of the objective function value to stop if no significative change occurs. This model can be applied for CBC as is described in algorithm 5, where data points are assigned to certain cluster if its membership value is highest.

Algorithm 5 General iterative model for CBC

1. Initialization: set initial centers $\boldsymbol{Q}^0 = (\boldsymbol{q}_1^{(0)}, \ldots, \boldsymbol{q}_k^{(0)})^\mathsf{T}$, maximum number of iterations N_{iter}. Do $r = 0$.
2. Compute the membership value $m(\boldsymbol{q}_j^{(r-1)} \mid \boldsymbol{x}_i)$ and weight $w(\boldsymbol{x}_i)$ for each data point
3. Update centers using: $\boldsymbol{q}_j^{(r)} = \dfrac{\sum_{i=1}^{n} m(\boldsymbol{q}_j^{(r-1)} \mid \boldsymbol{x}_i) w(\boldsymbol{x}_i) \boldsymbol{x}_i}{\sum_{i=1}^{n} m(\boldsymbol{q}_j^{(r-1)} \mid \boldsymbol{x}_i) w(\boldsymbol{x}_i)}, \; j = 1, \ldots, k$
4. $r \leftarrow r + 1$ and repeat the steps 2 and 3 until a number of iterations N_{iter} or convergence
5. Assign all data points so: $\boldsymbol{x}_i \in \boldsymbol{C}_l^{(r)}$ if $l = \arg\max_j m(\boldsymbol{q}_j^{(r)}, \boldsymbol{x}_i)$

The convergence of a GIM-based algorithm can be measured by comparing the objective function value obtained from the current partition $f(\boldsymbol{C}^{(r)})$ and the value obtained with the partition generated in immediately previous iteration $f(\boldsymbol{C}^{(r-1)})$, via a difference criterion $|f(\boldsymbol{C}^{(r)}) - f(\boldsymbol{C}^{(r-1)})| < \delta$, a quotient criterion $f(\boldsymbol{C}^{(r)})/f(\boldsymbol{C}^{(r-1)}) \approx 1$, among others.

Next are described some clustering methods from GIM.

7.2.4 H-means based on GIM

The objective function that the H-means algorithm optimizes is:

$$HM(\boldsymbol{X}, \boldsymbol{Q}) = \sum_{i=1}^{n} \min_{j \in 1,\ldots,k} ||\boldsymbol{x}_i - \boldsymbol{q}_j||^2 \tag{7.7}$$

where $||\cdot||$ represents the euclidian norm in case of MSSC.

H-means has a hard membership function and fixed weights, as is shown in the

following expressions:

$$m_{HM}(\boldsymbol{q}_j \mid \boldsymbol{x}_i) = \begin{cases} 1 & \text{if } l = \arg\min_j ||\boldsymbol{x}_i - \boldsymbol{q}_j||^2 \\ 0 & \text{otherwise} \end{cases} \tag{7.8}$$

and

$$w_{HM}(\boldsymbol{x}_i) = 1 \tag{7.9}$$

Note that the objective function of H-means is the same as MSSC. Furthermore, by replacing the expressions (7.8) and (7.9) in (7.6), the same center updating function as MSSC is obtained (see (7.2)). Therefore, the general iterative model is demonstrated.

7.2.5 Gaussian expectation maximization

The objective function of Gaussian expectation maximization (GEMC) is a linear combination of gaussian distributions centered at each centroid $\boldsymbol{q}_j$. The objective function to be maximized can be written as [153].:

$$GEMC(\boldsymbol{X}, \boldsymbol{Q}) = -\sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} p(\boldsymbol{x}_i \mid \boldsymbol{q}_j) p(\boldsymbol{q}_j) \right) \tag{7.10}$$

where $p(\boldsymbol{x}_i \mid \boldsymbol{q}_j)$ is the probability of $\boldsymbol{x}_i$ since it is generated by a Gaussian distribution centered at $\boldsymbol{q}_j$, and $p(\boldsymbol{q}_j)$ is the prior probability of the cluster associated to center $\boldsymbol{q}_j$. The log function is used for simplicity, and the minus sign accounts for minimization. This method employs soft membership and fixed weights, given by the following expressions:

$$m_{GEMC}(\boldsymbol{q}_j \mid \boldsymbol{x}_i) = \frac{p(\boldsymbol{x}_i \mid \boldsymbol{q}_j) p(\boldsymbol{q}_j)}{p(\boldsymbol{x}_i)} \tag{7.11}$$

and

$$w_{GEMC}(\boldsymbol{x}_i) = 1 \tag{7.12}$$

Given the nature of this method, Bayes rule is used to compute m_{GEMC}, where term $p(\boldsymbol{x}_i)$ is the evidence or total probability defined as follows:

$$p(\boldsymbol{x}_i) = \textstyle\sum_{j=1}^{k} p(\boldsymbol{x}_i \mid \boldsymbol{q}_j) p(\boldsymbol{q}_j)$$

In the parametric case, term $p(\boldsymbol{x}_i \mid \boldsymbol{q}_j)$ can be computed as a normal gaussian centered at centroid $\boldsymbol{q}_j$ and covariance $\boldsymbol{\Sigma}_j$, as follows:

$$p(\boldsymbol{x}_i \mid \boldsymbol{q}_j) = \mathcal{N}(\boldsymbol{q}_j, \boldsymbol{\Sigma}_j) = \frac{1}{\det(\boldsymbol{\Sigma}_j)^{\frac{1}{2}}}(2\pi)^{-p/2} e^{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})\boldsymbol{\Sigma_j}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})^\intercal}, \tag{7.13}$$

where $\boldsymbol{\Sigma}_j$ is the covariance matrix, which can be unique $\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma} = \text{cov}(\boldsymbol{X})$. It could also be computed for each cluster and updated for each iteration using $\boldsymbol{\Sigma}_j = \text{cov}(\boldsymbol{C}_j)$, being preferable because, in this way, change of clusters variance per iteration is taken into account. Thus, the objective function of this method could converge to a better value.

As variant of this method, Parzen's method can be used to estimate the membership function. Then, method becomes non-parametric density-based clustering (NPDBC). Thereby, for NPDBC using Parzen's estimation, the membership function is the same as GEMC, except that the term $p(\boldsymbol{x}_i \mid \boldsymbol{q}_j)$ is computed as follows:

$$p(\boldsymbol{x}_i \mid \boldsymbol{q}_j) = \frac{1}{nh} \sum_{i=1}^{n} \mathcal{K}\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right) \tag{7.14}$$

where $\mathcal{K}$ is a Gaussian kernel which is given by:

$$\mathcal{K}(\boldsymbol{z}) = \frac{1}{(2\pi)^{-p/2}} e^{-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^\intercal} \tag{7.15}$$

7.3 Initialization Algorithms

One of the biggest problems of CBC algorithms is the convergence to a local optimum distant from the global optimum. This can be attributed to the sensitive to initialization that this kind of algorithms present. For this reason, there exist several initialization algorithms that guarantee a proper initial partition. In this study, J-means and max-min algorithms are explored.

7.3.1 Max-min algorithm

The aim of max-min algorithm is to find, into the set of data $\boldsymbol{X}$, the k elements that are further away from each other, improving the number of necessary groups to classify the classes and the convergence value [130]. This algorithm starts with a random data point of $\boldsymbol{X}$ as the first center and the rest of them are chosen following an strategy,

in which selected element in the i-th iteration is the element that is the further one among the $i-1$ chosen elements. Then, the first center $\boldsymbol{q}_1$ is chosen randomly from $\boldsymbol{X}$, and the second center $\boldsymbol{q}_2$ is the data point which presents the maximum distance between $\boldsymbol{q}_1$ and remaining points $\{\boldsymbol{X}-\boldsymbol{q}_1\}$. Since these centers, the rest of them can be obtained using the max-min criterion, as follows

$$f(\boldsymbol{x}_l) = \max_{\boldsymbol{x}_i \in \{\boldsymbol{X}-\boldsymbol{Q}\}} \left\{ \min_{\boldsymbol{q}_j \in \boldsymbol{Q}} \|\boldsymbol{x}_i - \boldsymbol{q}_j\|^2 \right\}, \quad j = 1, \ldots, k \tag{7.16}$$

where $\|\cdot\|$ represents the euclidian norm.

The max-min algorithm is described in Algorithm 6

Algorithm 6 Max-Min algorithm

function $\boldsymbol{Q}$:= maxmin($\boldsymbol{X}$,k) {$\boldsymbol{X} \in \mathbb{R}^{n \times p}$, k is the number of clusters}

Require: $1 < k \leq n$;

$\boldsymbol{q}_1$:= Random(X); $\boldsymbol{x}_f := \boldsymbol{q}_1$; {It is the first centroid}

$l^* := \underset{\substack{1 \leq l \leq n \\ l \neq f}}{\arg\max} \|\boldsymbol{x}_f - \boldsymbol{x}_l\|^2$; $\boldsymbol{q}_2 := \boldsymbol{x}_{l^*}$; {Is the second centroid}

$\boldsymbol{Q} := \{\boldsymbol{q}_1\} \cup \{\boldsymbol{q}_2\}$;

if $k > 2$ **then**

for i =3 to k **do**

len := size(Q);

for j=1 to n **do**

$$\text{distance}_j := \min_{\substack{\boldsymbol{q}_i \in \boldsymbol{Q} \\ 1 \leq i \leq len;}} \|\boldsymbol{x}_j - \boldsymbol{q}_i\|^2;$$

end for

$$j^* := \underset{1 \leq j \leq n}{\arg\max}\{\text{distance}_j\};$$

$\boldsymbol{q}_i := \boldsymbol{x}_{j^*}$; $\boldsymbol{Q} := \boldsymbol{Q} \cup \{\boldsymbol{q}_i\}$;

end for

end if

7.3.2 J-means algorithm

J-means algorithm consist of updating the centers trough local assessment of objective function, i.e., only taking into consideration a certain region around the centers instead

of all data space [151]. This algorithm works as follows. After a random initialization, every point $\boldsymbol{p}_i$ out of a sphere of radius ε with center $\boldsymbol{q}_j$ is considered as a centroid candidate. Thus, $\boldsymbol{p}_i$ replaces a current centroid $\boldsymbol{q}_j$. After updating, the objective function value is calculated using only the new centroid. Then, the original objective function (previous value f^1) is compared with the new objective function value (f^2). Thereby, if $f^1 > f^2$, the process stops; otherwise the algorithm starts again using the same initial partition and its updates.

Parameter ε is chosen in such way that no intersections among spheres occurs, for that reason is a necessary condition that $\varepsilon < \frac{1}{2}\min\|\boldsymbol{q}_j - \boldsymbol{q}_i\| \quad i \neq j$. In conventional J-means, MSSC conditions are employed; therefore spheres are defined from distances based criteria.

In summary, the general J-means algorithm is described in algorithm 7.

Algorithm 7 J-means algorithm

1. Initialization: chose the initial partition $\boldsymbol{C}^0 = \{\boldsymbol{C}_j^0\}_{j=1}^k$ associated to $\boldsymbol{Q}^0$, $\boldsymbol{C}^1 \leftarrow \boldsymbol{C}^0$.

2. Find unoccupied points, i.e., entities which do not coincide with a cluster centroid: points out of a sphere of radius ε ($\varepsilon < \frac{1}{2}\min\|\boldsymbol{q}_j - \boldsymbol{q}_i\| \quad i \neq j$) with center $\boldsymbol{q}_j$.

3. Find the best partition $\boldsymbol{C}_k^2$ and corresponding objective function value f^2 in the jump neighborhood of the current solution $\boldsymbol{C}_k^1$.

   If $f^1 > f^2$

4. The process stop with the solution $\boldsymbol{Q}^1$

   otherwise

5. Move to the best neighboring solution $\boldsymbol{C}_k^2$ ($\boldsymbol{C}_k^1 \leftarrow \boldsymbol{C}_k^2, f^1 \leftarrow f^2$) and return to step 2.

   End If

The J-means variants consist of changes in the sphere definition, for example, using statistical measures instead of distances. For example, employing a covariance matrix as is done in GEM-based methods. In this case, method could be called J-GEM.

7.4 Estimation of the Number of Groups

In the field of unsupervised analysis, there are no many studies about the automatic estimation of the number of the groups, being established manually. Among the differ-

ent techniques, spectral analysis has demonstrated remarkable results in the estimation of the number of clases into the data set. In this work, different alternatives for this task are described. First one is based on singular value decomposition (SVD) [154]. Secondly, it is analyzed the eigenvalues of the normalized affinity matrix. Thirdly, an approach using the eigenvectors of affinity matrix [150]. Finally, it is introduced a new approach that is based on a relevance analysis procedure [3].

7.4.1 Estimation using SVD

Many SVD properties are useful in a variety of pattern recognition problems and applications. For example, mapping a data set in a space where data are better represented for posterior classification tasks, improving processing time and classifier performance. Another application, that is not usual enough, is the estimation of the number of groups, as described below. By letting $\boldsymbol{X}$ be a $n \times p$ data matrix, then its decomposition in singular values is

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}} = \sum_{i=1}^{d} \sigma_i \boldsymbol{u}_i \boldsymbol{v}^{\mathsf{T}} \tag{7.17}$$

where $d = \min(n, p)$, $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m]$ is an $n \times n$ orthonormal matrix, $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m]$ is a $p \times p$ orthonormal matrix, $\boldsymbol{S} = \text{diag}(\boldsymbol{\sigma})$ is a diagonal matrix and $\boldsymbol{\sigma}$ is a singular value vector. Matrix $\boldsymbol{U}$ corresponds to eigenvectors of $\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}$ and matrix $\boldsymbol{V}$ holds the eigenvectors of $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$, and they are called singular right and left matrix, respectively.

The estimation of the number of groups is done under the principle of Frobenius, which establishes if a matrix is normalized with respect to Frobenius norm, i.e.,

$$\|\boldsymbol{X}\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{d} x_{ij} = n,$$

is evident that

$$\|\boldsymbol{X}\|_F^2 = \sum_{l=1}^{\rho(\boldsymbol{X})} \sigma_l^2 \tag{7.18}$$

where $\rho(\cdot)$ represents the matrix rank.

Given this, the number of groups is chosen as

$$k = \arg\min \left\{ \alpha_{svd}\, n \leq \sum_{l=1}^{k} \sigma_l^2 \right\} \tag{7.19}$$

where α_{svd} is a parameter to be tuned.

In [154], it is widely explained and discussed this application and other interesting properties of SVD.

7.4.2 Estimation analyzing eigenvalues

One possible approach to discover the number of groups is to analyze the eigenvalues of the normalized affinity matrix given by: $\widetilde{\boldsymbol{A}} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, where $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}$ is the trivial affinity matrix, $\boldsymbol{D} = \text{diag}(\widetilde{\boldsymbol{A}}\mathbf{1}_n)$ is a diagonal matrix that represents the degree of $\widetilde{\boldsymbol{A}}$ and $\mathbf{1}_n$ is a vector of all 1's, as described in [150].

In [155], it is demonstrated that the first eigenvalue (highest magnitude) of the normalized affinity matrix $\widetilde{\boldsymbol{A}}$ of magnitude 1 is repeated with multiplicity equal to the number of groups. Then, the number of groups can be estimated by counting the eigenvalues of $\widetilde{\boldsymbol{A}}$ equaling 1. However, if the groups are not clearly separated, once noise is introduced, the values can deviate from 1, thus the criterion cannot be used. An alternative approach can be reached searching for a drop in the magnitude of the eigenvalues, however, this approach lacks a theoretical justification [156].

7.4.3 Estimation using eigenvectors

A better approach can be obtained by analyzing the eigenvectors of normalized affinity matrix. First, the matrix $\widetilde{\boldsymbol{A}}$ must be divided into c submatrices or blocks $\{\widetilde{\boldsymbol{A}}^1, \ldots, \widetilde{\boldsymbol{A}}^c\}$, then, the $n \times c$ blocks diagonal matrix is formed. The parameter c ($c \leq n$) represents the initial tentative number of groups with which the heuristic starts, it is fixed manually. The blocks diagonal matrix $\hat{\boldsymbol{V}}$ is formed, in that way the eigenvalues and eigenvectors of its blocks are padded appropriately with zeros, as follows

$$\hat{\boldsymbol{V}} = \begin{bmatrix} \boldsymbol{v}^1 & \mathbf{0}_n & \mathbf{0}_n \\ \mathbf{0}_n & \cdots & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{0}_n & \boldsymbol{v}^c \end{bmatrix}_{n \times c} \tag{7.20}$$

where $\boldsymbol{v}^i$ is a n-dimensional vector which represents the i-th eigenvector of block $\widetilde{\boldsymbol{A}}^i$, $\mathbf{0}_n$ is a n-dimensional vector of all 0's.

As was mentioned above, the eigenvalue equaling 1 has multiplicity equalling the number of groups, thus, the eigensolution can be generated by any orthogonal vectors, no only eigenvectors, spanning the same subspace of $\hat{\boldsymbol{V}}$. Therefore, $\hat{\boldsymbol{V}}$ can be replaced by $\boldsymbol{V} = \hat{\boldsymbol{V}}\boldsymbol{R}$, where $\boldsymbol{R}$ is any $c \times c$ orthonormal matrix. This implies that there exists a rotation matrix $\hat{\boldsymbol{R}}$, such that each row in the matrix $\hat{\boldsymbol{V}}\hat{\boldsymbol{R}}$ has a single non-zero entry.

Let $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{R}$ be a the matrix obtained after rotating the eigenvectors matrix and denote $\beta_i = \max_j z_{ij}$. To find a new space where groups will be well represented in terms of separability, the following cost function can be minimized:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{c} \frac{z_{ij}^2}{\beta_i^2} \tag{7.21}$$

Minimizing this cost function over all the possible rotations will provide the best alignment with the canonical coordinate system, therefore the number of groups k ($1 < k \leq c$) is chosen taken one providing the minimal. If several group numbers yield the same minimal cost, the largest of those is selected. The optimization process is carried out using a stochastic descent gradient scheme, as described in [156].

7.4.4 Estimation based on spectral relevance analysis

Spectral analysis has shown to be a powerful tool in many pattern recognition applications, such as data projection and weighting. For instance, as is described in chapter 6, spectral analysis is useful to study the feature relevance. By combining the definitions given by [155] and [156] with the relevance procedure described in Section 6.3, it can be achieved a new approach to estimate the number of groups. By recalling Section 6.3, the solution of the following optimization problem provides information to determine the relevance of features:

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

From previous equation, it can be written the rotated affinity matrix given by $\boldsymbol{A}_{\boldsymbol{Q}-\boldsymbol{\alpha}} = \boldsymbol{Q}^{\top}\boldsymbol{A}_{\alpha}\boldsymbol{A}_{\alpha}\boldsymbol{Q}$, where $\boldsymbol{A}_{\boldsymbol{\alpha}}$ is the affinity matrix weighted by vector $\boldsymbol{\alpha}$ and $\boldsymbol{Q}$ is an orthonormal rotation matrix. Since the trace of $\boldsymbol{A}_{\boldsymbol{Q}-\boldsymbol{\alpha}}$ represents the sum of squared eigenvalues of $\boldsymbol{A}_{\alpha}$ and the affinity matrix holds the relation among observations, it can be infer that its corresponding diagonal provides substantial information to determine the number of groups into data matrix. Then, the number of groups can be estimated from the values into the diagonal of $\boldsymbol{A}_{\boldsymbol{Q}-\boldsymbol{\alpha}}$ and an accumulated value criterion applied over the trace of the same matrix. Mathematically, the number of groups can be chosen as the amount of elements that satisfy that truncated-trace value is less than a certain value δ, i.e.,

$$k = \text{numel}(\lfloor \text{tr}(\boldsymbol{A}_{\boldsymbol{Q}-\boldsymbol{\alpha}}) < \delta) \rfloor) \tag{7.22}$$

where numel$(\cdot)$ represents the number of elements of its argument and $\lfloor \cdot \rfloor$ denotes the elements that satisfy condition given by its argument. When values are normalized with regard to amplitude, parameter δ is chosen to be near to 1.

In other words, by defining a vector of accumulated value as

$$\boldsymbol{z} = \frac{\text{diag}(\boldsymbol{A}_{\boldsymbol{Q}-\boldsymbol{\alpha}})}{\text{tr}(\boldsymbol{A}_{\boldsymbol{Q}-\boldsymbol{\alpha}})},$$

the value of k is chosen considering that the following is satisfied:

$$\sum_{i=1}^{k} z_i \approx \frac{N}{100}, \tag{7.23}$$

over $N\%$ criterion of accumulated value. Thus, $\delta = N/100$.

7.5 Segment Analysis

Further decreasing of computational load can be reached if sectioning the whole input data into segments for localized processing (segment analysis). An intuitive way to carry out this kind of analysis consist of dividing into N_s subsets, called segments, and later applying a clustering procedure for each segment. Segmented data set is denoted by $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{N_s}\}$, where $\boldsymbol{X}_l$ is a $n_l \times p$ matrix corresponding to the l-th segment, $n_l = \text{round}(n/N_s)$ and round$(\cdot)$ represents the entire nearest to its argument. Using

directly resultant groups per each group does not represent a significant improvement in data processing because it can yield a high number of groups. Therefore, a cluster merger stage must be included into the segment procedure. There are two basic approaches for clusters merger. First one consist of using the whole set of partitions in order to merge all groups from each segment. In the second approach, clusters are formed per segment into a sequential method, being preferable because it represents a real-time oriented sequential scheme.

At the beginning, a proper length of segment to be clustered is estimated. Selection of proper number of localized clustering segments is constrained by following restrictions: twice of number of features must exceed the amount of observations per segment ($n_l \geq 2p$), and the minimum of computational cost should be reached. Then, at the end of grouping step, combination of clustered segments is considered, based on estimation of the proximities between each considered cluster and the remaining clusters. By considering that $\boldsymbol{P}^i = \{\boldsymbol{C}_1^i, \ldots, \boldsymbol{C}_{k^i}^i\}$ is the partition estimated for i–th iteration, where k^i is the number of assumed groups for the same partition and $\boldsymbol{Q}^i = \{\boldsymbol{q}_1^i, \ldots, \boldsymbol{q}_{k^i}^i\}$ are the centroids of i-th cluster, then, combination of clusters follows the next rule:

$$\vartheta(j^i, j^{i+1}) = \vartheta\left(\boldsymbol{q}_{j^i}^i, \boldsymbol{q}_{j^{i-1}}^{i-1}\right) = \mathrm{d}\left(\boldsymbol{q}_{j^i}^i, \boldsymbol{q}_{j^{i-1}}^{i-1}\right) \tag{7.24}$$

that is, if estimated measure $\vartheta(j^i, j^{i+1})$ lies within assumed proximity interval $\vartheta(j^i, j^{i+1}) < \epsilon$, then both considered clusters are to be combined. Otherwise, following comparison of cluster is accomplished. Nonetheless, if there is any cluster not fulfilling the proximity measure during the actual i-th iteration, it is no discarded but considered later during coming next iterations. Therefore, incorrect clustering of minority classes is avoid as well as computational load is decreased.

Algorithm 8 explains the steps of the segment grouping of data.

7.6 Clustering Performance Measures

As cluster validity measure the following clustering index is suggested that is expressed as the relationship f_1/f_2 between the expected value of the objective function f_1, assessed if considering an ideal partition, and the actual value, f_2, estimated for the resultant partition. Objective function can be taken as any described in Section 7.2.

Algorithm 8 Segment grouping

1. Divide the data set: $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{N_s}\}$, where $\boldsymbol{X}_l$ is a $n_l \times d$ matrix and $n_l = \text{round}(n/N_s)$

2. Gather the elements of each one of the segments into k^l groups: $\boldsymbol{P}^l = \{\boldsymbol{C}_1^l, ..., \boldsymbol{C}_{k^l}^l\}$ associated to centers $\boldsymbol{Q}^l = \{\boldsymbol{q}_1^l, \ldots, \boldsymbol{q}_{k^l}^l\}$

For $l = 2$ **until** N_s **do**

3. Compare the centers of each one of the segments: $\vartheta(j^l, j^{l-1}) = \vartheta\left(\boldsymbol{q}_{j^l}^l, \boldsymbol{q}_{j^{l-1}}^{l-1}\right)$, $j^l = 1, \ldots, k^l$, $j^{l-1} = 1, \ldots, k^{l-1}$

If $\vartheta(j^l, j^{l-1}) < \epsilon$

4. The group $\boldsymbol{C}_{j^l}^l$ is jointed to the other group $\boldsymbol{C}_{j^{l-1}}^{l-1}$

otherwise

The group $\boldsymbol{C}_{j^l}^l$ keeps as an independent group in order to analyze following one

End If

End For

This index measure the ratio of objective function change. Since $f_2 \geq f_1$, it can be infered that index is regarded to a proper clustering if its value lies some close to 1. It must be quoted that the proposed above measure is no sensitive to the assumed number of clusters.

On the other hand, as another cluster validity measure to be considered, clustering quality is assessed that is based on spectral graph partitioning [150], when a good clustering desires both tight connections within partitions and loose connections between partitions. Thus, the cluster coherence is calcules as follows:

$$\epsilon_M = \frac{1}{k} \sum_{l=1}^{k} \frac{\boldsymbol{M}_l^{\mathsf{T}} \boldsymbol{A} \boldsymbol{M}_l}{\boldsymbol{M}_l^{\mathsf{T}} \boldsymbol{D} \boldsymbol{M}_l}$$

where $\boldsymbol{M}$ is the matrix formed by the membership values of all elements to each cluster: $m_{ij} = m(\boldsymbol{q}_j/\boldsymbol{x}_i)$, $\boldsymbol{M}_l$ denotes a membership submatrix associated to the cluster l, $\boldsymbol{A}$ is the affinity matrix and $\boldsymbol{D}$ is the degree of matrix $\boldsymbol{A}$. The matrix $\boldsymbol{M}$ is binary, then, when smooth clustering is implemented, the following conversion must be performed, $M_{ij} = \langle \max \arg m(\boldsymbol{q}_j/\boldsymbol{x}_i) \rangle$, $j = 1, \ldots, k_r$, where $\langle \cdot \rangle$ is 1 if its argument is true and 0 otherwise. Due to normalization with respect to the affinity matrix, the maximum value of $\varepsilon_{_M}$ is 1, therefore, it indicates a good clustering if its value is near 1. Furthermore, because of the nature of the function, a large set of groups is penalized.

Nonetheless, this work takes advantage of the data set labels and therefore supervised measures are accomplished. Thus, performance outcomes can be contrasted with another similar works. In particular, each assembled cluster can be split into two clases: one holding the majority elements regarding to the class of interest (IC), and another having the minority elements being of different classes (OC). In general, heartbeats associated to IC correspond to abnormal beats and can appear suddenly in the recording, while OC correspond to normal heartbeats.

Therefore, the following quantitative measures are defined:

- True Positive (T_P), the number of heartbeats IC classified correctly.
- True negative (T_N), the number of heartbeats OC, classified correctly.
- False positive (F_P), the number of heartbeats OC classified as IC.
- False negative (F_N), the number of heartbeats IC classified as OC.

After computing the above described measures, the following values of sensitivity (Se), specificity (Sp), and clustering performance (CP) are estimated as:

$$S_e = \frac{T_P}{T_P + F_N}$$
$$S_p = \frac{T_N}{T_N + F_P}$$
$$C_P = \frac{T_N + T_P}{T_N + F_P + T_P + F_N}$$

The sensibility and specificity quantify the proportion of elements from IC and theOC that are correctly classified, respectively. Both indexes measure the partition quality with respect to ideal case, when the quantity of clusters equates to the number of classes, but each cluster holding just one class.

Nonetheless, there is no ideal partition, i.e., either, the number of clusters is lower than the number of classes, because the variables considered do not discriminate distinct classes or it should be expected more clusters than classes. Besides, some clusters may contain majority and minority elements from another classes. Therefore, the partition might be penalized when holding a relatively large number of groups regarding number of classes, for instance, by means of a factor as

$$e^{-\eta k_r / k_a} \tag{7.25}$$

where k_r is the number of groups resulting from the clustering, k_a, is the admissibility value of groups, and $\eta,\ 0 < \eta \leq 1$, is an adjusting value. In this way, the measure $\mathfrak{m}$ that can be S_e, S_p or C_P is weighted as follows:

$$\mathfrak{m} = \begin{cases} \mathfrak{m} e^{-\eta k_r / k_a}, & k_r > k_a \\ \mathfrak{m}, & k_r \leq k_a \end{cases}$$

The value of η must be greater than 0, and it can be less than 1 for a less rigorous penalization ($0 < \eta \leq 1$).

Table 7.1 shows the compilation of clustering performance measures considered in this study.

Table 7.1: Clustering performance measures

Measure	**Notation**	**Math expression**
Sensitivity	Se	$\frac{TN}{TN+FP} \times 100$
Specificity	Sp	$\frac{TP}{TP+FN} \times 100$
Clustering performance	CP	$\frac{TN+TP}{TN+TP+FN+FP} \times 100$
Objective function ratio	f_1/f_2	$\frac{f_1}{f_2}$
Cluster coherence	$\varepsilon_{_M}$	$\frac{1}{k}\sum_{l=1}^{k} \frac{\boldsymbol{M}_l^{\mathsf{T}} \boldsymbol{A} \boldsymbol{M}_l}{\boldsymbol{M}_l^{\mathsf{T}} \boldsymbol{D} \boldsymbol{M}_l}$

Part III

Experiments and Results

Chapter 8

Experimental Set–Up

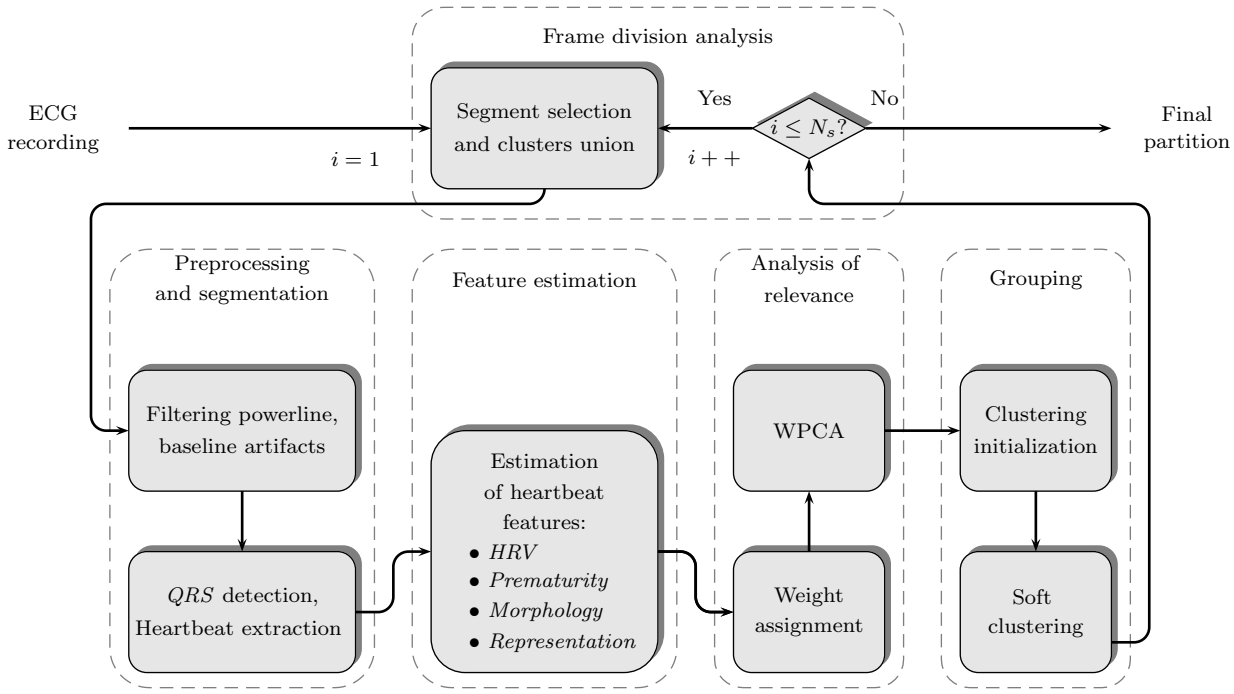


Figure 8.1: Block diagram of proposed unsupervised methodology for Holter monitoring of cardiac arrhythmias.

Figure 8.1 depicts the methodology proposed for Holter arrhythmia analysis that includes the next stages: a) Preprocessing, b) Feature estimation, c) Analysis of relevance, and c) Clustering. As input data, Holter recordings initially are preprocessed to reduce the influence of interferences and artifacts. Next, recordings are segmented based on calculation of *QRS* complex. Heartbeat features, which are estimated us-

ing variability, prematurity, morphology and representation measurements of the QRS complex and heart rate variability, are extracted by weighted linear projection. Lastly, projected data is grouped by a clustering algorithm. Because of restrictions for reducing computational load, the proposed methodology is carried out by framing along the time axis the input data into N_s successive divisions of Holter recordings, where each frame is separately processed. Therefore, according to the assumed criterion of homogeneity between two given consecutive frame divisions, resulting clusters can be either merged or split. In the next sections, the stages of the procedure are described in detail.

8.1 Data Set

8.1.1 ECG Validation Database

The experiments are carried out over the whole set of recordings from the ECG MIT-BIH database, which holds different types of arrhythmia as discussed in Chapter 2. In accordance to the AAMI standard (ANSI/AAMI EC57:1998/(R)2003) [4], the following groups shown in Table 2.2 are of interest to be examined: normal–labeled heartbeat recordings (termed N), Supraventricular ectopic beat (Sv), Ventricular ectopic beat (V), Fusion beat (F), as well as unknown beat class (Q) is taken into consideration. All classes listed are assumed to be present during holter analysis.

The used specific recordings are shown in Table 9.4. Name and total number of beats per arrythmia group are included in the first six columns. As can be seen, some recordings exhibit strong unbalanced number of observation per class. Namely, recording #215 holds just one heartbeat of class F, and 2 of S, whereas its number of normal heartbeats is 3194.

8.2 Preprocessing Procedures

Attained heartbeat set from the ECG signal, $s(t)$, that is subject to discrete time transformation, $\boldsymbol{s} = \{s_k\}$; where $s_k \triangleq s[kT_s]$, being $k \in \mathbb{N}$, and T_s the sampling period, has to be preprocessed. In the beginning, recordings are normalized geometrically [11] to prevent biasing, i.e., $\boldsymbol{s}^0 = (\boldsymbol{s} - \mathbf{E}\{\boldsymbol{s}\})/(|\max\{\boldsymbol{s}\}|)$, where the notation $\mathbf{E}\{\cdot\}$ stands for the expectance operator. Then, unbiased vector $\boldsymbol{s}^0$ is filtered to reduce signal disturbances and artifacts.

Table 8.1: Set of recordings of the MIT/BIH database used in the experiments.

Group	N (#)					S (#)				V (#)		F (#)		Q (#)	
rec./Arr.	**N**	**L**	**R**	**e**	**j**	**A**	**a**	**J**	**S**	**V**	**E**	**F**	**f**	**P**	**Q**
MIT label	**1**	**2**	**3**	**34**	**11**	**8**	**4**	**7**	**9**	**5**	**10**	**6**	**38**	**12**	**13**
100	2237					33				1					
101	1858					3									2
102	99									4			56	2026	
103	2080					2									
104	163									2			666	1378	18
105	2524									41					5
106	1506									520					
107										59				2076	
108	1738				1	4				16		2			
109		2490								38		2			
111		2121								1					
112	2535					2									
113	1787						6								
114	1818					10		2		43		4			
115	1951														
116	2300					1				109					
117	1532					1									
118			2164			96				16					
119	1541									444					
121	1859					1				1					
122	2474														
123	1513									3					
124			1529		5	2		29		47		5			
200	1742					30				825		2			
201	1623				10	30	97	1		198		2			
202	2059					36	19			19		1			
203	2528						2			444		1			4
205	2569					3				71		11			
207		1457	85			106				105	105				
208	1585								2	992		372			2
209	2619					382				1					
210	2421						22			194	1	10			
212	922		1824												
213	2639					25	3			220		362			
214		2000								256		1			2
215	3194					2				164		1			
217	244									162			260	1540	
219	2080					7				64		1			
220	1953					94									
221	2029									396					
222	2060				212	208		1							
223	2027			16		72	1			473		14			
228	1686					3				362					
230	2253									1					
231	314		1252			1				2					
232			396		1	1381									
233	2229					7				830		11			
234	2698							50		3					
Total	74989	8068	7250	16	229	2542	150	83	2	7127	106	802	982	7020	33

Table 8.2: Different wavelet functions to ECG denoising

wavelet(*order*)	*level*
daubechies(1-10)	3–8
coiflet (2-5)	3–8
biorthogonal(1.1, 1.3, 1.5, 2.2, 2.4, 2.6, 2.8, 3.1, 3.3, 3.5, 3.7, 3.9, 4.4)	3–8
symlet(2-9)	3–8

Table 8.3: Working value set of *QRS* detector

Parameter	L_1	L_2	l_s	l_{ov}	m	n	Refractory period
Value	5	4	40	20	8	8	200 ms

Specifically, power line interference is reduced using adaptive sinusoidal interference multiple canceller that is assumed to provide significant signal–to–noise ratio improvement [157]. The parameters $\mu_{a_i} = \mu_{\phi_i} = 0.12$ of (5.15) that control the convergence rate in the ASIMC algorithm, were obtained in a experimental way for optimal localization of interferences.

In order to remove the high-frequency noise such as EMG, some wavelet functions at different levels of decomposition were analyzed. Table 8.2 depicts the families used.

Also, the baseline wandering is canceled out by the method described in [139] taking into account the DWT approximation coefficients $a_i(l)$ of the expression (5.11).

Although the signal is also partially filtered, this preprocessing is assumed not to affect the separability of among the underlying heartbeat groups.

Since the analysis of arrhythmias under consideration is supported on fixed changes of both *QRS* complex as well as the heart rate variability (*HRV*), *R*–peak locations are previously estimated accordingly to the procedure given in Section 5.3 including the following sequential procedures: band–pass filtering, *R* peak enhancement and adaptive thresholding. Parameters of the proposed algorithm were experimentally adjusted to improve the enhancement of R peak detection getting the better performance over considered database, as shown in Table 8.3.

The performance of detector was evaluated using the standard measures of the sensitivity ($Se_{qrs} = TP/(TP + FN)$) and positive predictivity ($P_{qrs} = TP/(TP + FP)$), where true positive (TP) is the total number of *QRS* correctly located by the detector. 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.

Furthermore, to avoid analysis over *QRS* complexes of different length, their seg-

Table 8.4: Feature set considered for Holter monitoring of cardiac arrhythmias.

Index	Type	Description
x_1	HRV and Prematurity [10]	• RR interval
x_2		• pre-RR interval
x_3		• post-RR interval
x_4		• Difference between RR and pre–RR intervals, $x_4 = x_1 - x_2$
x_5		• Difference between post–RR and RR intervals, $x_5 = x_3 - x_1$
x_6		• Continuous APB^* heartbeat type, (eq. 8.2)
x_7	Morphology and representation	• QRS matching by Dynamic time warping
x_8	[8], [10], [9], [136]	• Polarity of QRS complex
x_9		• Energy of QRS complex
$x_{10}, \ldots, x_{19}$		• First 10 Hermite–based coefficients
$x_{20}, \ldots, x_{90}$		• Db2 (A4: $20-25$, D4: $26-31$, D3: $32-41$, D2: $43-58$, D1: $59-90$)
$x_{91}, \ldots, x_{100}$		• var$\{A4, D4, D3, D2, D1\}$, max$\{A4, D4, D3, D2, D1\}$

* The notation APB stands for Atrial Premature Beat, being a sort of S heartbeats.

mentation is carried out for a fixed window length, that is, each j–th complex $\boldsymbol{d}_j$ is accomplished as follows,

$$\boldsymbol{d}_j = \{s_k^0\}; \forall k \in [l_j - aF_s, l_j + bF_s], \tag{8.1}$$

where l_j is the R–peak time location of the j-th heartbeat and $F_s = 1/T_s$ is the sampling frequency. Nonetheless, it must be quoted that some morphologies might exhibit S–waves lasting exceptionally more than usual, and therefore they can be missed if using such a short processing window, then, as usually recommend, QRS width is fixed to be of 200 ms length, i.e., $a = b = 0.1$.

8.3 Feature Estimation

Heartbeat characterization is achieved by taking into consideration the wide set of features that had been proposed early for arrhythmia analysis over Holter ECG recordings [8–10, 136]. Usually, the whole set of studied features can be divided into the following groups, as shown in Table 8.4:

8.3.1 Prematurity and Variability based Features.

When considering S labeled arrhythmias, their morphology looks highly similar to the normal heartbeat shape, and therefore, the following set of features, which are extracted from variability of cardiac rhythm, is mainly considered [10]:

– *HRV–derived features* (x_1, x_2, x_3): Interval parameters providing information about

sequences of heartbeats with unusual timing, namely [4]:

$$\begin{aligned} x_1 &= l_j - l_{j-1}, && (RR \text{ interval}) \\ x_2 &= l_{j-1} - l_{j-2}, && (pre-RR \text{ interval}) \\ x_3 &= l_{j+1} - l_j, && (post-RR \text{ interval}) \end{aligned}$$

It should be remarked that atrial (S) and ventricular (V) ectopic beats manifest abrupt changes on fiducial point intervals, which in turn, affect the respective values of heartbeat interval features.

– *Prematurity features* (x_4, x_5, x_6): Defined parameters, $x_4 = x_1 - x_2$ and $x_5 = x_3 - x_1$, are assumed to be relevant since they make possible the identification of S arrhythmia type, when reflecting the increase or decrease of the heart rate. Besides, if any heart beat occurring after another S–labeled event is regarded as normal, the above couple of features will change of sign. Feature x_6 accounts for the number of consecutive S that is also sensitive to an increase of the heart rate, exceeding the normal range set for x_4. The parameter x_6 is expressed as follows:

$$x_6 = \left(\frac{x_3}{x_1}\right)^2 + \left(\frac{x_2}{x_1}\right)^2 - \left(\frac{1}{3}\sum_{i=1}^{3} x_i^2 \log(x_i)^2\right). \tag{8.2}$$

Besides, the first and second squared terms in (8.2) are sensitive to abrupt changes of heart rate, whereas, the last addend is inferred as unnormalized Shannon entropy, which increases the value of x_6 whenever heart rate is steadily increasing.

8.3.2 Morphological and Representation Features $(x_7, \ldots, x_{100})$.

Since most of analyzed arrhythmias change the shape of QRS complex, their characterization can be attained by commonly used time and spectral–based techniques [130]. Therefore, regarding the former techniques, the following features are worth to be considered: A couple of features that are sensitive to abnormal QRS complexes: x_7, which computes a morphological dissimilarity by means of Dynamic Time Warping (DTW) approach between current QRS complex, and a linearly averaged QRS complex of the last n heartbeats [130]. Next, a parameter, which is sensitive to ventricular arrhythmias exhibiting abnormal QRS complexes such as ventricular extrasystoles (V) or branch blocks (N) [9], is defined as $x_8 = |\max\{\boldsymbol{d}_j\}/\min\{\boldsymbol{d}_j\}|$, being $\boldsymbol{d}_j$ the current QRS complex. In addition, since the morphological notoriety of branch block heartbeats, the

QRS energy, which is a straightforward feature to detect previously described type of heartbeats, is estimated as $x_9 = \sum_{i=1}^{L_d} d_j[i]^2$, where L_d is the processing length of the j-th *QRS* complex.

On the other hand, spectral–based representation features that have been used in the field of signal compression are also taken into account, because only few coefficients are needed to reconstruct the signal [8]. In this line of analysis, the Hermite coefficient are used and can be computed as described in section 5.4.2. The elements of Hermite base are ranged in the interval $(-t_0, t_0)$ with $t_0 = 100$ ms, in order to set the length of window to be 200 ms.

As discussed in the Section 5.4.1, wavelet decomposition coefficients are also considered. Specifically, 4–level coefficients of Daubechies–2 class (*dB2*) are computed, which had been proved to describe properly different heartbeat morphologies, as discussed in [136]. The following statistical descriptors of decomposition coefficients are calculated: mean value, variance, and maximum values are estimated.

As a result, given an i-th observation heartbeat, the respective feature vector $\{\boldsymbol{x}_i \in \mathbb{R}^p : i = 1, \ldots, n\}$, where $p = 100$, is assumed to be the input training space toward arrhythmia classification purpose.

8.4 Analysis of Relevance Procedures

In Section 6.3, an iterative algorithm to calculate the relevance matrix from feature space, which converges before a certain number of iterations is reached. In [3] is suggested as proper number of iteration as $r \leq 5$ (see algorithm 1) and is also demonstrated that algorithm converges to an optimal value of objective function (6.16). Experimentally, the fact of value of r was assessed over artificial data reaching the convergence in 4 iterations [148].

Algorithm 1, in step 2, requires to calculate the elements g_{ij} of matrix $\boldsymbol{G} \in \mathbb{R}^{p \times p}$. This calculation can signify a high computational cost when a great number of features p is considered, besides matrix $\boldsymbol{G}$ is computed per iteration. In order to improve this procedure, here it is proposed an element-to-element matricial operation, avoiding the calculation per matrix element, following the next expression,

$$\boldsymbol{G} = \text{times}(\boldsymbol{A}, \boldsymbol{B}), \tag{8.3}$$

where $\text{times}(\cdot, \cdot)$ is an array-wise multiplication, it is to say, element-by-element product

Algorithm 9 Fast Power-embedded $Q-\alpha$ method

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 $\boldsymbol{G}^{(r)} = \text{times}((\boldsymbol{M}\boldsymbol{M}^{\mathsf{T}}), (\boldsymbol{M}\boldsymbol{Q}^{(r-1)}\boldsymbol{Q}^{(r-1)\mathsf{T}}\boldsymbol{M}^{\mathsf{T}}))$
3. Compute $\boldsymbol{\alpha}^{(r)}$ as the eigenvector associated with the major eigenvalue of $\boldsymbol{G}^{(r)}$.
4. Compute matrix: $\boldsymbol{A}_{\boldsymbol{\alpha}}^{(r)} = \boldsymbol{M}^{\mathsf{T}}\text{diag}(\boldsymbol{\alpha}^{(r)})\boldsymbol{M}$
5. Compute the orthonormal transformation: $\boldsymbol{Z}^{(r)} = \boldsymbol{A}_{\boldsymbol{\alpha}}^{(r)}\boldsymbol{Q}^{(r-1)}$
6. Compute QR decomposition: $[\boldsymbol{Q}^{(r)}, \boldsymbol{R}] = \text{qr}\ (\boldsymbol{Z}^{(r)})$
7. Make $r \leftarrow r+1$ and return to the step 2

of the arrays $\boldsymbol{A}$ and $\boldsymbol{B}$. By letting $\boldsymbol{A} = \boldsymbol{M}\boldsymbol{M}^{\mathsf{T}}$ and $\boldsymbol{B} = \boldsymbol{M}\boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{M}^{\mathsf{T}}$.

By replacing, previous calculation in algorithm 1, it can be re-written as shown in algorithm 9.

Relevance analysis methods are assessed carrying out the procedure described in Sections 6.2 and 6.3. Variants of the same methods are also considered, namely, taking into consideration only the first eigenvector (i.e., $p = 1$) in case of method described in 6.2 and a free parameter scheme of method described in Section 6.6.

Relevance analysis results can be used to project data employing the procedure described in algorithm 6.7.

8.5 Clustering Procedures

8.5.1 Estimation of number of groups

Estimation of number of groups methods are evaluated in two schemes. First scheme consists of analyzing the whole data set, denoted as $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, where n is the number of heartbeats into recording and p is the number of features ($p = 100$). Second one, consist of estimating the number of groups of selected data using $Q-\alpha$ algorithm. In this case, selected data matrix, that is weighted by relevance value of each feature, is $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{n\times q}$, where q is the number of relevant features.

For each scheme, the number of groups is computed varying the number of segments, first $N_s = 1$ (i.e, analyzing the whole recording) and secondly $N_s = 6$ (i.e., by dividing the recording into 6 parts and estimating per each one of them).

By dividing the recording can be achieved a better local estimation of number of

groups and also computational cost can be decreased, being advantageous, in particular, in case of algorithms that use iterative procedures.

Used methods are described in Section 7.4, specifically, those explained in Section 7.4.1, 7.4.2 and 7.4.4. Method given by Section 7.4.3 is not considered, because applies the same principle as algorithm 7.4.2 providing similar results, but it spent more processing time because of the computation of eigenvectors.

In the case of the SVD estimation method, it is considered that α_{svd} from equation (7.19) is directly proportional to the number of groups, i.e., when decreasing α_{svd} the estimated value could be 1 and vice versa, that is when increasing its value the number of groups can be considerably increased. Experimentally, it was proved that a proper value of parameter α_{svd} is 0.6.

For eigenvalues based estimation (Section 7.4.2), in order to avoid the fact of strong changes in the data matrix that can affect the normalization of the trivial affinity matrix, is introduced a soft affinity matrix $\boldsymbol{A}$ type exponential of the form:

$$a_{ij} = \mathrm{e}^{-d^2(x_i,x_j)/(\sigma_i\sigma_j)}, \quad a_{ii} = 1,$$

where $\sigma_i = \mathrm{d}(x_i, x_n)$ is the N-th nearest neighbor. This affinity measure is widely described in [156].

Finally, in the case of $Q-\alpha$ algorithm based estimation, the only parameter to be set is the number of iterations r from algorithm 1, which is advisable to be $r \leq 5$, as explained in the relevance analysis scheme from Section 8.4.

8.5.2 Segment analysis

According to the described in Section 7.5, data space is divided into N_s equal segments, which are clustered using techniques discussed in Chapter 7. Then, it must estimated the proper number of segments taking into account that a good trade-off between quality of partition and computational cost can be reached. Carrying out this process, some tests are performed to assess the partitions by varying the parameter $N_s = 1, \ldots, N_{s_{max}}$ and measuring, in each case, the partition quality and the computation time. The reference unit TU for the time used to analyze whole recording without any division.

By following the general scheme for cluster union described in algorithm 8, it is proposed a method for 2-contiguous segment union as shown in Figure 8. To that end, segments correspond to partitions denoted as $\boldsymbol{P}^A = \{\boldsymbol{C}_1^A, \ldots \boldsymbol{C}_m^A\}$ and $\boldsymbol{P}^B =$

$\{\boldsymbol{C}_1^B, \ldots \boldsymbol{C}_n^B\}$, where the first one (A) is the accumulated grouping from first segment until current, and partition B is the corresponding to segment to be merged. Term $\boldsymbol{C}_i^j$ is the i-th cluster from j-th partition and $\boldsymbol{q}_i^j$ is its corresponding center. This method is described in algorithm 10.

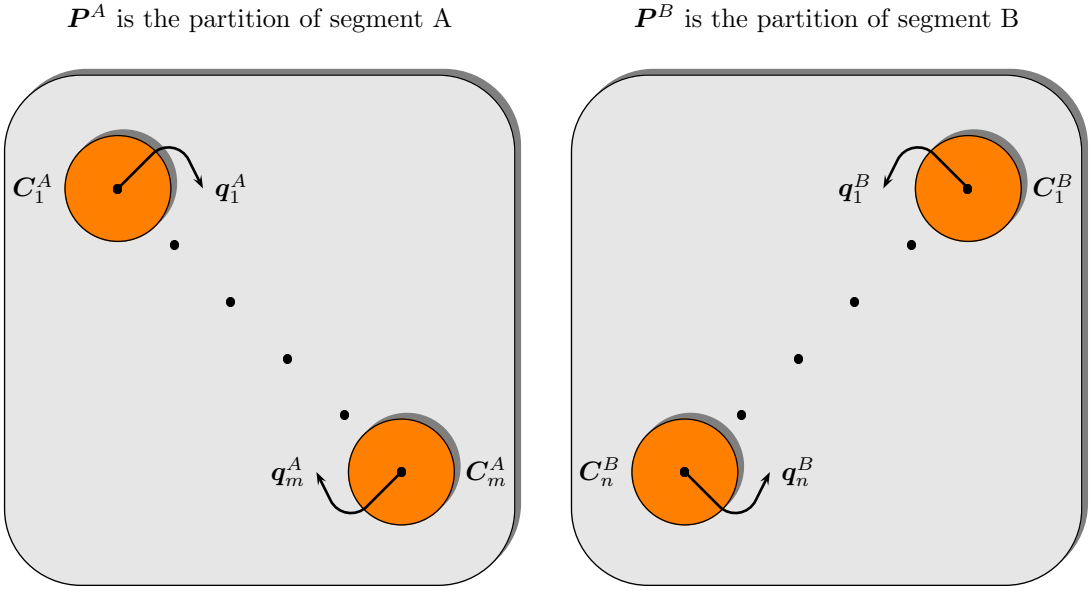


Figure 8.2: Two contiguous segments and their corresponding clusters

8.6 General Methodology

Following results are obtained by employing the full general methodology. It is the whole feature set. Feature selection stage is carried out using $Q-\alpha$ algorithm (Section 6.3) and MSE-based approach (Section 6.2). Clustering is applied over projection of weighted data as is described in Section 6.7. As clustering method, NPDBC with Parzen estimation is applied. Performance indices are those described in Section 7.6, besides the number of resultant groups k_{end} and the number of relevance features p. Value of $\varepsilon_{\scriptscriptstyle M}$ is computed by employing resultant matrix from feature selection stage, in case of $Q-\alpha$, and it is used the trivial affinity measure in case of MSE-based feature selection. Measures Sp and Se are penalized by applying equation (7.25), with $\eta = 0.05$ and $k_a = 12$.

Algorithm 10 Segment grouping algorithm

Objective {To achieve the union between two contiguous segments that contain a set of clusters}

Inputs $\boldsymbol{P}^A = \{\boldsymbol{C}_1^A, \ldots, \boldsymbol{C}_m^A\}$, $\boldsymbol{P}^B = \{\boldsymbol{C}_1^B, \ldots, \boldsymbol{C}_n^B\}$, $\boldsymbol{q}_i^x$, m and n. {$\boldsymbol{P}^A$ and $\boldsymbol{P}^B$ correspond to the couple of input partitions, $\boldsymbol{q}_i^x$ corresponds to the center associated with i-th cluster from segment x, m and n correspond to the number of clusters in the segments B and A. respectively. }

Output $\boldsymbol{P}_r$ {Output partition}

Variables $\boldsymbol{u}$, $\boldsymbol{g}$, ϵ { $\boldsymbol{u}$ represents the indices of centroid of $\boldsymbol{P}^B$ that fulfil some conditions, $\boldsymbol{g}$ is the vector of distances among centroid of $\boldsymbol{P}^A$ and $\boldsymbol{P}^B$, ϵ represents a decision threshold to merge two clusters. }

Method:

$k = 1$; {counter for new clusters in $\boldsymbol{P}^A$}
$\epsilon \leftarrow \min\{\min(d(\boldsymbol{q}_i^A, \boldsymbol{q}_j^A)), \min(d(\boldsymbol{q}_i^B, \boldsymbol{q}_j^B))\}$, $\forall i, j$ and $i \neq j$
while $j \leq n$ **do**
  while $i \leq m$ **do**
    $g[i] \leftarrow d(\boldsymbol{q}_i^A, \boldsymbol{q}_j^B)$ {calculates the distance between two centroids.}
    $i++$
  end while
  if $\min(\boldsymbol{g}) < \epsilon$ **then**
    $\boldsymbol{C}_{\min(\boldsymbol{g})}^A \leftarrow \boldsymbol{C}_{\min(\boldsymbol{g})}^A \cup \boldsymbol{C}_j^B$ { Clusters are merged. }
    $\boldsymbol{q}_{\min(\boldsymbol{g})}^A \leftarrow F\{\boldsymbol{q}_{\min(\boldsymbol{g})}^A, \boldsymbol{q}_j^B\}$ { The new centroid is recalculated taking into account the centroid $\boldsymbol{q}_j^B$.}
  else
    $u[k] \leftarrow j$
    $k++$
  end if
  $j++$
  $i \leftarrow 1$
end while
$\boldsymbol{P}_r \leftarrow \boldsymbol{P}^A \cup \{\forall u, \boldsymbol{C}_{\boldsymbol{u}}^B\}$ { The clusters of segment B that do not satisfy the condition are added to the partition $\boldsymbol{P}^A$. }

Chapter 9

Results and Discussion

In this chapter, results of the proposed methodology are presented and discussed. They are obtained following the scheme shown in Figure 8.1, that encompass filtering, R-peak estimation, feature extraction, relevance analysis and clustering stages.

9.1 Preprocessing and Feature Estimation

9.1.1 ECG filtering

Adaptive filtering

Figures 9.1 and 9.2 exhibit the ability of the ASIMC adaptive filter to remove interferences. The test signal is contaminated with artificial noise that has the following characteristics:

- SNR: -6dB(i.e. approximately 3 times the value of signal amplitude).
- Interference frequencies: 60, 120 y 180 Hz. (interference and 2 harmonics)
- Frequency offset: $\pm 60 mHz$.
- Phase offset: Normal distribution $N(0, 1)$.

In Figure 9.1 the original signal (a), noisy signal (b) and filtered signal (c), are shown. Likewise, in subfigures (a), (b) and (c) from Figure 9.2, amplitude spectra for original, noisy and filtered signal respectively are presented.

Quantitatively, it is observed that the filter performance does not affect the fundamental components of signal, by considering that the signal from Figure 9.1 presents

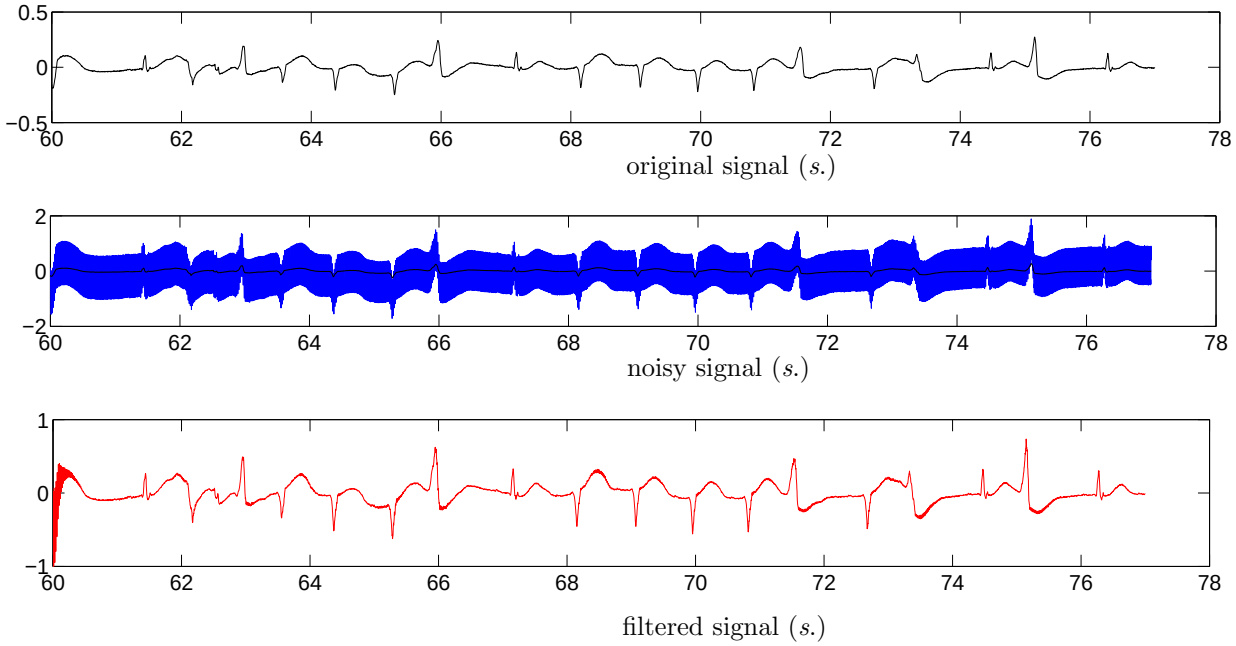


Figure 9.1: Original, noisy and filtered signal, corresponding to recording 207 in the interval between 60 and 77 *s.* (Channel 0).

three kind of arrhythmias (Left bundle branch block (L), right bundle branch block (R) and ventricular ectopic beat (V)). The same fact can also be observed in Figure 9.2, where the amplitude spectrum is similar to the original signal, except at the beginning while algorithm is adapting.

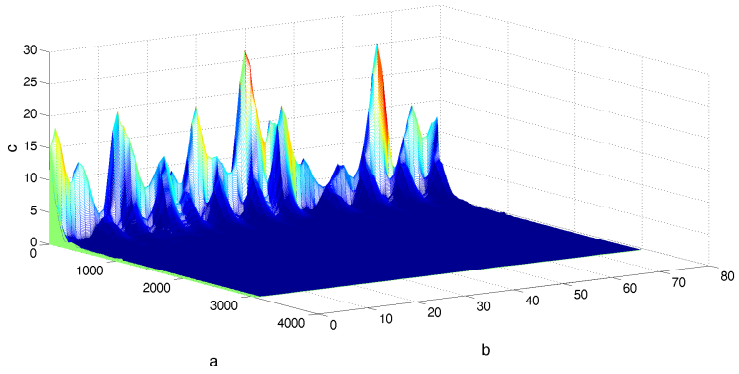
To quantify the filter effectiveness, some distortion measures well-known in literature, are employed [56]. They measure the signal-to-noise ratio between the original and filtered signals and correspond to:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}|x(i) - \tilde{x}(i)|^2 \tag{9.1}$$

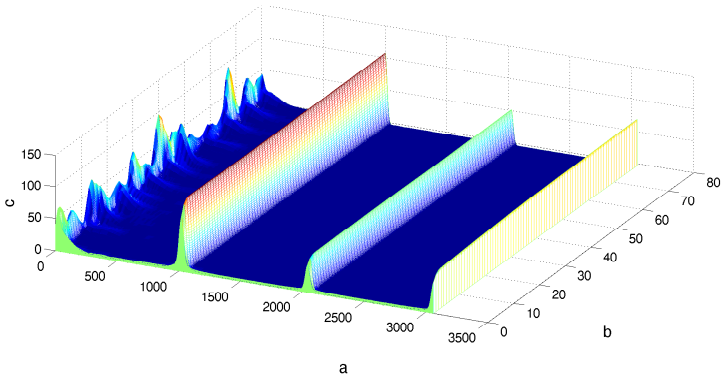
$$\text{PRD} = \sqrt{\frac{\sum_{i=1}^{N}(x(i) - \tilde{x}(i))^2}{\sum_{i=1}^{N}x^2(n)}} \times 100 \tag{9.2}$$

The MSE measure (9.1) is employed as a weighting factor for signal approximation that is represented by orthogonal expansions. As value of MSE is lower, adjustment between signals is better.

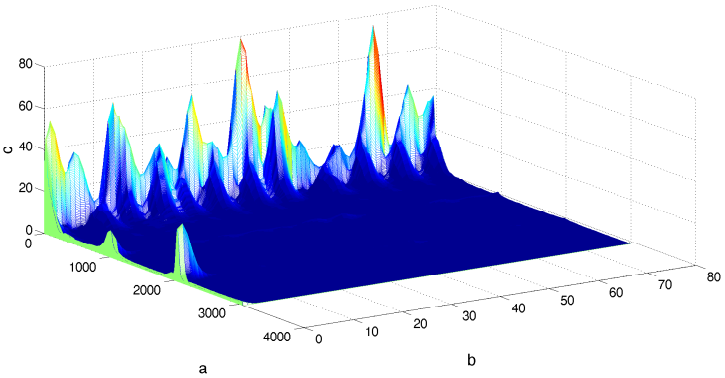
Meanwhile, PRD (9.2) is a distortion measure commonly used in data compression. This measure represents a relation between error energy and original signal energy.



(a) original spectrum



(b) contaminated spectrum



(c) reconstructed spectrum

Figure 9.2: Amplitude spectrum of the original, noisy and filtered signals.

Then, as value of PRD is lower, there exist a lower distortion.

Table 9.1 shows the quantitative difference between original and filtered signal according to standard error measures, such as MSE or PRD [56], by analyzing the channel 0 of the whole MIT/BIH database described in Section 8.1.

Table 9.1: Performance of the powerline filtering algorithm with respect to other algorithms

Measure	ASIMC	IE [56]	Digital Filter [56]
MSE $(\mu - \sigma)$	2.85E-40 – 2-74E-4	5.61E-4 – 4.26E-4	0.022 – 0.008
PRD $(\mu - \sigma)$	11.26 – 3.42	15.58–3.89	105 –10.94

In the Table 9.1, it can be noted that the adaptive algorithm presents good performance regarding other reference algorithms from literature [56]. Because of its easiness for calculation, each algorithm sampling interval requires $3M$ additions, $5M$ multiplications and $2M$ look-up operations to update the parameters estimates, where M is the number of harmonics of interference (See Eq. 5.13).

The proposed filter has the advantage of estimating interference signal harmonics, as well as, slight variations of frequency and phase values, making it a good choice to filter powerline interferences.

Moreover, an adaptation phase approximately equal to 1 s. is required by the algorithm, which is determined by the parameters $\mu_{a_i} = \mu_{\phi_i} = 0.12$ of (5.15) that control the convergence rate in the ASIMC algorithm.

It should also be noted that Holter recordings are less likely to be affected by power line, because the most time Holter recorders are powered by batteries; however interference from near electrical apparatus can significantly affect the signal quality.

Wavelet filtering

The mother wavelet selection is a fundamental stage in denoising approach, then, to carry out the EMG filtering process, wavelet functions depicted in Table 8.2 were evaluated. After experiments, it was proved that wavelet *coiflet* 2 at third level of decomposition with rigorous thresholding, presented better performance results.

For baseline wandering, the method proposed in [139] (see Section 5.2.2) was followed. In that method, the wavelet mother *Daubechies* for levels between 8 and 11, was used.

The Figure 9.3 shows an example of baseline wandering and EMG noise filtering in an ECG signal contaminated with artificial noise taking into account the following constraints:

- SNR of EMG noise: 18dB
- SNR of baseline wandering: 0dB

– Baseline frequency: 0.3 Hz

– Baseline phase and EMG noise: Normal distribution with zero mean and unit variance $N(0, 1)$.

Since the signals from database have inherent both baseline wandering and EMG noise, resultant filtered signal might present slight differences with respect to the original signal. Therefore, with the aim to assess quantitative and qualitatively the filter performance, synthetic signals are employed from a model described in [158]. This dynamic model consists of a three dimensional state equation which parameters are tuned in order to get an undisturbed signal.

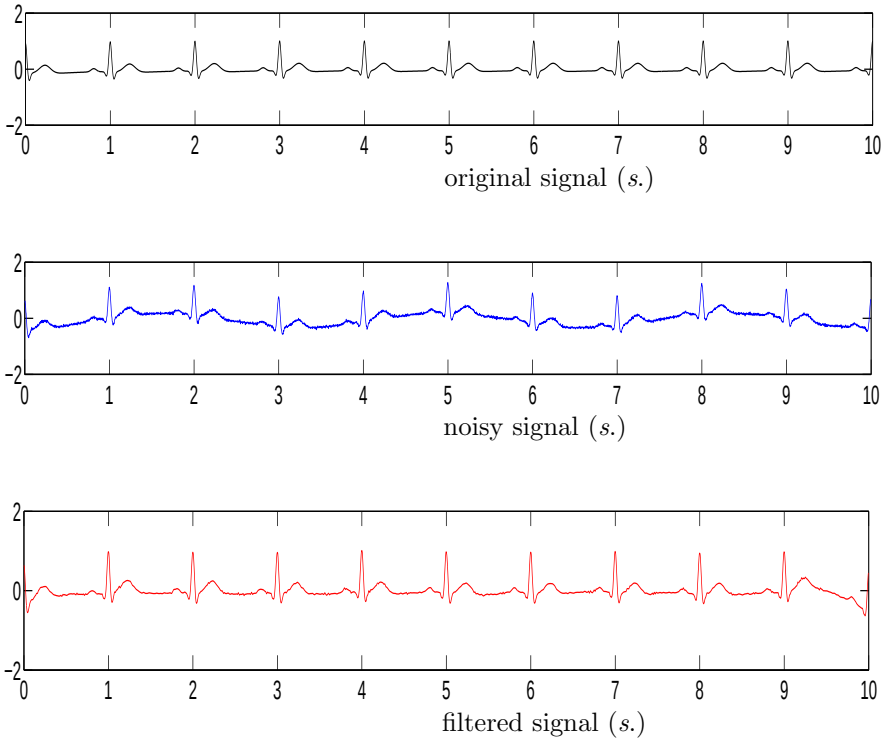


Figure 9.3: Original, noisy and filtered signal, corresponding to a synthetic ECG signal.

Nevertheless, to indicate the filter effectiveness over real signals, Figure 9.4 shows an example of applying the filter on a segment of recording 217 from the database.

Table 9.2 depicts values of the distortion measures ((9.1) and (9.2)) between the synthetic and original signal, employing wavelet filtering and other reference methods [11].

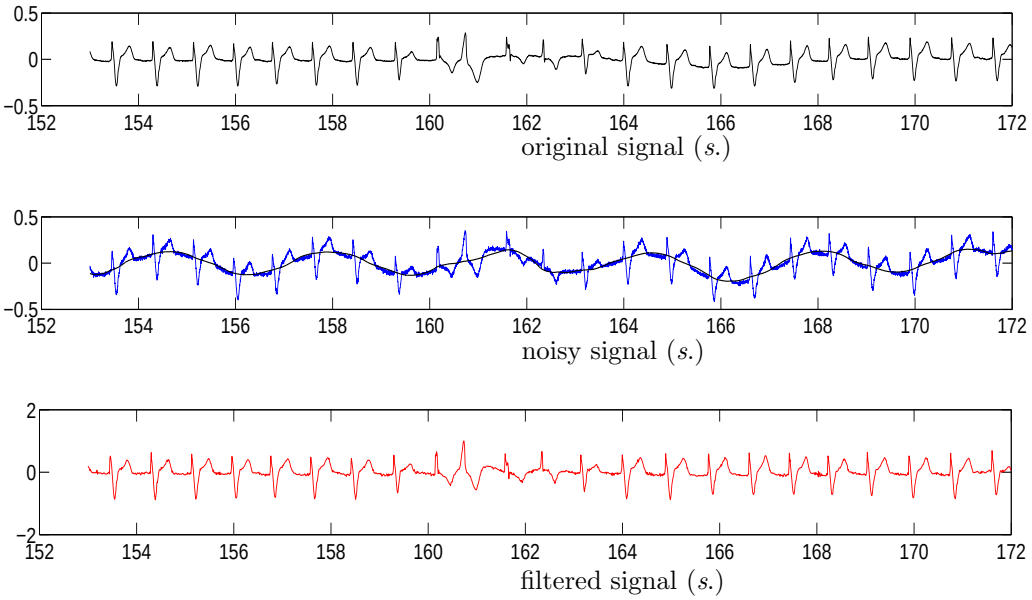


Figure 9.4: Original, noisy and filtered signal, corresponding to recording 217 in the interval between 153 and 172 *s*.

Table 9.2: Wavelet filtering performance with respect to other works

Measure	WT-based	Adaptive filter [11]	Digital Filter [11]
ECM	3.68E-5	6.84E-4	0.002
PRD	4.53	8.24	20.12

The Table 9.2 demonstrates that wavelet-based method presents better performance than other reference algorithms published in literature.

The adaptive wavelet denoising based on Donoho's estimator at different levels and types, gives better performance than classical techniques, and permits to estimate a wide range of baseline wander fundamental frequency as well as EMG noise. However the disadvantage of the approach is the computational cost if the ECG is analyzed in realtime.

9.1.2 *R*-peak detection

Figures 9.5 and 9.6 show the result obtained after ECG signal processing to the R fiducial point estimation, as described in Section 5.3. Figures corresponding to the original ECG signal (a), bandpass filtered signal (b), Shannon energy envelope with

square window (c) and Shannon energy (d), over two segments from recording 217.

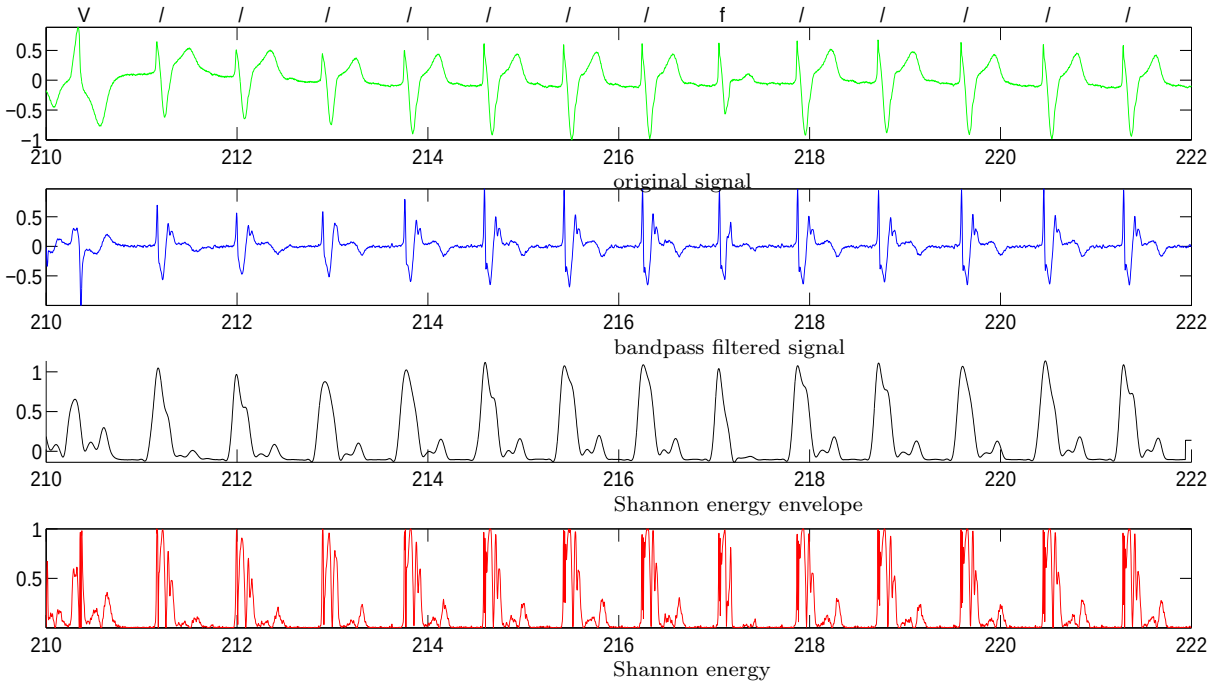


Figure 9.5: Original (a) and filtered (b) signal, Shannon energy envelope with rectangular window (c) and Shannon energy (d). The signal corresponds to recording 217 in the interval between 210 and 222 *s*.

In Figures 9.5 and 9.6, it can be observed the importance of using an analysis window (see Eq. 5.33) to estimate the search region of R-peak in the nonlinear transformation (9.5(c)) regarding calculating the Shannon energy without using window analysis (9.5(d)). The last one can be affected by noise and high frequency components present in the signal.

Figures 9.7 and 9.8 show an example where it is remarkable the advantage of using the Shannon energy envelope with square window (see 9.7 (c)) regarding common transformations as signal energy (9.8 (d)). Test signals correspond to two segments from recording 217. It can be seen the effect of non-linear transformation on R-peaks, which gives greater weighting to peaks with lower amplitude and viceversa (see 9.7(d), between 211 and 213 *s* and 9.8(d), between 845 and 848 *s*).

Performance of the detector was evaluated using Se_{qrs} and P_{qrs} (Section 8.2) standard measures, giving as general results $Se_{qrs} = 99.71\%$ and $P_{qrs} = 99.49\%$.

Results per recording are described in the last two columns of Table 9.4. The

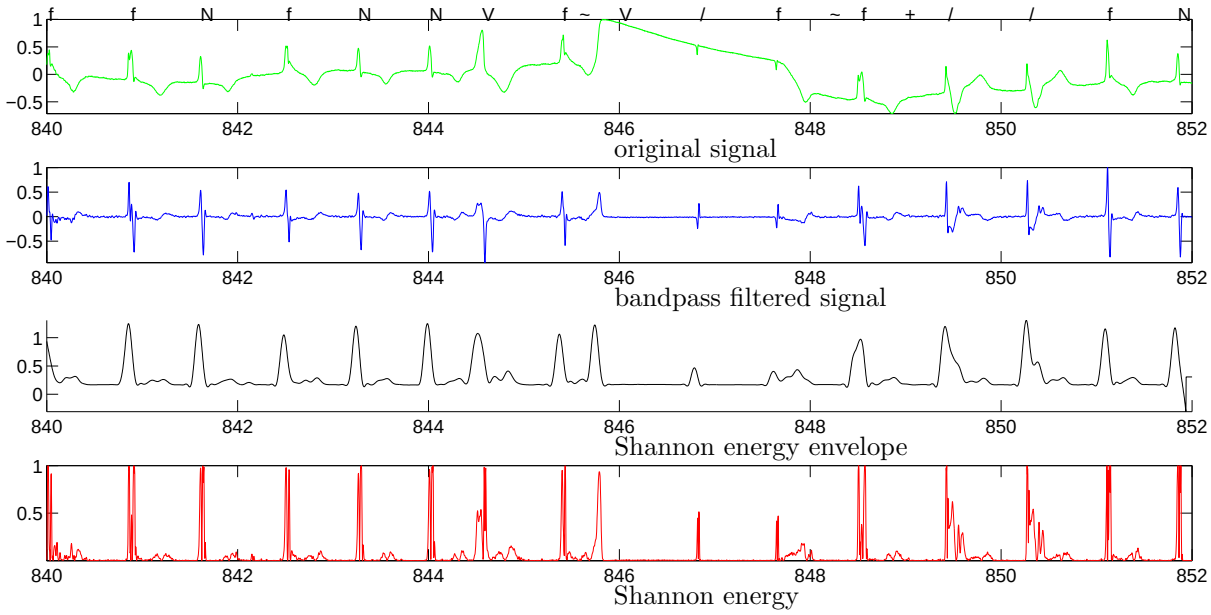


Figure 9.6: Original (a) and filtered (b) signal, Shannon energy envelope with rectangular window (c) and Shannon energy (d). The signal corresponds to recording 217 in the interval between 840 and 852 *s*.

performance of the detector is comparable regarding other works in the literature, as shown in Table 9.3.

Table 9.3: Summary of the total performance of the designed *QRS* detector and some reference *QRS* detectors on the MIT/BIH arrhythmia database.

Detector	$Se_{qrs}(\%)$	$P_{qrs}(\%)$
This work	99.71	99.49
Poli [159]	99.60	99.64
Paoletti [133]	99.65	99.48
Okada's [48]	98.32	98.34
Engelese and Zeelemberg's [160]	98.42	98.39

All paremeteres (Table 8.3) of the *R*-peak algorithm were fitted to obtain an optimal performance taking into account the sensibility and predictivity measures (Section 8.2), which do not vary linearly concerning the parameters.

The last stage based on interval-dependent threshold, takes into account, higher T-wave amplitude and lower *QRS*-complex amplitude, improving the efficiency of *QRS* detector.

Non-linear transformation proposed in this work improves the *R*-peak detection

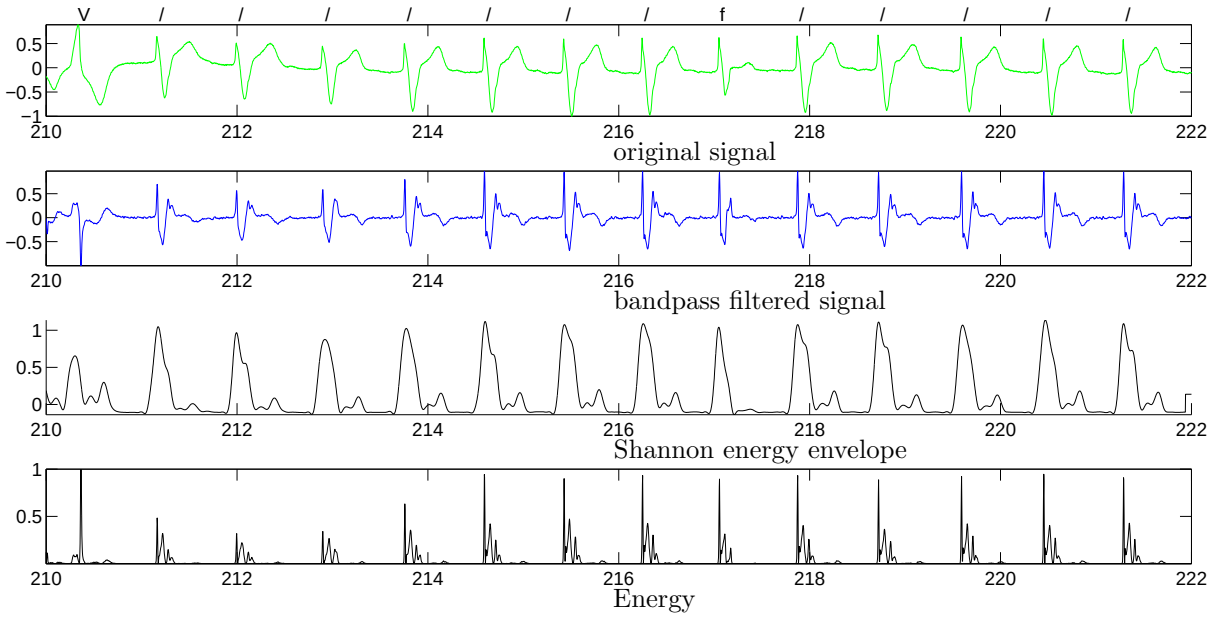


Figure 9.7: Original (a) and filtered (b) signal, Shannon energy envelope with rectangular window (c) and energy (d). The signal corresponds to recording 217 in the interval between 210 and 222 *s*.

in case of the low amplitude complexes could be detected as false negative (FN), increasing in this way the sensitivity of algorithm Se_{qrs}.

9.1.3 Characterization results

Hermite based characterization results

Feature estimation does not require fitting of parameters excepting the Hermite coefficients ($x_{10} \dots x_{19}$, Table 8.4). For Hermite model should be calculated the optimal value of scale-parameter σ_{opt} (eq. 5.38), which is found through a dissimilarity measure between original and reconstructed signal. The dissimilarity measure corresponds to dynamic time warping (DTW) analyzed in [9].

Experimentally, from optimization process described in 5.4.2, it was found that $\sigma = 25 \pm 5$ ms and $6 \leq N \leq 12$ are enough to represent the complexes. Figure 9.9 shows an example of original and reconstructed heartbeats extracted from the recordings 100 and 207 with $\sigma = 25$.

For all cases in this study, the difference between original spectrum of original signal and spectrum of reconstructed signal with $N \geq 6$ and $\sigma = 25$ ms is reasonably small

Table 9.4: Performance of the QRS detector using the recordings from the MIT/BIH database.

Performance of detector	**R-peak detector performance**	
recording		
measure	$Se_{qrs}(\%)$	$P_{qrs}(\%)$
100	100.00	100.00
101	100.00	99.23
102	100.00	100.00
103	99.65	99.03
104	100.00	99.25
105	99.48	99.24
106	100.00	99.34
107	99.25	99.23
108	99.67	99.82
109	99.65	99.34
111	99.54	100.00
112	99.47	99.59
113	99.49	97.59
114	99.69	99.00
115	99.81	98.80
116	100.00	99.55
117	99.47	99.23
118	100.00	99.17
119	100.00	99.78
121	99.62	99.44
122	99.45	99.32
123	99.74	100.00
124	99.96	99.54
200	99.33	99.43
201	99.68	99.40
202	99.94	100.00
203	99.54	99.37
205	100.00	100.00
207	99.56	99.66
208	99.59	99.68
209	99.38	99.64
210	99.45	99.64
212	99.49	99.28
213	99.45	99.24
214	99.65	99.08
215	100.00	99.44
217	99.78	99.33
219	100.00	100.00
220	99.23	99.62
221	99.70	99.36
222	99.07	99.51
223	99.89	99.40
228	99.84	100.00
230	100.00	99.75
231	99.65	99.71
232	100.00	99.65
233	99.74	99.61
234	100.00	100.00
μ(perf.).	99.71	99.49

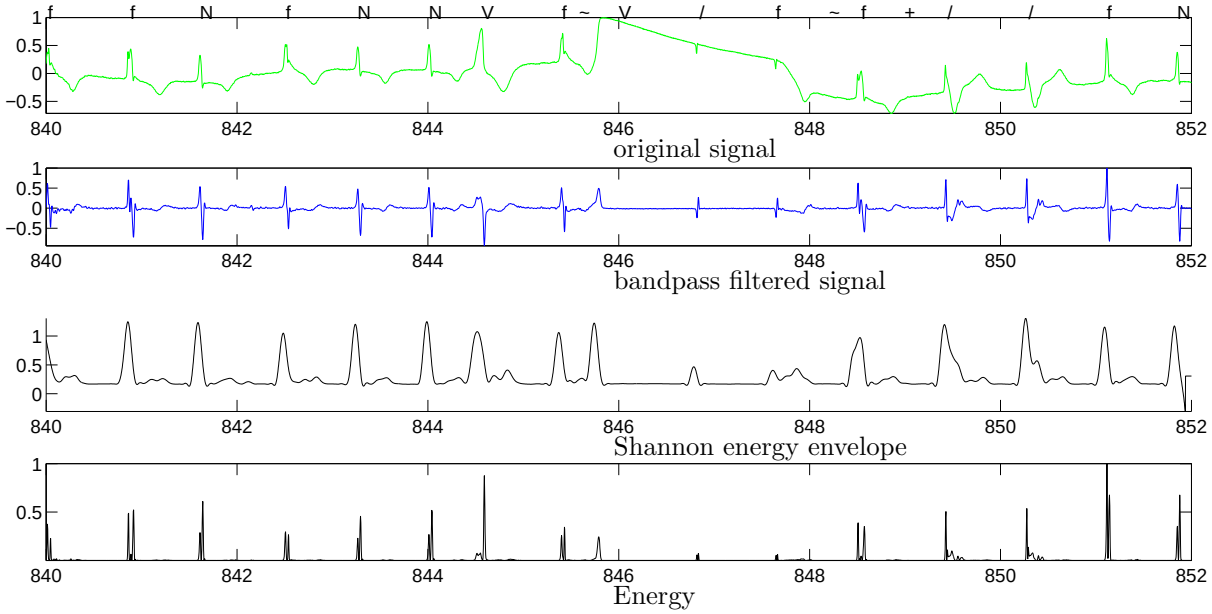


Figure 9.8: Original (a) and filtered (b) signal, Shannon energy envelope with rectangular window (c) and energy (d). The signal corresponds to recording 217 in the interval between 840 and 852 *s*.

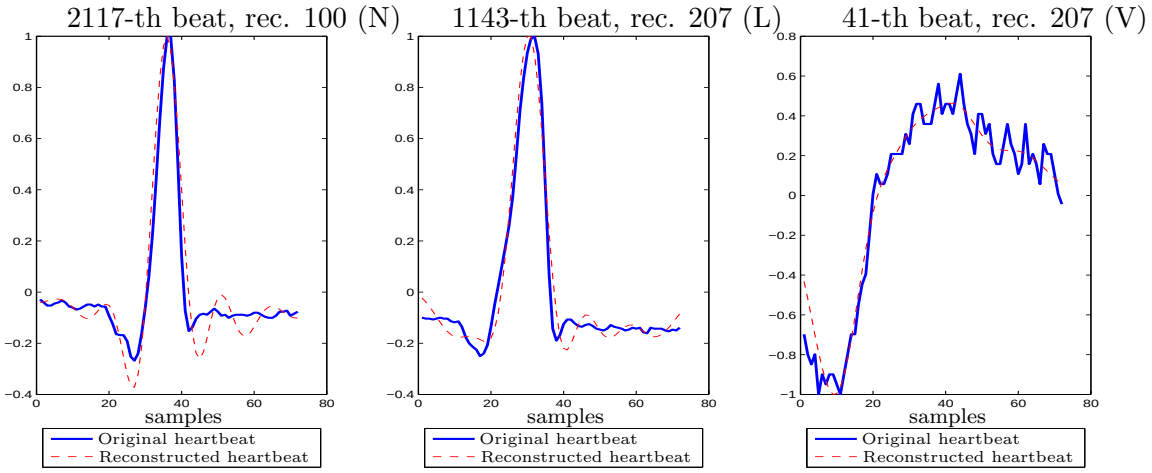


Figure 9.9: Original and reconstructed heartbeats with 10 first Hermite coefficients and $\sigma = 25$

(see Figure 9.11).

Figures 9.10 and 9.11 show, respectively, the spectrum of reconstructed signal using $N = 11$ and $N = 9$ respectively.

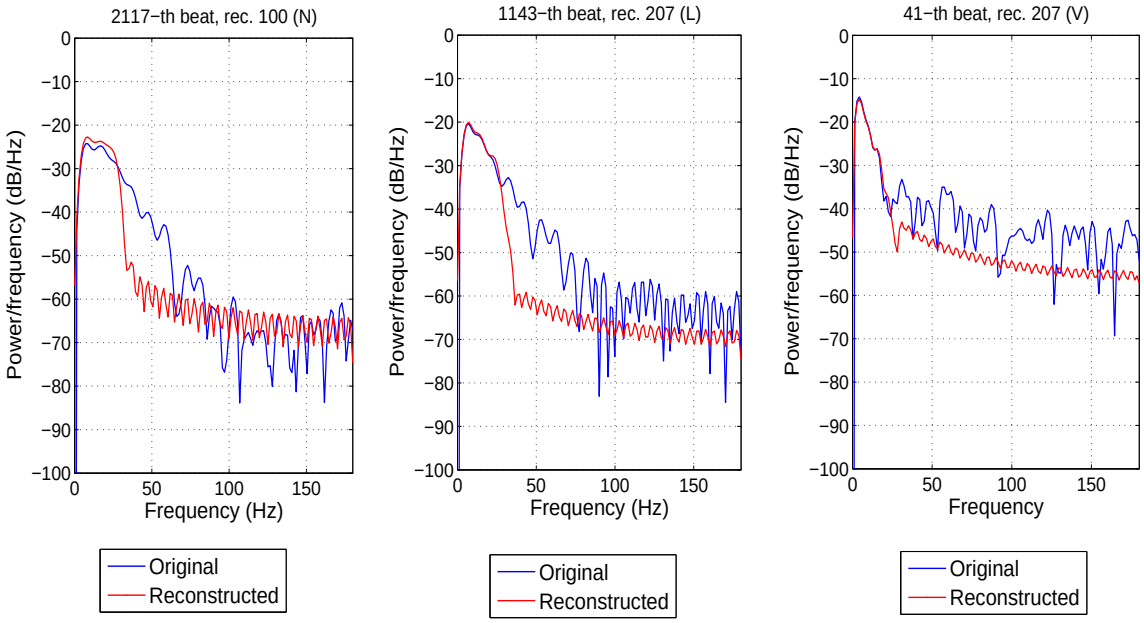


Figure 9.10: Spectrum of reconstructed signal employing first 11 elements of Hermite base

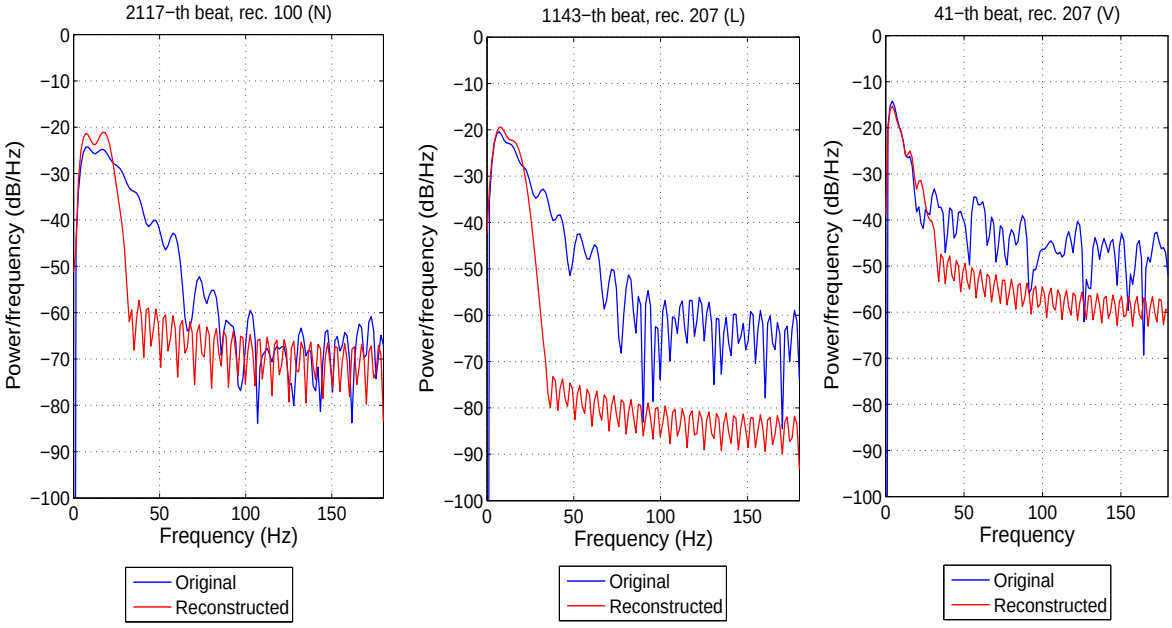


Figure 9.11: Spectrum of reconstructed signal employing first 9 elements of Hermite base

Experimentally it was found that an admissible value of the spectral difference

between original signal and its reconstruction (see section 5.4.2) is constrained by $\max|\text{diff}_n| \leq 5$.

In conclusion, Hermite model-based methodology for *QRS* characterization allows to reduce the search space of the optimal scale parameter σ_{opt} by minimizing the dissimilarity of spectra between reconstructed and original signal and, simultaneously, decrease computational load with the minimum number of elements to generate a proper reconstruction.

HRV features

As was described in Section 8.3, the supraventricular arrhythmias (S) and ventricular extrasystoles (V) have a common temporal pattern that can be well characterized using HRV information that discriminates over different type of arrhythmias such as Normal class (N).

The Figure 9.12 shows an example of separation between two classes for recording 232 from the database, using some HRV features (Table 8.4), such as, RR (x_1), pre-RR (x_2) and post-RR (x_3), described in Section 8.3.

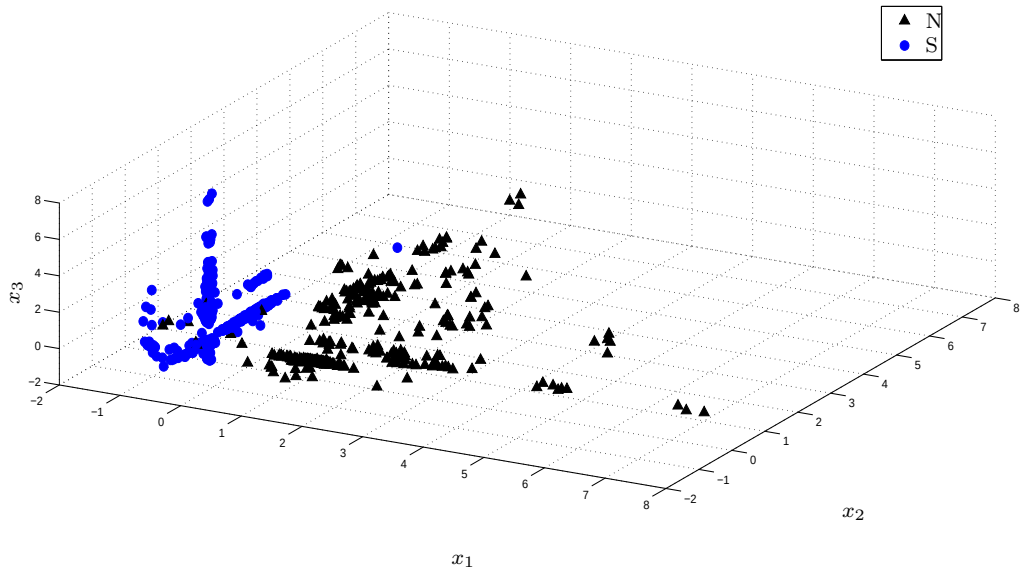


Figure 9.12: Separation between two classes (N and S) for recording 232 using some HRV features

However, all types of arrhythmias studied in this work, does not have the time variant pattern, such as the case of fusion (N), normal (F) or unknown (Q).

This aspect is depicted in Figures 9.13 and 9.14 by using features x_4 to x_6. The first one, presents separability between N and V classes, but, the second one, does not present separability between V and S classes, which have the same pattern.

In this way, a feature selection stage is necessary in order to select a proper set of features that discriminate the types of arrhythmias present into a Holter recording.

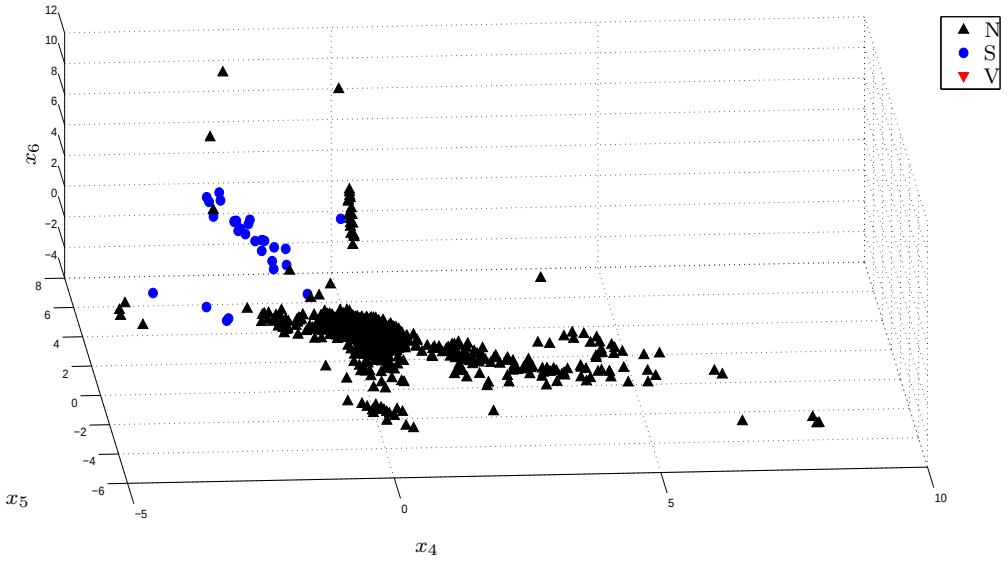


Figure 9.13: Separation between two classes for recording 213 using some HRV features

WT-based features

WT-based features have the capability to separate morphologies such as ventricular extrasystoles (V), fusion heartbeats (F) and unknown beats (Q). Nevertheless, arrhythmias characterized by time variant patterns, require HRV features.

Figure 9.15 shows a better separability between V and S type heartbeats from recording 213, than Figure 9.14 using the wt-based features x_{91}, x_{92} and x_{93} (See Table 8.4).

Figures 9.16 and 9.17 exhibit an example of good and bad separability among features, which are related to N, V and S arrhythmia types. While the first figure gives a good separation between N and V types, the second one does not, regarding N and S types, mainly because of both arrhythmias are characterized by HRV features.

Finally, Figure 9.18 shows a case where 3 types of arrhythmias can be separated using WT-based features. The classes correspond to N, V and Q types from recording

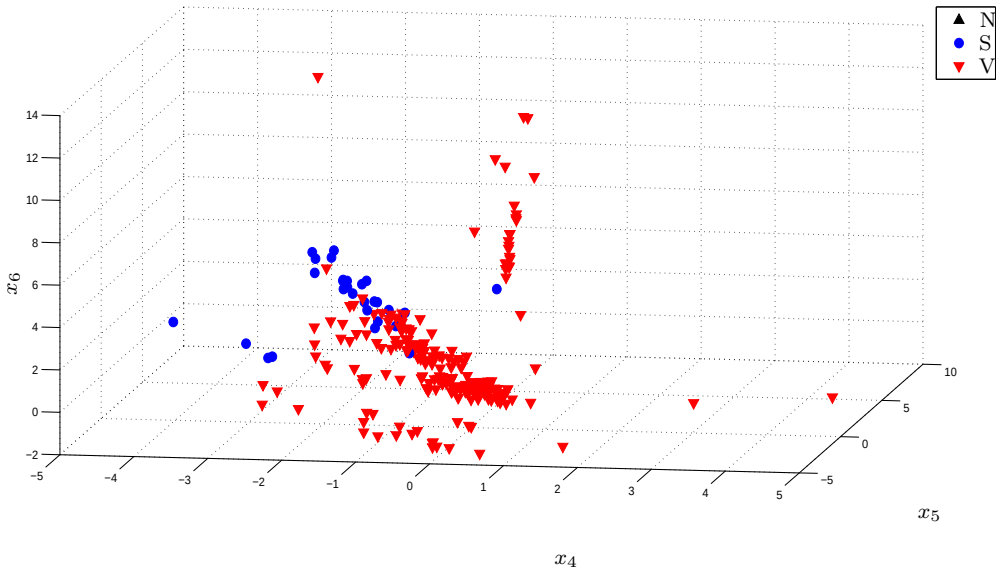


Figure 9.14: Separation between two classes for recording 213 using some HRV features

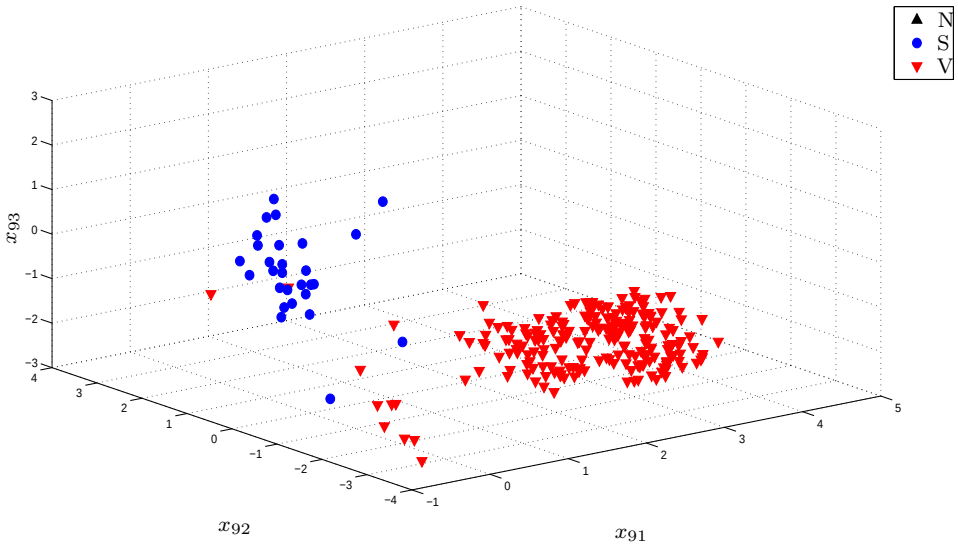


Figure 9.15: Separation between two classes for recording 213 using WT-based features

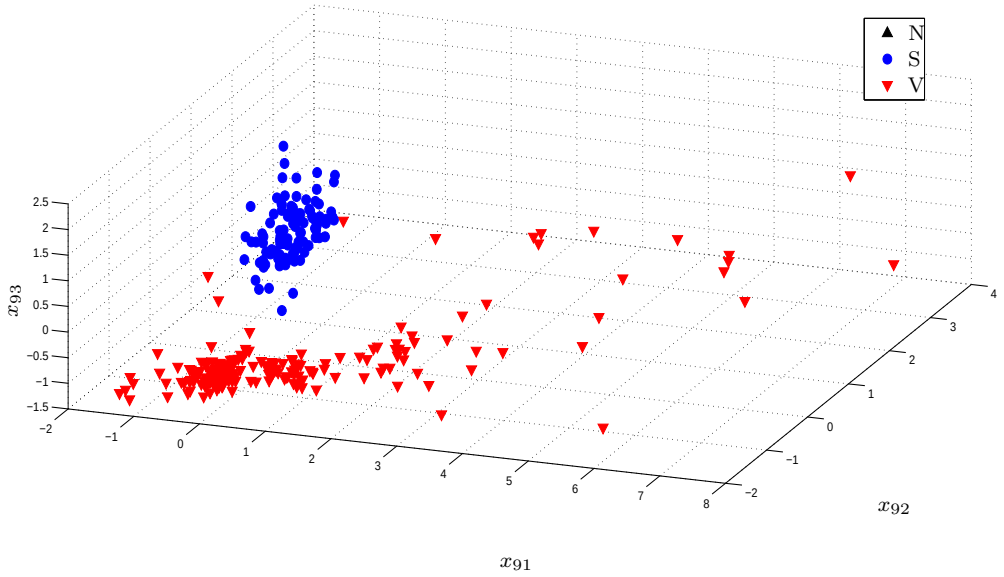


Figure 9.16: Separation between two classes for recording 207 using WT-based features

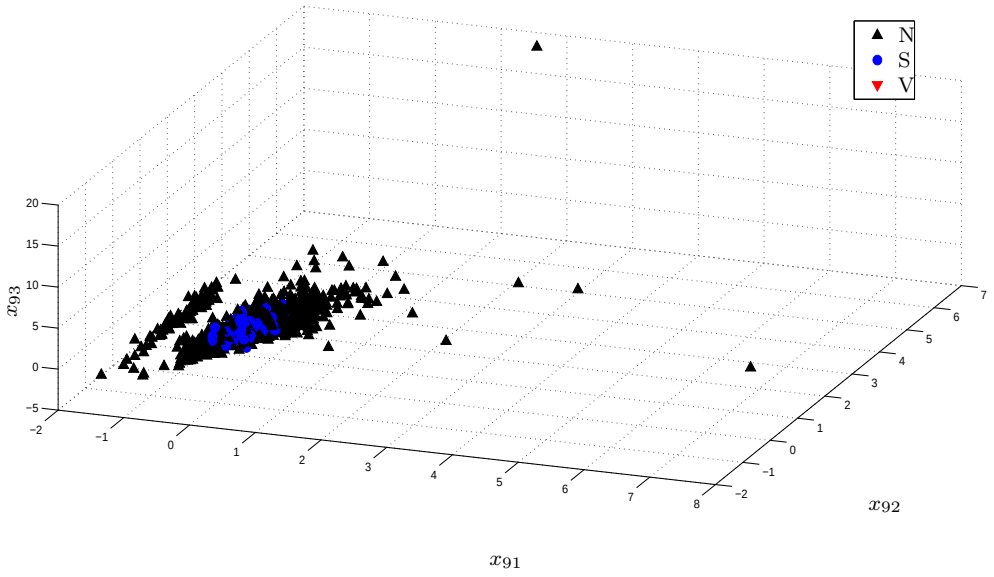


Figure 9.17: Separation between two classes for recording 207 using WT-based features

9.2 Analysis of Relevance Results

Figure 9.19 shows an example for relevance analysis stage using the proposed scheme, taking into account the last 5 minutes of record 217. It can be observed that occurs a short separation of first 3 principal components.

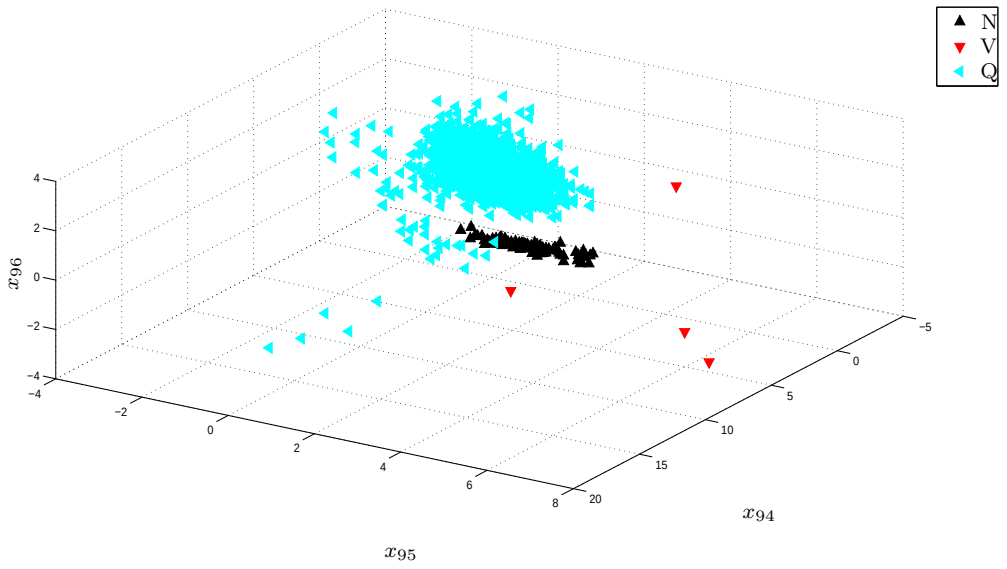


Figure 9.18: Separation among three classes for recording 102 using WT-based features

Remaining subfigures show the transformed data employing the studied methods where can be noted a better separability when using $\boldsymbol{w} = \sqrt{\widehat{\boldsymbol{\rho}}}$ and $\boldsymbol{w} = \sqrt{\boldsymbol{\alpha}}$. Particulary, in case of $\widehat{\boldsymbol{\rho}}$, the ignored eigenvectors (see (6.9)) for computing the relevance, generate a homogeneous weighting of the set of analyzed features, resulting in a lower selectivity, i.e., $\boldsymbol{w} = \sqrt{\boldsymbol{\rho}}$, where its separability is similar to $\boldsymbol{w} = \mathbf{1}$.

The variable weighting using the analyzed methods is shown in Figure 9.20, where the last five minutes of recording 217 are assessed. All methods excepting $\boldsymbol{w} = \sqrt{\boldsymbol{\rho}}$, give more relevance to some variables while leave without effect to others. Regarding $\boldsymbol{w} = \sqrt{\boldsymbol{\alpha}}$ and $\boldsymbol{w} = \sqrt{\widehat{\boldsymbol{\rho}}}$, there exist a similarity among relevant feature groups, e.g. some coefficients of WT-based features are relevant, mainly due to the arrhythmia types present in the recording as is depicted in Figure 9.19. Analysis with $\boldsymbol{w} = \sqrt{\boldsymbol{\rho}}$ gives equal relevance to almost all features, as was mentioned above.

Figure 9.21 shows the dynamic of calculated relevance of variables according to morphology type of each recording. Three segments of recording 207 are analyzed. The first segment corresponds to the first 5 minutes of recording, which contains beats type L, R and V. The second one corresponds to a time period between 20 and 25 minutes, that only have beats type L and V. The last one contains beats type A and E and corresponds to the last 5 minutes of recording.

It can be seen that in the first segment, the relevant variables correspond to some groups of the WT-based features (Table 8.4), while in the second one, besides, the WT-

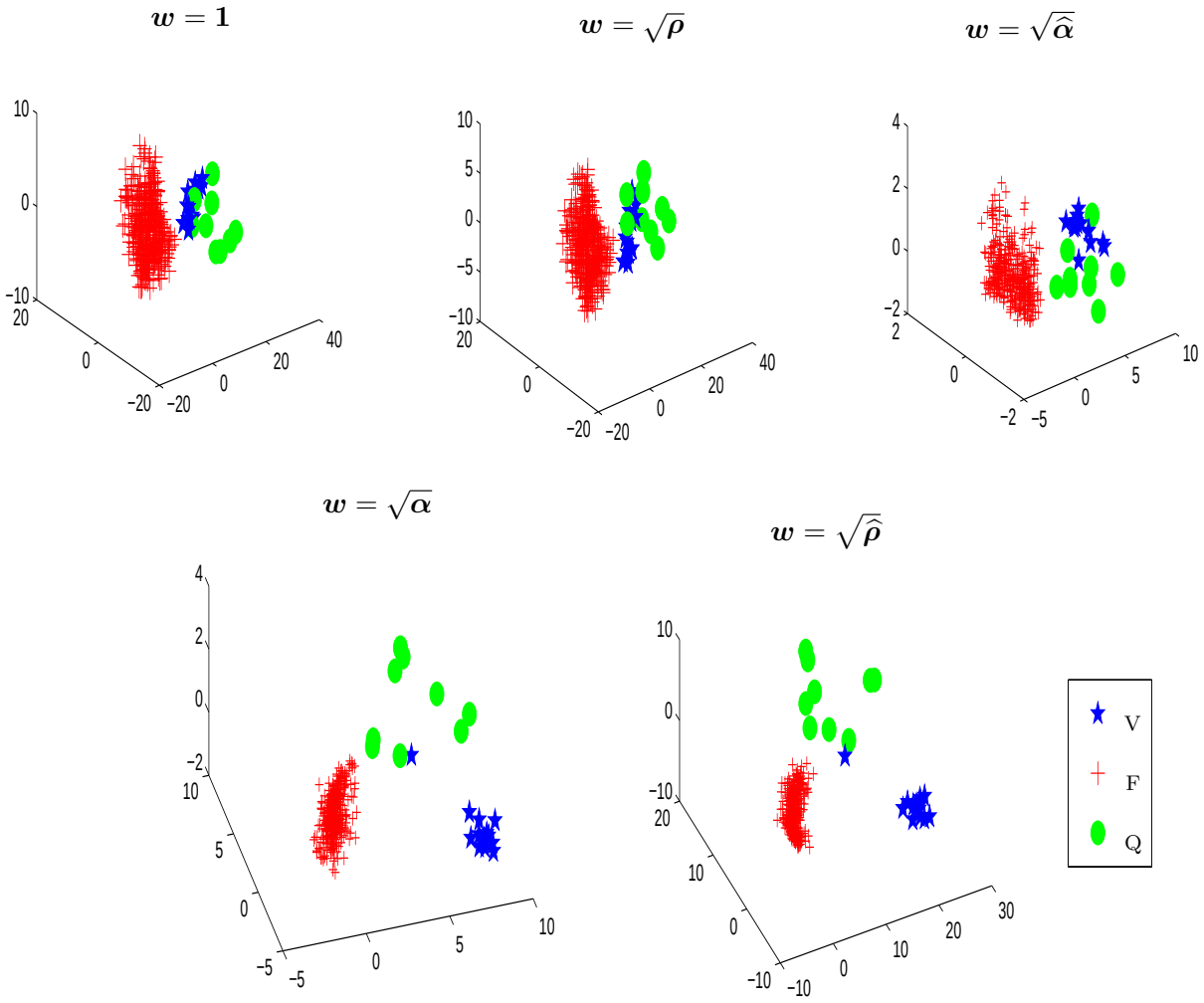


Figure 9.19: First 3 principal components after weighting the data matrix for recording 217 with different $\boldsymbol{w}$ (Algorithm 2)

based features, the Hermite coefficients have more weighting because these coefficients characterize appropriately the morphology of beats type L and V.

Finally, in the last analyzed segment the weighting for the first 3 variables (HRV features) is increased.

According to this, it can be concluded that segment analysis allows a local analysis of relevance and achieves a better performance after the final division, as will be shown in Section 9.3.

It should be highlighted that the variability features (HRV, Table 8.4) are essential to discriminate between normal heartbeats and supraventricular ectopic beats, whose

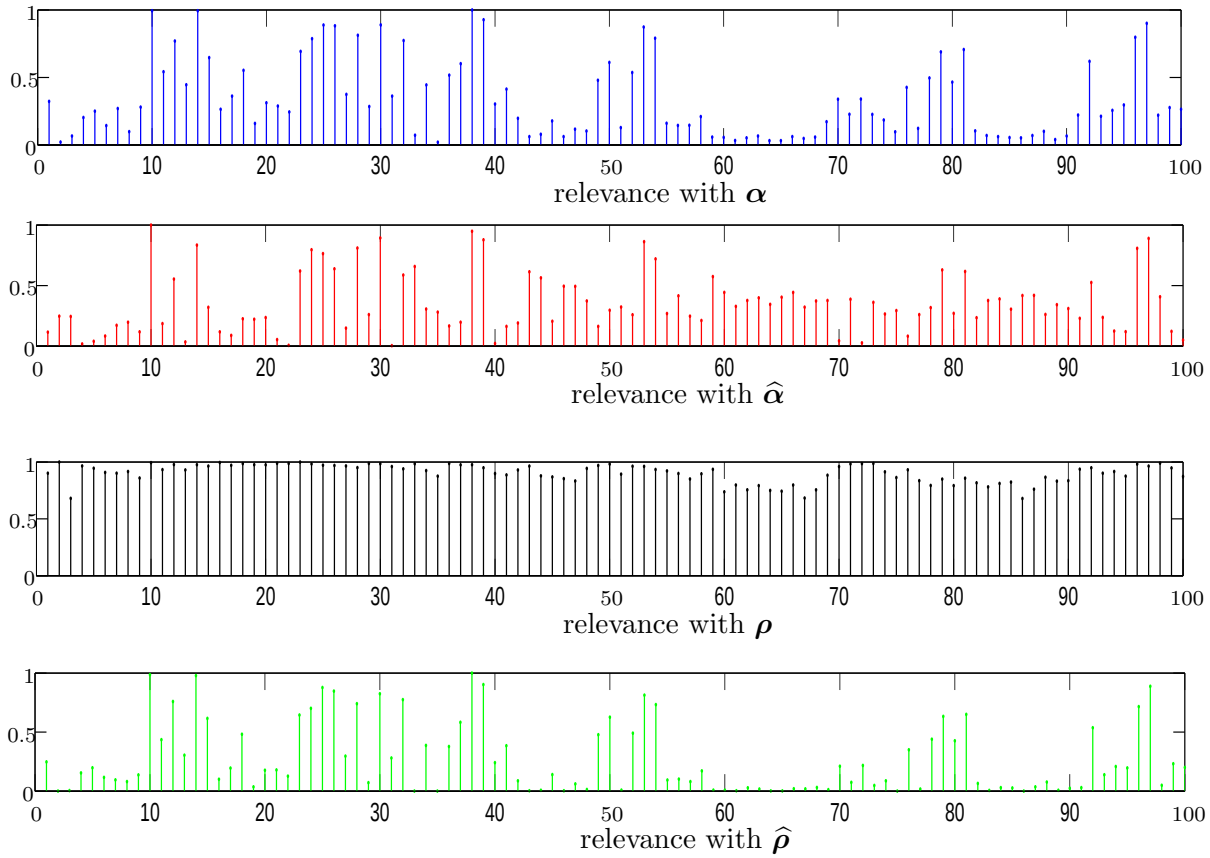


Figure 9.20: Relevance of the features (Table 8.4) for the last 5 minutes of recording 217 using all methods to estimate $\boldsymbol{w}$ (Algorithm 2)

morphology are similar. As to the relevant group of morphological features, the WT-based features have the ability of discriminating among heartbeats of type V, F and Q.

With the aim to quantify the capability of feature selection methods in cardiac arrhythmias analysis, a clustering algorithm with fixed parameters was used (Section 7.3.2), which is assessed by means of a sensitivity measure (Se), proposed in Section 7.6. This measure quantifies the proportion of heartbeats belonging to an Interest Class (IC) that are classified correctly. In this case, sensitivity measure is taken into consideration because if some classes are significantly minority, i.e., have a very small amount of heartbeats in a specific ECG recording, e.g., F or Q types, they might present low values of Se (close to 0), representing a weakness of the system to detect abnormal heartbeats, which can be mainly attributed to a lack in the performance of feature selection.

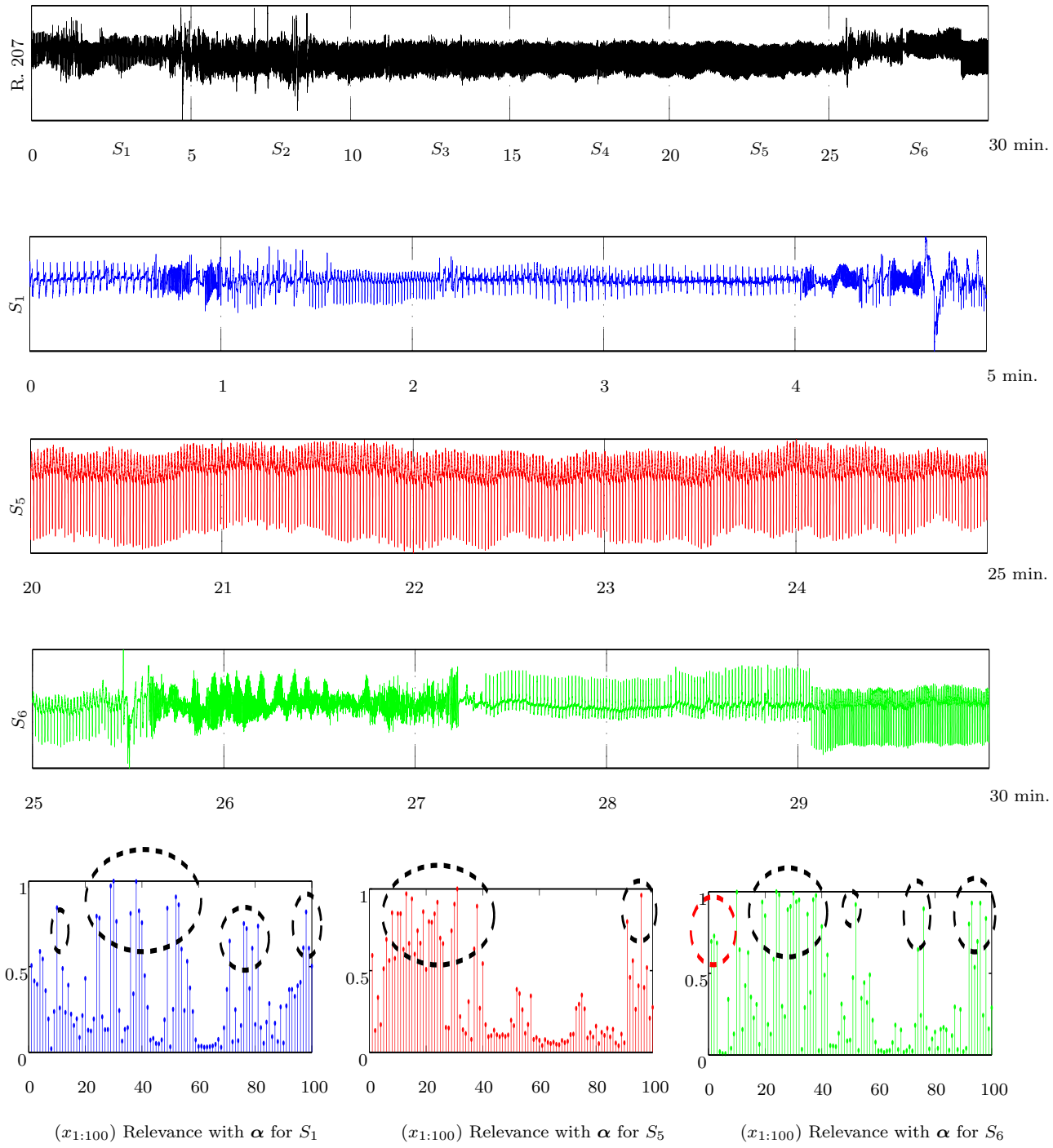


Figure 9.21: Results of the relevance of the features with Q-α method, for 3 segments of the recording 207

Detailed results are shown in Figures 9.22-9.26, where are depicted the measure Se for recordings that containing the heartbeats of interest, i.e, 5 groups of arrhythmias, as is described in Table 2.2. Recordings from the MIT/BIH database that do not appear in Figures 9.22-9.26 achieve a performance of 100 %.

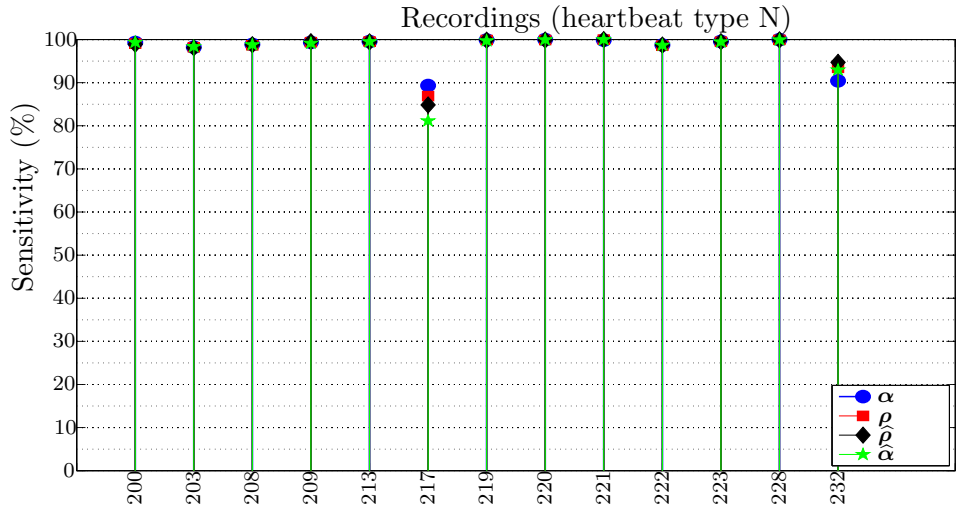


Figure 9.22: Sensitivity for Normal (N) heartbeats from the MIT/BIH database

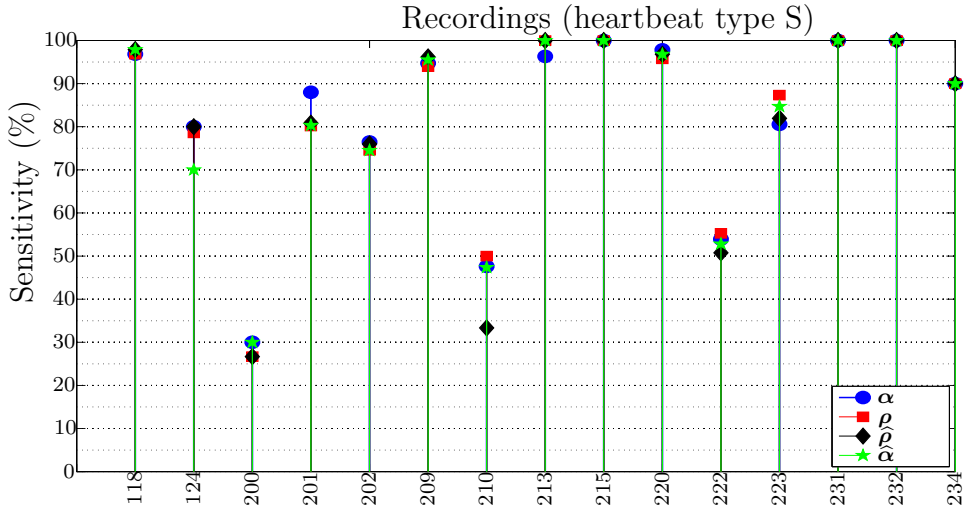


Figure 9.23: Sensitivity for Supraventricular (S) heartbeats from the MIT/BIH database

Figures 9.22-9.26 show a similar performance except for some recordings, e.g. 217, 201, 210, 230, among others; in which the Q-α method shows better performance.

There exist cases when the sensitivity is $Se = 0$, which corresponds to clusters where their heartbeats of the class of interest are mixed in other clusters. From a

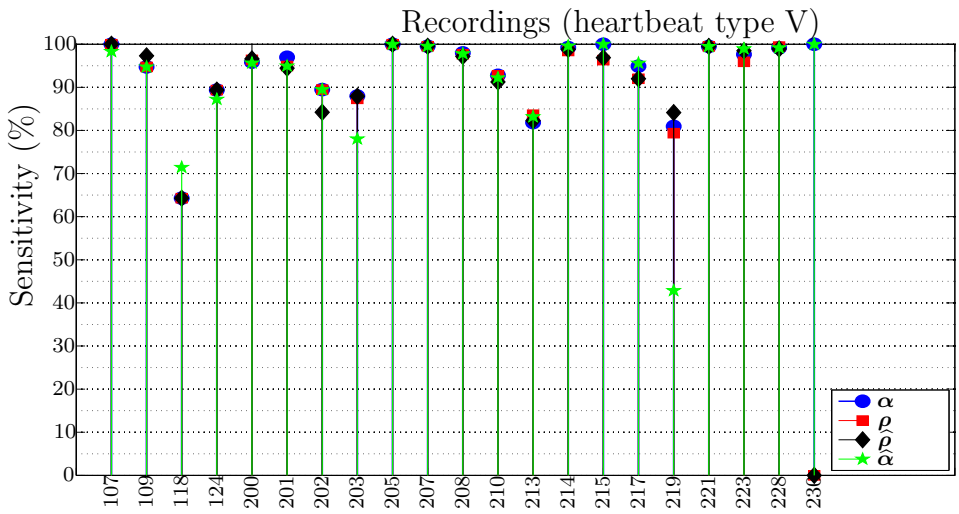


Figure 9.24: Sensitivity for Ventricular (V) heartbeats from the MIT/BIH database

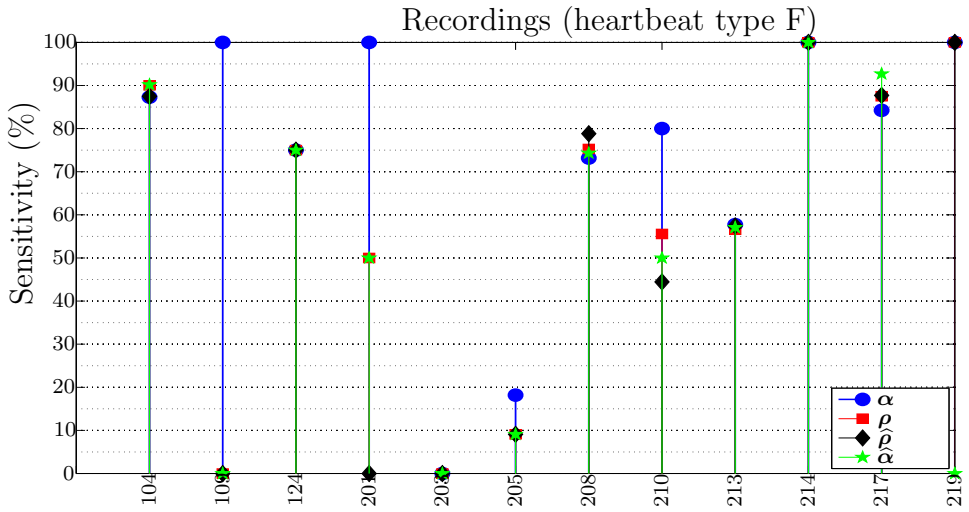


Figure 9.25: Sensitivity for Fusion (F) heartbeats from the MIT/BIH database

medical perspective, this represents a problem because high-risk heartbeats can be mixed in other clusters with normal classes. Thus, it is necessary that the sensibility measure should be greater than zero, i.e. $Se > 0$.

In conclusion, the weighting obtained from iterative $Q - \alpha$ algorithm (1 and 2), stands out mainly due to both the quadratic nature of the objective function to be maximized that employs M-inner product as distance measure, and the capability to select features in a unsupervised fashion that provides better separability between classes present in the recordings, as is discussed throughout this chapter. For these reasons, most of tests in the following section 9.3, were performed using the $Q - \alpha$

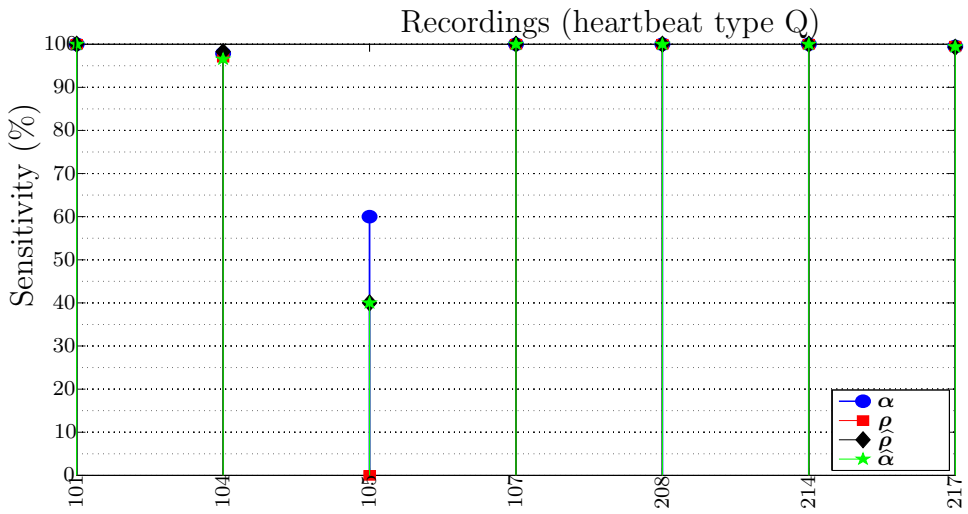


Figure 9.26: Sensitivity for Unknown (Q) heartbeats from the MIT/BIH database

algorithm.

9.3 Clustering Results

In this section, clustering results are presented. Clustering performance is assessed taking into account the automatic estimation of number of groups as is described in Section 7.4. It is also taken into consideration the effect of center initialization in comparison with random initialization, as discussed in 7.3. Furthermore, soft clustering performance is studied in order to analyze its behavior in connection with K-means based traditional grouping, using the algorithms described in 7.2. Finally some results for clustering performance are presented by varying the number of segments used to divide the recording. Clustering is quantified by performance measures (described in Section 7.6) and processing time.

9.3.1 Estimation of the number of groups

Estimation without relevance analysis

In Figure 9.27 are shown the estimated values of the number of groups (k) for all recordings from data base MIT/BIH, where heartbeats from each whole recording are analyzed without divisions, i.e. $N_s = 1$. Figure 9.28 shows the time spent per each considered methods for automatic estimation of the number of groups by using whole

recording.

In Figure 9.29 can be seen the estimated values for the number of groups over all recordings, by dividing each recording into 6 parts ($N_s = 6$) and estimating k per each one. Finally, Figure 9.30 shows its the corresponding processing times with $N_s = 6$.

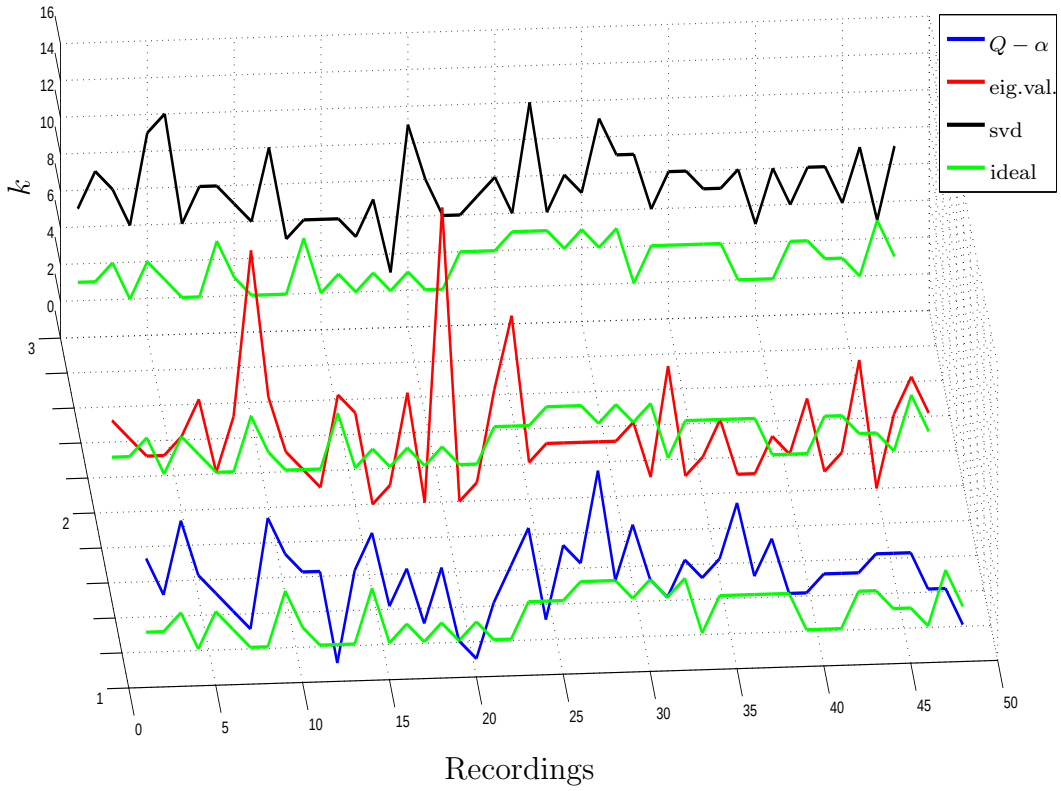


Figure 9.27: Estimation of number of groups without relevance analysis for all recordings from MIT/BIH database. $N_s = 1$.

Estimation using relevance analysis

As is depicted in Figure 9.28, time employed by the methods for estimation of number of groups when $N_s = 1$ is very high, including reaching up to about 2000 s. Meanwhile, in case of segment analysis processing time is decreased, as can be illustrated in Figure 9.30.

Because of the above discussed, the case of estimation using relevance analysis with $N_s = 6$, is analyzed, as can be seen in Figure 9.31. Processing time to estimate the number of groups for each method is shown in Figures 9.32 and 9.33.

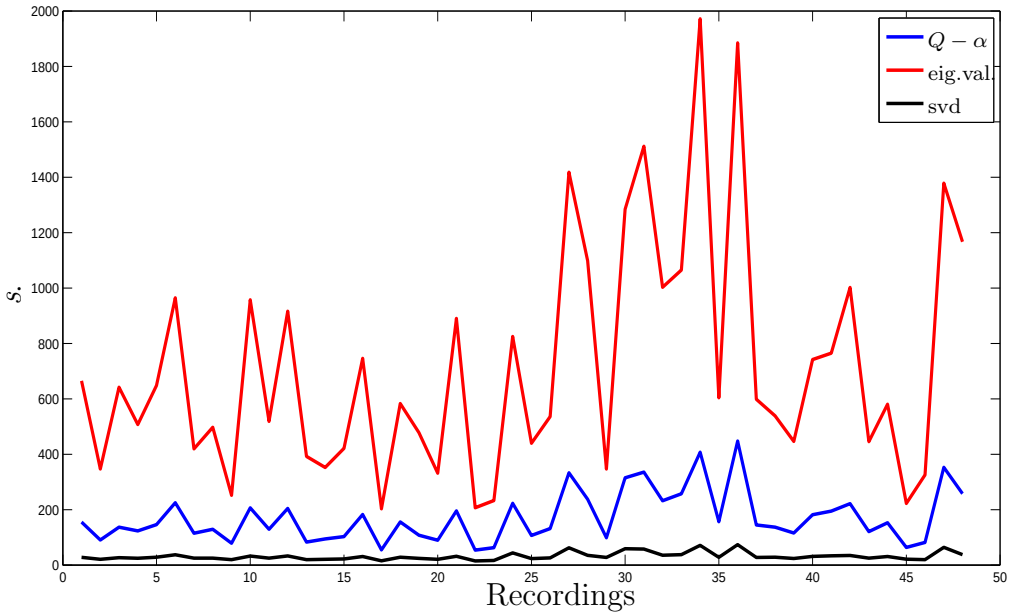


Figure 9.28: Time used in estimating the number of groups without relevance analysis using $N_s = 1$ segment, for all recordings from MIT/BIH database

As depicted in Figure 9.27, there exist a correspondence relation between the ideal number of groups and the estimated by the proposed methods.

In case of SVD-based estimation method, the parameter α_{svd} (See eq. 7.19) is experimentally tuned between the interval $0 \leq \alpha_{svd} \leq 1$, obtaining an estimated number between 3 and 10 with $\alpha_{svd} = 0.6$. Nonetheless, in most cases, the estimated is greater than the ideal value. From point of view of computational cost, this method have better performance because requires less amount of operations in comparison with the remaining methods. In this way, to analyze high-dimensional data set as those accomplished for ECG recordings, SVD-based approach provides a processing time less than others in case of $N_s = 1$, as shown in figure 9.28.

Eigenvalues-based approach, although improved by using a soft scaled affinity matrix, is one with highest dispersion as can be observed in Figure 9.27. Also, computation of affinity matrix implies a high computational cost, including reaching up towards 2000 $s.$ in the estimation process.

Method based on $Q - \alpha$ showed that can give substantial information to estimate the ideal number of groups as can be illustrated in Figure 9.27. Also, it spends a reasonable time regarding other methods. Performance of this method is related to the

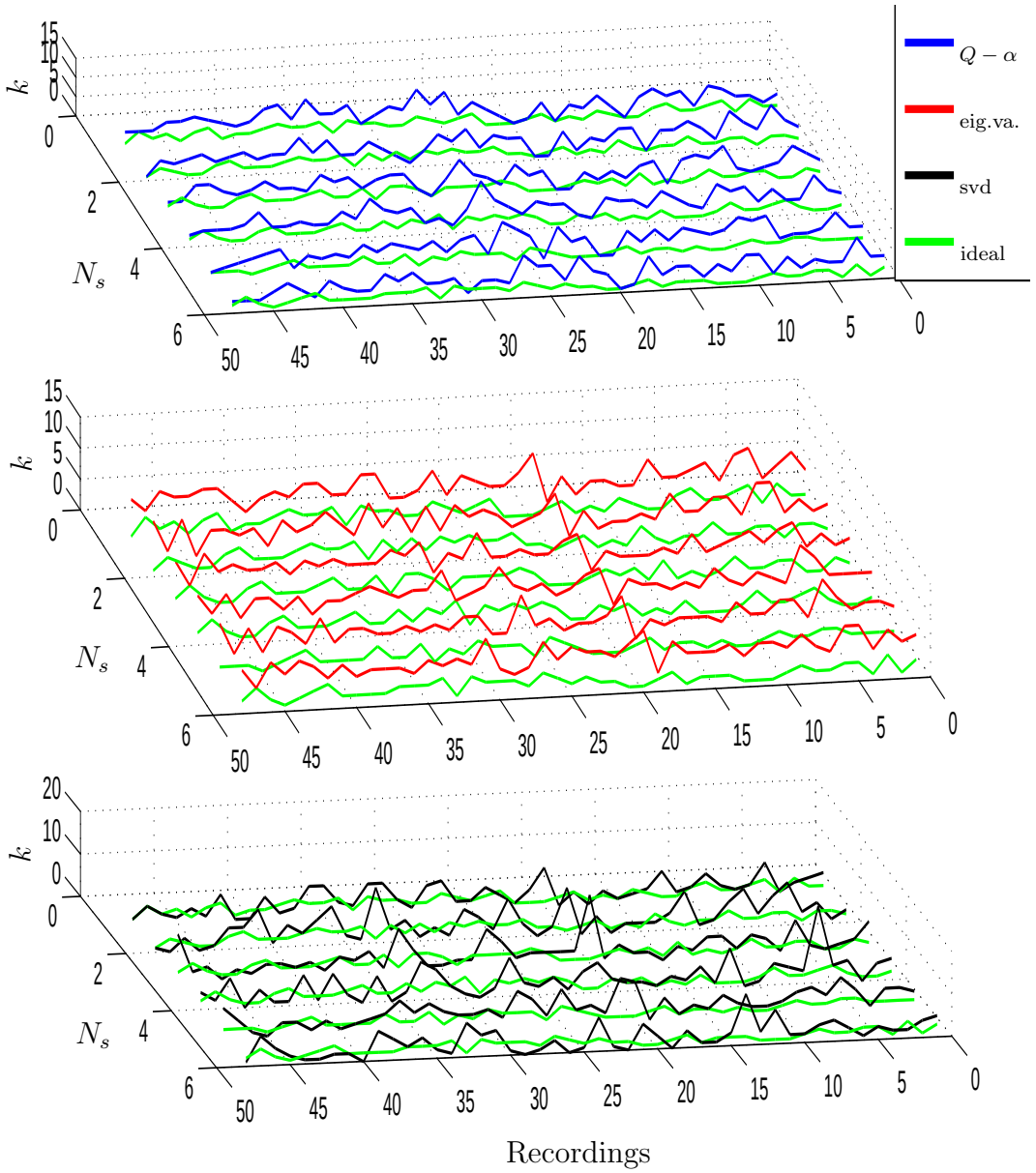


Figure 9.29: Estimation of number of groups without relevance analysis for all recordings from MIT/BIH database, with $N_s = 6$ segments.

optimization problem given by equation (6.16), which is solved with optimal values for vector $\boldsymbol{\alpha}$ and matrix $\boldsymbol{Q}$, which are tuned in an iterative fashion described in Algorithm 1.

However, it is important to highlight that SVD-based estimation is the method that best works when it is analyzed the whole data set without divisions in a high dimensional space, where the parameter α_{svd}, due to its dependence on analyzed data

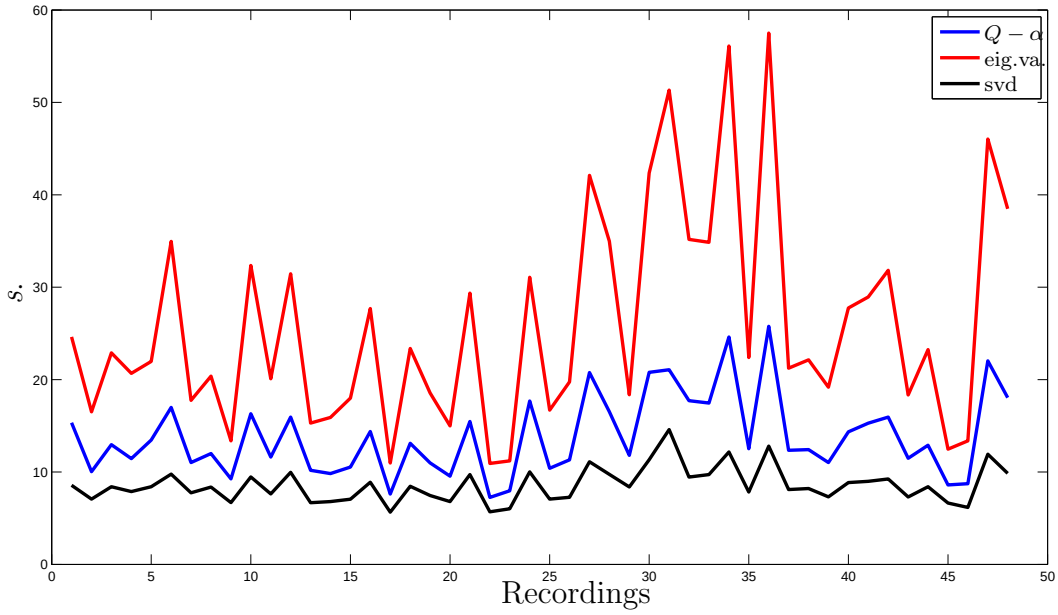


Figure 9.30: Time used in estimating the number of groups without relevance analysis using $N_s = 6$ segments, for all recordings of the MIT/BIH database

type, is tuned per each experiment.

When recording is divided into various segments, all methods present similar behavior. In the case of SVD approach, parameter α_{svd} is set to a fixed value in order to study the sensitivity of this method. Then, in Figure 9.29 can be seen that the estimated number of groups by using SVD was 1 for several recordings. In connection with processing time, the Figure 9.30 shows a considerable time decrease, reducing the estimation process from 2000 to around 60 s, as to eigenvalues method (recording 36). However, from all methods, SVD-based approach presents least computation cost, although also presents dependence with value of parameter α_{svd}.

Because proposed methodology includes a relevance analysis stage and considering that $Q-\alpha$ method has shown the best performance, it can be used to compute the number of groups. Then, $Q-\alpha$ algorithm, in addition to provide useful information to determine the feature relevance, obtains an estimate to the number of clusters without applying further procedures. Thereby, it becomes in the best method in terms of both correspondence with number of groups estimation (see Figure 9.31) as well as computational cost that is slightly less than SVD approach, as can be seen in Figure 9.33. Again, in case of SVD method, because of sensitivity to data matrix, it can be observed that the estimated number of groups for some recordings is 0. By other hand,

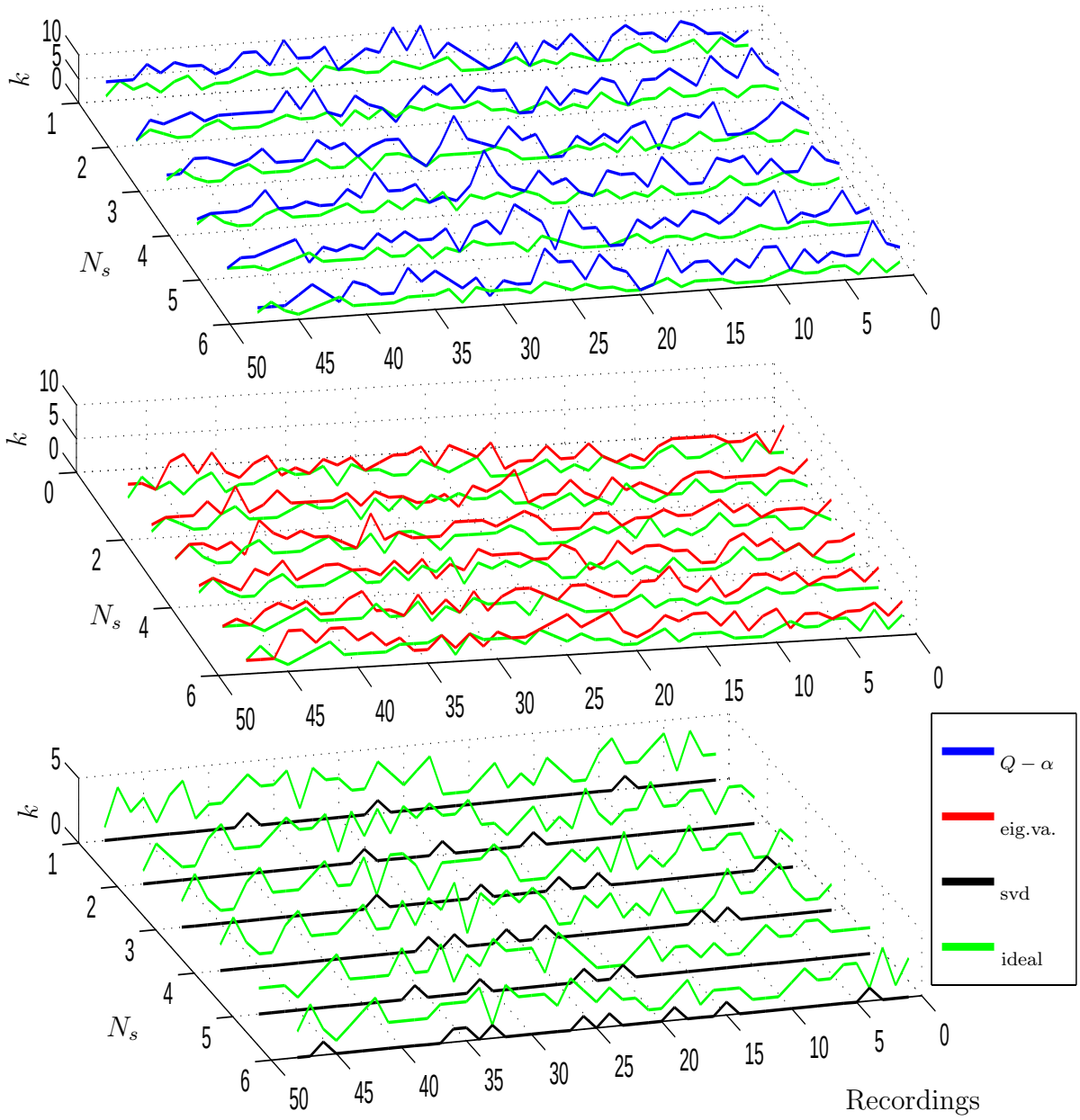


Figure 9.31: Estimation of number of groups with relevance analysis for all recordings the MIT/BIH database, with $N_s = 6$.

$Q - \alpha$ is absolutely unsupervised and therefore does not require of parameter tuning.

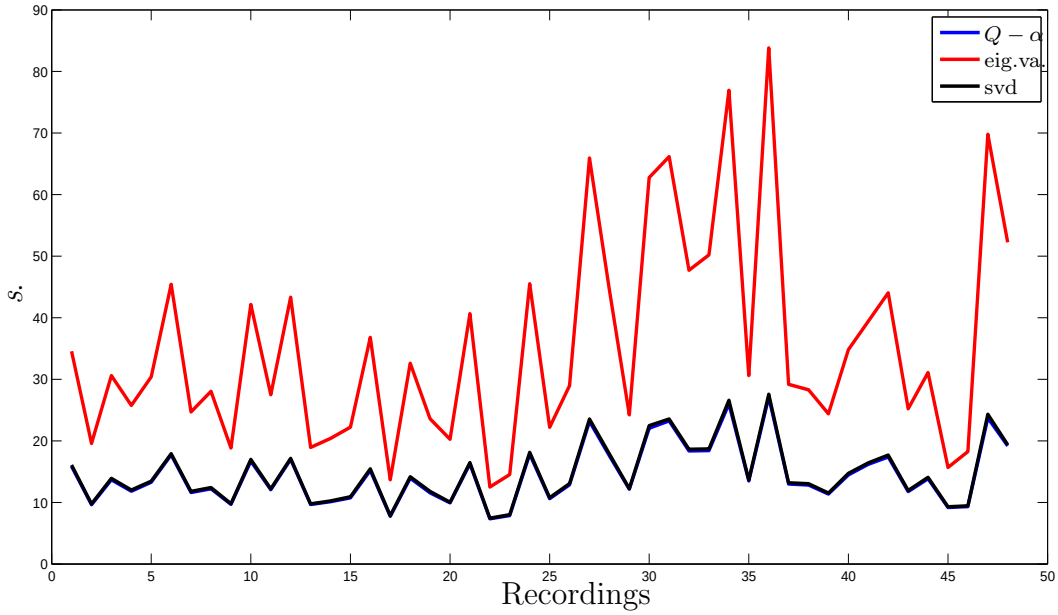


Figure 9.32: Time used in estimating the number of groups with relevance analysis using $N_s = 6$ segments, for all recordings from MIT/BIH database

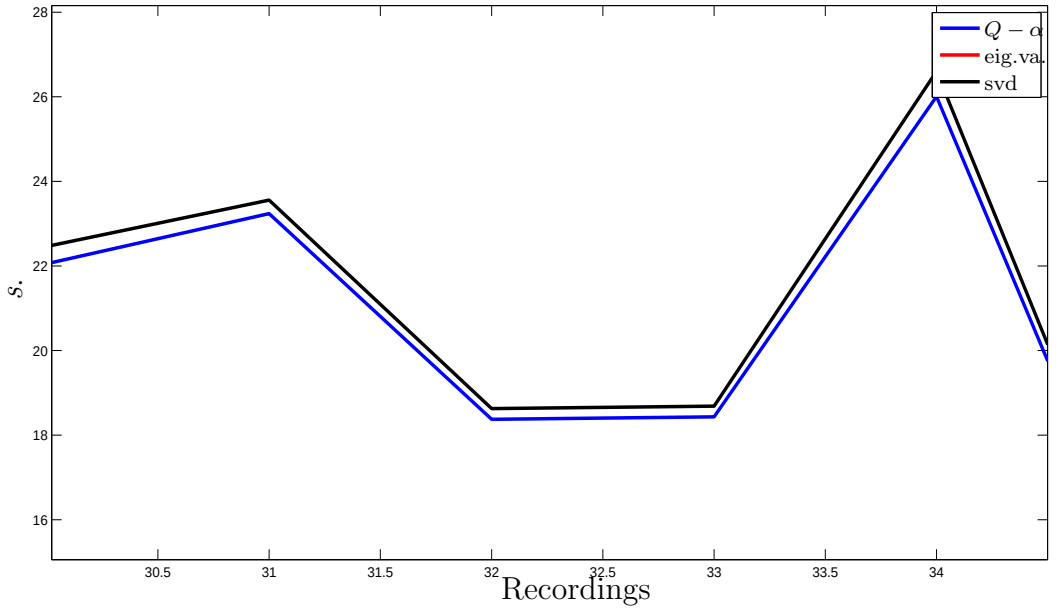


Figure 9.33: Time difference between $Q - \alpha$ method and *svd*, for some recordings of the MIT/BIH database. The analysis is carried out using $N_s = 6$.

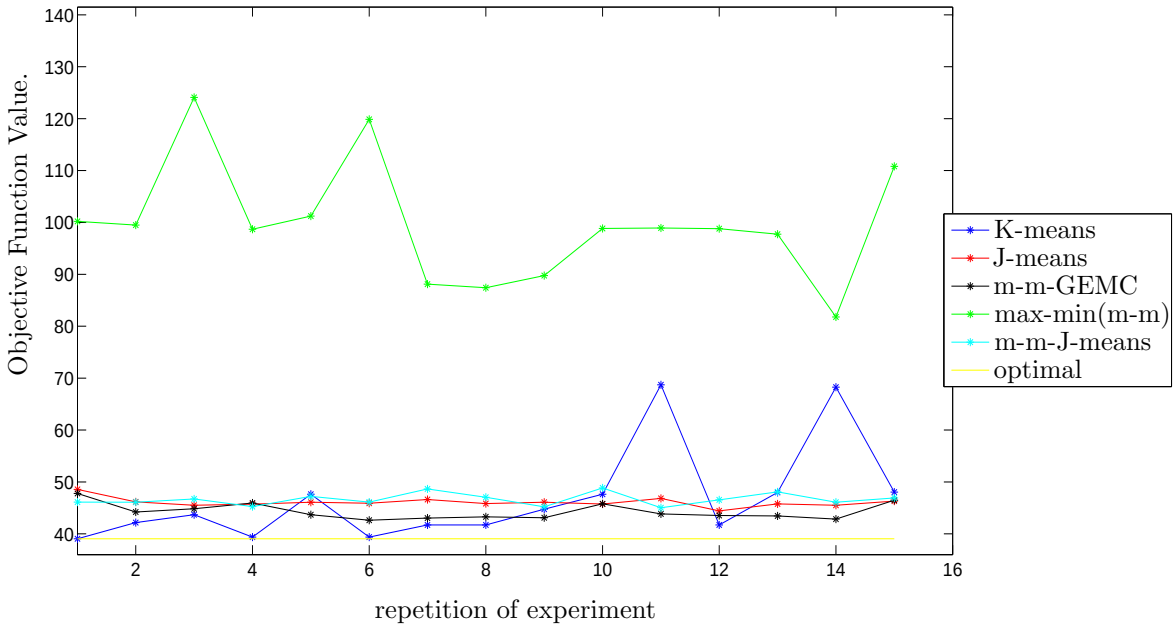


Figure 9.34: MSSC results by applying the initialization criteria with different methods, on the Fisher Iris data set with $k = 6$.

9.3.2 Initialization criteria

In this work, it is proposed a methodology for center initialization that performs better than the well-know random standard initialization. The problem with random initialization is that it can converge to a local minimum which corresponding partition could be unsuitable for the desired results, for instance, it could generate a clustering of different classes in a single cluster and hiding important elements to achieve a proper resultant partition. Thus, the max-min algorithm was explored that has shown good results in related works [64]. The disadvantage of this method is the sensitivity to outliers. Therefore, in this work is proposed, as an alternative to initialization stage, the clustering algorithm called J-means, whose initial parameters correspond to output of max-min algorithm. The advantage of J-means algorithm is that despite a bad initialization may converge to an objective function value near the global optimum value. The processing time that J-means algorithm spends is linear ($O(n)$), and it is therefore easy to implement into the proposed methodology with low computational cost.

To test the effectiveness of proposed procedure, Fisher Iris data set is employed that is a standard database commonly used for assessing pattern recognition algorithms. About this data base, it is known the optimal value of objective function with respect

to the number of clusters according to the minimum sum of squares (7.1), eg with $k = 2$, the optimal value is 152.3470 and with $k = 7$, the optimal value is 34.2982, as is explained in [151]. By taking into account these values, the following tests are performed.

Figure 9.34 shows a comparative of different methods to centroid initialization. Fifteen experiment repetitions were developed, where, for each experiment the objective function was evaluated.

Because the optimal value of objective function is 39.0399 for $k = 6$ [151], it can be noted that max-min algorithm presents the lowest performance, reaching values between 80 and 125. K-means algorithm, when a bad initialization occurs, increases the objective function value until 70. Therefore, although sometimes good performance can be achieved, K-means algorithm is very sensitive to initialization. Meanwhile, J-means and max-min-J-means present a similar performance that is not affected by initialization. In addition, max-min-GEMC, which is an algorithm used as a clustering stage in Section 9.3.3), presents even better performance that the other considered approaches due to its soft nature of the element assignment function (see Section 7.10).

Another test is presented which is oriented to assess several initialization criteria applied over a specific heartbeat (recording 207). This recording was chosen because it has 4 of classes of interest (A, V, L and R), and therefore is a representative recording from considered data base. In Table 9.5 is shown the performance of GEMC algorithm before applying max-min, J-means and J-GEMC criteria. Table values correspond to average and standard deviation of sensitivity (Se) and specificity (Sp) indices, computed to evaluate whole data set in 10 clustering procedure iterations. First column shows the name of used methods. Second column shows the corresponding results obtained without using convergence control. Finally, in third column are shown results by calculating the change of objective function. As can be seen, it was employed a DBC algorithm, that is because soft algorithms present better performance and are less sensitive to initialization, as it will be discussed in 9.3.3.

In general, it can be concluded that initialization algorithms improve the clustering performance, providing a proper partition initial. J-means based algorithms present better results in comparison with those obtained from max-min, this is because the continuos and systematic assessment of the change of objective function carried out by J-means procedures. In addition, such objective function can be adjusted to a specific grouping method (DBC or MSSC) in proper form. By the other hand, max-min method uses a distance preestablished criterion and then does not take into consid-

Table 9.5: Results for GEMC with different initialization criteria

Method	Iterations			Objective function			
	Iter.	Se $\mu - \sigma$	Sp $\mu - \sigma$	δ	Iter. Resultantes round(μ) $- \sigma$	Se $\mu - \sigma$	Sp $\mu - \sigma$
max - min	10^1	0.52 - 0.016	0.87 - 0.02	10^{-1}	10 - 0.707	0.56 - 0.022	0.88 - 0.027
	10^2	0.96 - 0.027	0.99 - 0.032	10^{-2}	15 - 1.01	0.63 - 0.015	0.88 - 0.017
	10^3	0.94 - 0.03	0.99 - 0.025	10^{-3}	25 - 1.13	0.95 - 0.032	0.99 - 0.015
J-means	10^1	0.63 - 0.01	0.91 - 0.008	10^{-1}	5 - 0.48	0.72 - 0.014	0.85 - 0.01
	10^2	0.95 - 0.012	0.99 - 0.009	10^{-2}	11 - 0.53	0.95 - 0.018	0.99 - 0.01
	10^3	0.95 - 0.013	0.99 - 0.008	10^{-3}	14 - 0.39	0.95 - 0.016	0.99 - 0.01
J-GEM	10^1	0.64 - 0.009	0.91 - 0.004	10^{-1}	5 - 0.32	0.72 - 0.006	0.87 - 0.002
	10^2	0.95 - 0.008	0.99 - 0.004	10^{-2}	11 - 0.31	0.95 - 0.004	0.99 - 0.002
	10^3	0.95 - 0.007	0.99 - 0.004	10^{-3}	12 - 0.31	0.95 - 0.004	0.99 - 0.002

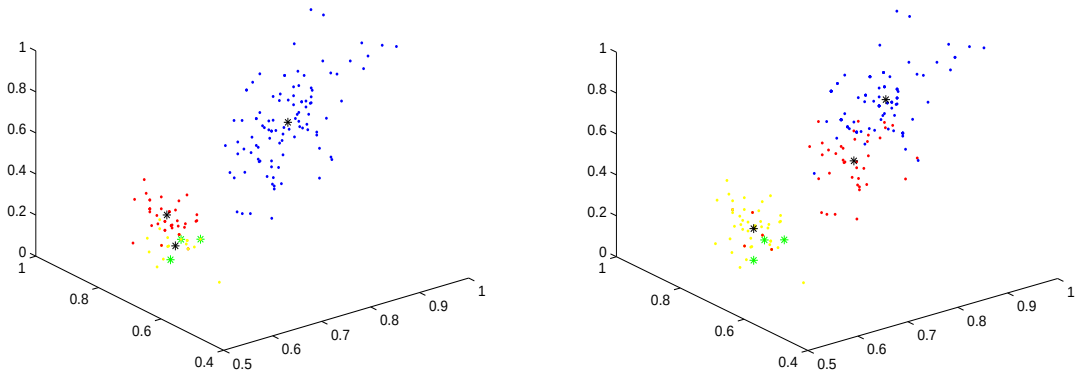
eration the nature of objective function corresponding to clustering procedure. Also, J-means algorithm reduces the computational cost by computing the objective function locally (see section 7.3.2).

As was mentioned above, soft clustering methods, as DBC, are less sensitive to initialization because the membership values have major probability to change than those hard methods. Such change can signify an orientation towards a better partition. In Figure 9.35 can be observed the low-sensitivity to center initialization of GEMC method. Green points represent initial centers and black points are the resultant centers. Feature space correspond to 3-class 4-dimensional artificial data set with 150 samples and balanced classes. Clustering performances achieved with this test using hard and soft methods are, respectively, $CP = 95.4\%$ and $CP = 40.3$. These values were calculated with $N_{iter} = 20$ and without convergence control (i.e., without computing the change of objective function).

9.3.3 Grouping algorithm

This test is focused to grouping of ventricular arrhythmias (R, L and V) using as data set representation and morphological features obtained from Hermite analysis, as can be seen in 5.4.2. Considered feature set is:

- QRS energy .
- Optimal scale parameter σ_{opt}.
- Hermite coefficient 6 (C_σ^n with $n = 6$).



(a) Hard clustering (b) DBC clustering

Figure 9.35: Sensitive to center initialization for hard and soft clustering

- Difference bettween current QRS complex and a QRS template and a QRS template (see equation 5.45).

Clustering stage is carried out by means hard and soft clustering algorithms. Performance indices correspond to those mentioned in Section 7.6.

Tables 9.6 and 9.7 show the clustering results. First 3 columns correspond to algorithm performance with initial partition chosen at random and the remaining 3 columns hold the performance clustering using max-min criterion for initialization. Procedure iterated 10 times in both cases and the mean (μ) and standard deviation (σ) values of performance indices were computed. Tests without using initialization criterion are carried out without convergence control and N_{iter} is set to be 20. Remaining results max-min were obtained with $N_{iter} = 100$ by assessing the algorithm convergence through the objective function value.

In Figure 9.36 is shown an example of feature space corresponding to recording 207. Beats type N from another recording (215) are added because recording 207 has not normal beats.

QRS morphological and representation features provide a good separability of ventricular arrhythmias considered in this study due to the physiological nature of this kind of signals. Also, these arrhythmias are directly related with QRS wave. In this case, energy, Hermite based features (C_n^{σ}, σ_{opt}) and spectral differences with QRS templates shown to be proper features for classification task.

Table 9.6: Results for hard clustering

Rec.	Random initialization ($\mu - \sigma$)			max- min ($\mu - \sigma$)		
	$Sp(\%)$	$Se(\%)$	CP (%)	$Sp(\%)$	$Se(\%)$	CP (%)
118	99.26− 0	38.46− 16.97	98.75− 0.12	99.26− 0	51.61− 0	98.85− 0
124	99.48− 0.14	72.34− 6.73	95.77− 0.002	99.54− 0	74.75− 0.12	95.94− 0.002
207	98.4− 0.1	99.1− 0.01	99.5− 0.002	98.9− 0.01	99− 0.002	99.5− 0.001
214	69.5− 0.0474	77.94− 0.09	75.64− 0.009	70.12− 0.021	79.35− 0.09	78.2− 0.008
215	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0
217	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0
219	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0
221	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0
223	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0
228	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0
230	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0
233	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0
234	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0

Table 9.7: Results for soft clustering

Rec.	Random initialization ($\mu - \sigma$)			max - min ($\mu - \sigma$)		
	$Sp(\%)$	$Se(\%)$	CP (%)	$Sp(\%)$	$Se(\%)$	CP (%)
118	99.26− 0	63.19− 3.7	99.6− 0.02	99.26− 0	63, 2− 0.1	99.72− 0.001
124	99.53− 0.1	74.86− 1.2	95.94− 0.0009	99.54− 0	74.87− 0.08	95.94− 0.001
207	98.9− 0.1	99− 1.2	99.5− 0.0009	98.9− 0	99− 0.08	99.5− 0.001
214	69.9− 0.009	80.94− 0.05	78.14− 0.009	70.2− 0.008	80.94− 0.092	78.2− 0.002
215	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0
217	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0
219	100− 0	100− 0	100 0	100 0	100 0	100 0
221	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0
223	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0
228	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0
230	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0
233	100− 0−	100− 0	100− 0	100− 0	100− 0	100− 0
234	100− 0	100− 0	100− 0	100− 0	100− 0	100− 0

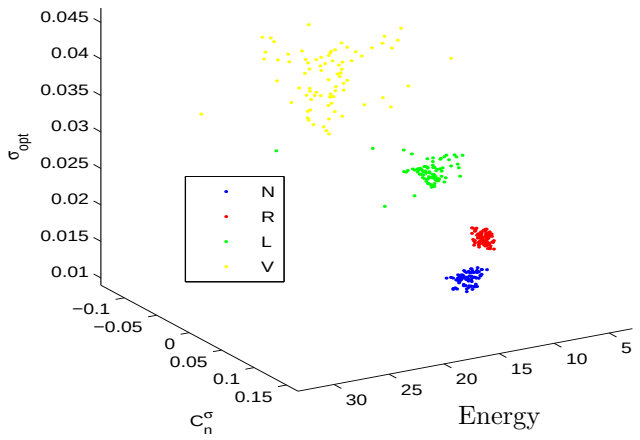


Figure 9.36: Features for recordings 207 (R, L, V) and 215 (N)

In general, two considered clustering methods generate similar results when is applied an initialization criterion, as can be seen in Tables 9.6 and 9.7. This is because of the leading advantage of soft clustering is the low sensitive to initialization. As can be easily noted, hard clustering performance without applying initialization criterion is decreased, while soft clustering keep practically the same performance. Thereby, it can be concluded that a good-tuned soft clustering could omit the initialization stage. Nonetheless, it should be said that a proper initialization criterion makes that clustering has a major probability to achieve a good convergence value.

Performance results are similar for almost all recordings (last 9 rows, recordings from 215 to 234), but recordings 118, 124 and 214 present low values of Sp because of considered features do not generate a proper separability in all cases. Also, in recordings 118 and 124, value of Se is high in contrast to low value of Sp. This is because of unbalanced number of observation per class, therefore only one beat wrongly classified could considerably affect the value of Sp.

9.3.4 Segment analysis results

The results of clustering are accomplished by framing each recording into 6 divisions and the resulting clusters are merged as described in Section 8.5.2. The number of segments is achieved experimentally, improving the trade–off among the number of segments, computational cost and quality of partition. Thus, the segment analysis enhances the performance if comparing to the whole data clustering. In fact, it reduces the