting vector, $\boldsymbol{\alpha}$, is refined iteratively, and the whole data set is to be used, where the orthonormal matrix is updated per iteration to get the subset of relevant features. As a result, the computational load may increase. Nonetheless, based on variance criterion, it can be infered that the first q components of $\widehat{\boldsymbol{x}}^{(l)}$ hold the most informative directions of weighting data,

thus, the l $(q+1 \leq l \leq p)$ directions do not significantly contribute to the explained variance. Then, time calculation when computing the vector $\boldsymbol{\alpha}$ can be reduced just to one iteration with no significant decrease of accuracy [3]. With this in mind, the feature relevance may be preserved optimizing the p original variables or the first q variables. Indeed, maximizing

$$\mathrm{tr}(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{Q}) \tag{6.22}$$

is equivalent to maximize

$$\mathrm{tr}(\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{A}_{\boldsymbol{\alpha}}) = \mathrm{tr}(\boldsymbol{X}\mathrm{diag}(\boldsymbol{\alpha})\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\mathrm{diag}(\boldsymbol{\alpha})\boldsymbol{X}^{\mathsf{T}}). \tag{6.23}$$

Since this expression is bilinear regarding $\boldsymbol{\alpha}$, the objective function can be re-written as

$$\boldsymbol{\alpha}^{T}\boldsymbol{H}\boldsymbol{\alpha} \tag{6.24}$$

where

$$\boldsymbol{H}_{ij} = \mathrm{tr}(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}_i\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{x}_j) = \boldsymbol{x}_i\boldsymbol{x}_j^{\mathsf{T}}\,\mathrm{tr}(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}_j) = (\boldsymbol{x}_i\boldsymbol{x}_j^{\mathsf{T}})^2. \tag{6.25}$$

Accordingly, it can be infered that the approximate vector of relevance $\widehat{\boldsymbol{\alpha}}$ is the eigenvector corresponding to the largest eigenvalue of $(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{\cdot 2}$ (where notation $(\boldsymbol{\chi})^{\cdot 2}$ stands for the square of each one of the elements of the involved matrix $\boldsymbol{\chi}$).

In conclusion, the weighting factor is related to either vectors: $\boldsymbol{\alpha}$ (complete case) and $\widehat{\boldsymbol{\alpha}}$ (approximate case). Thus, the weighting matrices become $\boldsymbol{W}_{\alpha} = \mathrm{diag}(\sqrt{\alpha})$ and $\boldsymbol{W}_{\widehat{\alpha}} = \mathrm{diag}(\sqrt{\widehat{\alpha}})$, respectively.

6.7 Projection of Weighted Data

As described above, the data is weighted by the diagonal matrix $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{w})$, where $\boldsymbol{w}$ is the weighting vector that can be calculated using either the MSE or the M inner-product-based approaches above explained. Therefore, weighting data $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{W}$ is linearly projected, so: $\boldsymbol{Y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{V}}$, where $\widetilde{\boldsymbol{V}}$ are the principal components of $\widetilde{\boldsymbol{X}}$, $\widetilde{\boldsymbol{V}} = \boldsymbol{V}$ if $\boldsymbol{W} = \mathrm{diag}(\mathbf{1}_p)$. The attained procedure for relevance analysis and rotation of weighted data based on described methods is shown in Algorithm 2.

Algorithm 2 Projection of weighted data.

1. (Initialization): Normalize $\boldsymbol{X}$, $\mu(\boldsymbol{x}_i) = 0$, $\|\boldsymbol{x}_i\| = 1$, $1 \leq i \leq p$
2. Choose a method to find the weighting vector $\boldsymbol{w}$

(a) $\boldsymbol{w} \leftarrow \sqrt{\boldsymbol{\alpha}}$, Eigenvector corresponding to the largest eigenvalue of $\boldsymbol{G}$ (algorithm 1, $r \leftarrow$ last iteration)

(b) $\boldsymbol{w} \leftarrow \sqrt{\widehat{\boldsymbol{\alpha}}}$, Eigenvector corresponding to the largest eigenvalue of $(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{\cdot 2}$.

(c) $\boldsymbol{w} \leftarrow \sqrt{\boldsymbol{\rho}}$, see (6.9), removing eigenvectors $[q+1, \ldots, p]$ that do not significantly contribute to variance.

(d) $\boldsymbol{w} \leftarrow \sqrt{\widehat{\boldsymbol{\rho}}}$, see (6.9), $q = 1$.

3. Weight original data: $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\mathrm{diag}(\boldsymbol{w})$
4. Compute principal components: $\widetilde{\boldsymbol{V}}$ of $\widetilde{\boldsymbol{X}}$
5. Project data: $\boldsymbol{Y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{V}}$

Chapter 7

Clustering

7.1 Introduction

Unsupervised analysis encloses all discriminative methods, which do not require a prior knowledge about classes for classification task. Generally, they only need some initialization parameters such as the number of groups to be formed or some other hint about the initial partition. Therefore, the unsupervised analysis, in classification terms, allows to group homogeneous patterns without using any information on the nature of classes in data set. For this reason, with unsupervised analysis is not achieved an automatic classification but subsets of homogeneous data generated from distances, dissimilarities or statistical measures based criteria. Then, term non-supervised classification is related to grouping of data into similar subsets.

There are several reasons for interest in unsupervised procedures, among them:

- Unsupervised methods are useful when collecting and labeling of a large set of sample patterns is surprisingly costly or non-feasible.
- In case of the variables or features do not change significatively over time, unsupervised algorithms converge fast and lead a proper partition.
- They allow to categorize and find hierarchical elements.

However, an unsupervised analysis system-generated solution can be affected because of factors as non-proper initial parameters, that can generate wrongly a convergence value.

Unsupervised grouping or clustering has shown to be useful and very versatile in data exploratory analysis. Then, different methods of clustering have been developed to solve several problems such as computational cost, sensitivity to initialization, unbalanced classes, convergence to a local optima distant from global optima, among others. However, choosing a method is not a trivial task, it must be taken into account the nature of the data and operating conditions in order to group similar patterns in such way as to get a good tradeoff between computational cost and effectiveness in the separability of classes.

There are exact methods to solve this problem, provided the dataset is very small. For large datasets, methods based on heuristic searchs are used instead, such as partitional algorithms [151].

In the literature, clustering algorithms that classify the given data into a single collection of homogeneous pattern subsets are called partitional, where subsets or clusters are refined, commonly, in a iterative fashion. The difference between one algorithm and another is given by the measure to quantify the grouping quality and the partition updating function. Regarding the context of this work, clustering is the most frequently used technique for automatic analysis of heartbeat patterns to detect pathologies into Holter recordings.

In this chapter, representative algorithms of partitional clustering, such as K-means (Section 7.2.1) and H-means (Section 7.2.2), are studied. Also, in Section 7.2.3, it is described a general iterative model for clustering that is based on H-means principle and allows to develop several clustering methods. In Section 7.3, some initialization algorithms for partitional methods are described. In addition, in Section 7.5, a sequential clustering scheme is proposed, where initial data are divided and processed into segments in order to decrease the computational cost and avoid wrong classification of minority classes.

Notation

Data matrix to be clustered will be denoted by $\boldsymbol{X} \in \mathbb{R}^{n \times p} : \boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\mathsf{T}$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\mathsf{T}$ is the i-th observation or sample. Array $\boldsymbol{P}_k$ will be the partition set of $\boldsymbol{X}$ where k is the number of clusters, and $\boldsymbol{C} = \{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_k\}$ will be the k-dimensional clusters set, with $\boldsymbol{Q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k)^\mathsf{T}$ as the corresponding centers set.

7.2 Center-based Clustering

The classical technique of unsupervised grouping is the partitional clustering or center-based clustering (CBC), which has the goal of minimizing an objective function to obtain an optimal solution via iterative center-updating [152]. The objective function defines how good a clustering solution is and must be related to the center updating function. At the end, the final partition that is is refined iteratively by means a center-updating procedure until satisfying a convergence criterion, represents the clustering solution. CBC algorithms are distinguished by their objective functions and corresponding center-updating functions.

For instance, in the minimum sum of squares based clustering (MSSC), explained widely in [151], the objective function can be expressed as:

$$\min_{\boldsymbol{\rho}_k \in \boldsymbol{P}_k} \sum_{j=1}^{k} \sum_{\boldsymbol{x}_l \in \boldsymbol{C}_l} \|\boldsymbol{x}_l - \boldsymbol{q}_j\|^2 \tag{7.1}$$

where $\|\cdot\|$ represents the Euclidean norm and the j-th center or centroid is given by:

$$\boldsymbol{q}_j = \frac{1}{\mathrm{n_e}(\boldsymbol{C}_j)} \sum_{l:\boldsymbol{x}_l \in \boldsymbol{C}_j} \boldsymbol{x}_l, \quad j = 1, \ldots, k \tag{7.2}$$

where $\mathrm{n_e}(\cdot)$ represents the number of points or elements of its argument. Henceforth, this notation will be used throughout this document.

The aim of this method is finding out the data partition that minimices the distance between elements belonging to each cluster and its respective center, i.e., the within-classes variance. The MSSC solution can be achieved from, the best known and most used, K-means and H-means algorithms.

7.2.1 K-means

In this method, a start partition associated to an initial center set is chosen and their center reassignments changes, that are done to generate new partitions, are assessed per each iteration. Then, once a center is moved, all reassignments are done and the objective function change due to this movement is computed.

By assuming a data point $\boldsymbol{x}_i$ that belongs to $\boldsymbol{C}_l$ for the current solution is reassigned to another cluster $\boldsymbol{C}_j$, the center updating can be accomplished applying the following

equations:

$$\boldsymbol{q}_l \leftarrow \frac{n_l \boldsymbol{q}_l - \boldsymbol{x}_i}{n_l - 1}, \quad \boldsymbol{q}_j \leftarrow \frac{n_j \boldsymbol{q}_j + \boldsymbol{x}_i}{n_j + 1} \tag{7.3}$$

donde $n_i = \mathrm{n_e}(\boldsymbol{C}_i)$ y $l \neq j$.

Changes of the objective function value caused by reassignments are computed using

$$v_{ij} = \frac{n_j}{n_j + 1} \|\boldsymbol{q}_j - \boldsymbol{x}_i\|^2 - \frac{n_l}{n_l - 1} \|\boldsymbol{q}_l - \boldsymbol{x}_i\|^2, \ \boldsymbol{x}_i \in \boldsymbol{C}_l \tag{7.4}$$

The previous equation is applied in case of MSSC objective function. In general, a specific objective function must be considered, so

$$v_{ij} = \frac{n_j}{n_j + 1} f(\boldsymbol{q}_j, \boldsymbol{x}_i) - \frac{n_l}{n_l - 1} f(\boldsymbol{q}_l, \boldsymbol{x}_i), \ \boldsymbol{x}_i \in \boldsymbol{C}_l \tag{7.5}$$

where f is the objective function expression corresponding to some criterion or clustering method. Such changes are computed for all possible reassignments. Then, if they are all non-negative ($v_{ij} \geq 0$) the procedure stops with a partition corresponding to a local minimum. Otherwise, the reassignment reducing most the objective function value is performed and the procedure iterated [151]. The general K-means heuristic for cluster updating is described in algorithm 3.

Because of continuous assessment of objective function changes, K-means algorithm could provide better convergence value than other algorithms, since each center is updated independently; but, in turn, it could represent a higher computational cost.

7.2.2 H-means

H-means algorithm is a variant of K-means, where centers are updated once per iteration. Therefore, under any criterion, all k centers are established before assessing the objective function change. Thus, computacional cost decreases in comparison a K-means based approach.

In brief, H-means works as follows. An initial partition $\boldsymbol{C} = \{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_k\}$ is chosen at random and the centers of each cluster $\boldsymbol{Q} = \{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k\}$ are computed. Then, each data point is assigned (reallocated) to its closest centroid $\boldsymbol{q}_j$ (according to some criterion), if no change in assignments occurs, the heuristic stops with a locally minimum partition. Otherwise, the centers are updated and the procedure iterated

Algorithm 3 K-means

1. Initialization: set k, initial partition $\boldsymbol{C}^{(0)}$ with their centers $\boldsymbol{Q}^{(0)}$ and maximum number of iterations N_{iter}. Do $r = 1$.

while $r < N_{iter}$ **do**

for $j = 1 \ldots k$ **do**

2. Reassign centers: $\boldsymbol{q}_l^{(r)} \leftarrow \dfrac{n_l \boldsymbol{q}_l^{(r-1)} - \boldsymbol{x}_i}{n_l - 1}, \quad \boldsymbol{q}_j^{(r)} \leftarrow \dfrac{n_j \boldsymbol{q}_j^{(r-1)} + \boldsymbol{x}_i}{n_j + 1}$

3. Compute change of objective function value: $v_{ij} = \dfrac{n_j}{n_j + 1} \|\boldsymbol{q}_j^{(r)} - \boldsymbol{x}_i\|^2 - \dfrac{n_l}{n_l - 1} \|\boldsymbol{q}_l^{(r)} - \boldsymbol{x}_i\|^2$, $\boldsymbol{x}_i \in \boldsymbol{C}_l^{(r)}$

if $v_{ij} \geq 0$ $(i = 1, \ldots, n$ and $j = 1, \ldots, k)$ **then**

Process stops with resultant partition $\boldsymbol{C}^{(r)}$

else

$r \leftarrow r + 1$

end if

end for

end while

[151]. In algorithm 4 are shown the details of how this method works.

Algorithm 4 H-means

1. Initialization: set k, initial centers $\boldsymbol{Q}^{(0)} = (\boldsymbol{q}_1^{(0)}, \ldots, \boldsymbol{q}_k^{(0)})^{\mathsf{T}}$, initial assignment $\boldsymbol{C}^{(0)}$, maximum number of iterations N_{iter} and precision threshold δ. Do $r = 1$.

while $r < N_{iter}$ **do**

2. Update centers: $\boldsymbol{Q}^{(r)} = \varphi_{\boldsymbol{Q}}(\boldsymbol{C}^{(r-1)}, \boldsymbol{Q}^{(r-1)})$

3. Assign each element: $\boldsymbol{C}^{(r)} = \varphi_{\boldsymbol{C}}(\boldsymbol{X}, \boldsymbol{Q}^{(r)})$

if $|\,\mathrm{d}(\boldsymbol{q_j}^{(r)}, \boldsymbol{q_j}^{(r-1)})| < \delta$ $(j = 1, \ldots, k)$ **then**

Process stops with a final partition $\boldsymbol{C}^{(r)}$

else

$r \leftarrow r + 1$

end if

end while

Function $\mathrm{d}(\cdot, \cdot)$ represents a distance or dissimilarity measure between those two

vectors of its argument. In this case, this function is applied to measure the change of current center set with respect to immediately previous center set, so when such change is less than a prefixed value δ, the algorithm converges. Variables $\varphi_{\boldsymbol{Q}}$ and $\varphi_{\boldsymbol{C}}$ represent the center updating function in terms of previous partition and the partition updating from current centers, respectively. In Section 7.2.4, it is described this method in detail.

7.2.3 General iterative clustering

In the general model for clustering algorithms that use iterative optimization, described in [152], centers are computed using a membership function $m(\boldsymbol{q}_j \mid \boldsymbol{x}_i)$ and a weight function $w(\boldsymbol{x}_i)$, which respectively define the proportion of data point $\boldsymbol{x}_i$ that belongs to center $\boldsymbol{q}_j$ and how much influence data point $\boldsymbol{x}_i$ has in recomputing the centroid parameters for the next iteration. By assuming that membership is a non-negative value and the absolute membership is 1, function m must satisfy two important conditions:

$$m(\boldsymbol{q}_j \mid \boldsymbol{x}_i) \geq 0 \text{ and } \sum_{j=1}^{k} m(\boldsymbol{q}_j \mid \boldsymbol{x}_i) = 1.$$

This function is called *hard* when it can only take discrete values, i.e., $m \in \{0, 1\}$. Otherwise, $0 \leq m \leq 1$ and it is called *soft*.

Both functions, m and w, are directly related to the nature of the objective function as can be seen in Sections 7.2.4 and 7.2.5.

According to this method, the center updating function can be written as:

$$\boldsymbol{q}_j = \frac{\sum_{i=1}^{n} m(\boldsymbol{q}_j \mid \boldsymbol{x}_i) w(\boldsymbol{x}_i) \boldsymbol{x}_i}{\sum_{i=1}^{n} m(\boldsymbol{q}_j \mid \boldsymbol{x}_i) w(\boldsymbol{x}_i)}, \quad j = 1, \ldots, k \tag{7.6}$$

Previous equation is associated with the expression commonly used in geometry to compute a centroid: $\boldsymbol{q} = \sum_i g(\boldsymbol{r}_i)\boldsymbol{r}_i / \sum_i g(\boldsymbol{r}_i)$, where $\boldsymbol{r}_i$ is the position vector corresponding to the i-th element and $g(\cdot)$ is the mass density function.

Given that the membership and weight functions can be adjusted to any objective function (taking into account the constraints discussed above) and the centers are refined iteratively, this method represents a general iterative model (GIM) for unsupervised grouping. The heuristic of this model is the same as H-means algorithm, therefore all centers are updated before applying the convergence control, i.e, computing change