probability that a minority class heartbeat might be clustered wrongly. Furthermore, in most cases, the sum of processing times over all segments turns to be considerably shorter than the time of analysis of whole recording data for one iteration. Therefore, the introduced framing approach significantly reduces the computational cost.

Selection of proper number of clustering segments is constrained by following restrictions: twice of number of features must exceed the amount of heartbeats per segment, and the minimum of computational cost should be reached.

Defining the unit time (TU) as the time necessary to process the whole record in one iteration ($N_s = 1$) using only the initial vector of features, resultant performance can be expressed as multiples or fractions of TU. Fig. 9.37 shows some error bars where it is noted that computational cost improve as N_s increase. This test was carried out using all the relevance analysis methods and the initial data set, however, in Figure 9.37 are shown the most relevant results making a comparison between Q-α and reference methods, i.e, by analyzing all variables, the relevant ones and PCA. When $N_s = 1$, WPCA and PCA methods spend more time to carry out the process regarding the reference methods, but when N_s increases, the processing time decreases. It can also see that error bar calculated as $1 - Se/100$, decreases as N_s increases.

In Figure 9.37 can be seen that after some number of segments, the computational cost increases again. This is because the initialization and feature selection routines that are performed for each segment, increase processing time.

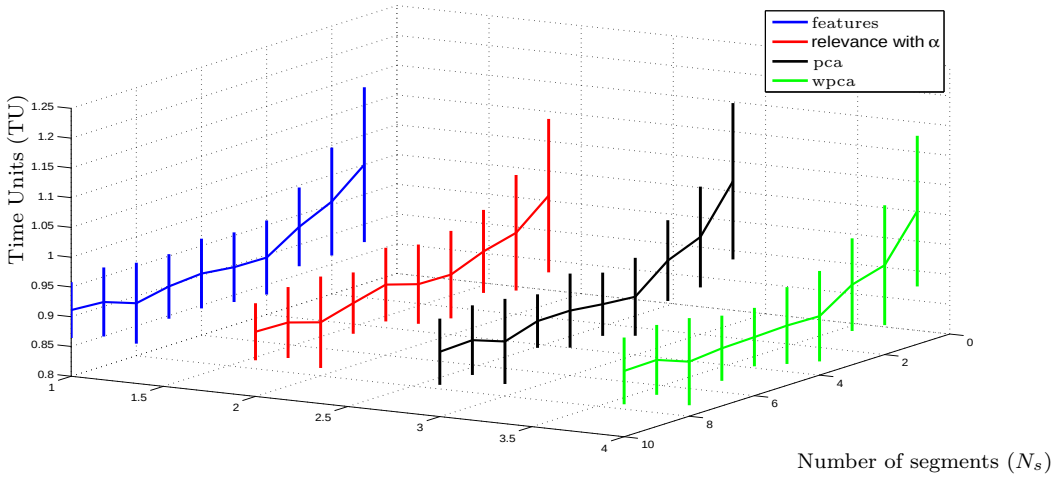


Figure 9.37: Comparison between proposed method and the reference methods using error bars, which measure processing time and performance

Table 9.8 depicts the total system performance using the non–supervised indices,

Table 9.8: Clustering performance using non–supervised indices (f_1/f_2, $\epsilon_{\boldsymbol{B}}$), number of resultant groups (k_{end}), number of relevant features (q) and processing time [s].

	f_1/f_2	$\epsilon_{\boldsymbol{B}}$	k_{end}	q	Time [s]
	$\mu - \sigma^2$				
$\boldsymbol{\alpha}$	0.95 - 0.08	0.86 - 0.12	9 - 2.43	6 - 0.86	37.20 - 7.65
$\boldsymbol{\rho}$	0.94 - 0.10	0.82 - 0.14	11 - 2.67	10 - 1.78	38.52 - 7.52
$\widehat{\boldsymbol{\rho}}$	0.94 - 0.10	0.83 - 0.14	11 - 2.66	9 - 1.77	37.47 - 7.66
$\widehat{\boldsymbol{\alpha}}$	0.95 - 0.08	0.84 - 0.12	9 - 2.44	6 - 1.16	36.20 - 7.54

as discussed in Section 7.6, but parameters for computational cost and the number of resulting groups are displayed, as well. The first column regards to the index f_1/f_2, the second one to $\epsilon_{\boldsymbol{B}}$, the third one corresponds to the resulting clusters (k_{end}) after processing all segments of each recording, the fourth one is the number of selected features before projecting the data, and the last one is the time fixed from filtering to clustering stages. The rows show the considered methods of weighting of variables.

Evaluation of the quality of partition is provided by the index f_1/f_2 that have been introduced in [153], where a maximum index value of $f_1/f_2 = 0.98$ was achieved over only 14 recordings of the entire MIT/BIH database. In this work, the maximum index attained is close to (~ 0.96), but over the whole recording set if providing framed division analysis, thus, pointing out to better generalization ability. If increasing the number of groups k, then index f_1/f_2 tends to 1. Nonetheless, a very high number of k leads to more difficult evaluation of resulting groups by the specialist. Specifically, in [8], the total number of groups representing the entire MIT/BIH database is assumed to be 25, while in [10], when considering only some recordings from MIT/BIH database, the average number of groups diminishes to 15. In this work, for the entire MIT/BIH database the value k ranges within $9 \leq k \leq 11$, still showing better performance.

As discussed above, the non–supervised measure $\epsilon_{\boldsymbol{B}}$ penalizes the number of groups after the final partition. In fact, if the clustering procedure is carried out for a value less than or equal to the proper value of k, then $\epsilon_{\boldsymbol{B}}$ tends to 1. Otherwise, this value becomes far from 1, as the amount of k is increased, due to the upper bound monotonicity theorem [150]. In concrete case, a value of k less than 6 can be taken as an admisible number of groups because 5 classes are considered in the present arrhythmia analysis.

Table 9.8 shows that the achieved values for $\epsilon_{\boldsymbol{B}}$ are ranging within interval $0.82 \leq \epsilon_M \leq 0.86$, which can be considered as a realistic outcome to quantify the resultant partition. Still, this measure does not reach the value 1, as if grouping were absolutely

Table 9.9: Clustering results by using supervised measures for all groups of arrhythmias

	N $\mu-\sigma^2$	S $\mu-\sigma^2$	V $\mu-\sigma^2$	F $\mu-\sigma^2$	Q $\mu-\sigma^2$
			α		
Se	99.25− 2.48	91.11−15.66	96.11− 8.24	70.73−32.05	91.9 −17.83
Sp	95.77− 9.12	99.36− 2.19	99.87− 0.24	99.59− 0.77	99.79− 0.47
CP	99.16− 2.5	96.18− 6.15	98.29− 3.37	93.29−10.62	99.91− 0.21
			$\widehat{\alpha}$		
Se	99.06− 3.5	90.52−16.54	96.31− 7.03	50.59−35.33	87.88−26.77
Sp	95.32− 9.49	99.49− 1.62	99.84− 0.31	99.51− 0.99	99.6 − 0.73
CP	99.25− 2.01	96.24− 6.67	98.13− 3.86	81.64−32.51	99.96− 0.06
			$\widehat{\rho}$		
Se	99.27− 2.55	91.26−15.26	91.5 −21.52	50.57−34.63	79.91−44.67
Sp	92.19−19.06	99.5 − 1.51	99.86− 0.26	99.6 − 0.76	99.85− 0.34
CP	99.14− 2.46	95.68− 7.47	93.71−20.77	81.05−33.04	79.93−44.68
			ρ		
Se	99.24− 2.79	90.08−18.57	91.62−21.61	42.44−39.57	87.86−26.75
Sp	92.17−19.37	99.58− 1.22	99.85− 0.28	99.53− 0.92	99.79− 0.47
CP	99.12− 2.58	96.16− 6.07	93.77−20.79	68.87−41.56	99.91− 0.21

correct, because of the effect of penalization regarding the number of groups and also the sensitivity to the affinity matrix. Besides, the average number is shown for the relevant features q before the WPCA projection (see Section 6) that most contribute to the clustering process. As seen, the range of values of q is $6 \leq q \leq 10$, showing admisible values with respect to the total number of features. Additionally, the average computing time needed to process the whole recording is 37 s, being a reasonable time for each considered recording.

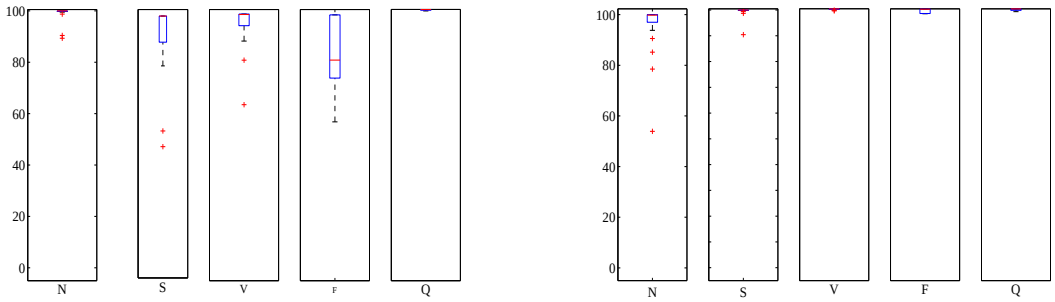
Lastly, Table 9.9 shows the arithmetic mean (μ) and variance (σ^2) of supervised measures, discussed in Section 7.6, over the entire database for each group of arrhythmias, using the proposed weighting methods (see Algorithm 2). In [10], the clustering performance is evaluated only considering three arrhythmia groups, namely, N, S and V (table 2.2). Nonetheless, the performance measures are calculated by couples of arrhythmias and, as consequence, the value of measures tends to increase. The reported results are $Se_S = 93.3\%$ and $Sp_S = 99.5\%$, which can be compared with the second column of Table 9.9, where the maximum values are $Se_S = 91.3\%$ and $Sp_S = 99.5\%$. Although sensitivity is minor than the reported in [10], it should be noted that among results of this work, all recordings from the database with S-type arrhythmias (the class of interest) and the remaining groups as another class are analyzed. Thus, the proposed method provides more robustness when considering other types of arrhythmias.

For some considered arrhythmia groups, the performance results in table 9.9 became

remarkable, e.g. $Se_N = 99.25\%$, $Se_V = 96.5\%$. In other cases, e.g. F group, the performance values decreases, $Se_F = 70.7\%$, due to the low number of representative heartbeats of some classes. Yet, one can infer that the best performance is provided when the data are weighted by using $\boldsymbol{\alpha}$.

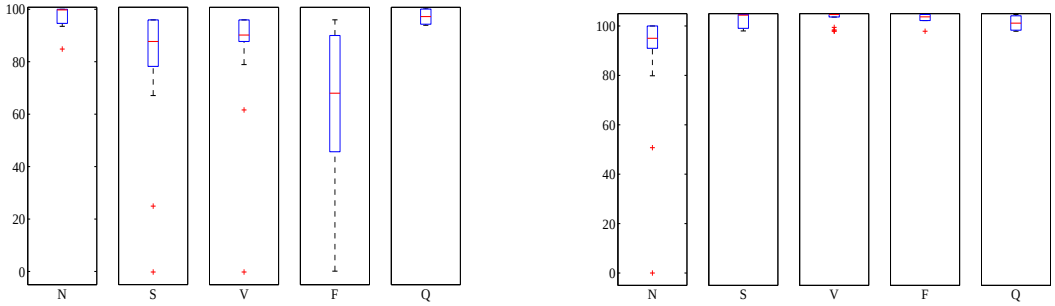
9.4 General Methodology Results

In Tables 9.10 and 9.11, are shown the obtained results. Figures 9.38 and 9.39 correspond to a graphic form for the same results.



(a) Se (b) Sp

Figure 9.38: Performance results of NPDBC with $Q - \alpha$



(a) Se (b) Sp

Figure 9.39: Performance results of NPDBC with MSE-based feature selection

As can be seen, segment analysis improves the clustering performance, in comparison with traditional analysis (i. e., analyzing whole data set in one segment), mainly

Table 9.10: Results for full segment methodology using NPDBC as clustering method and $Q - \alpha$ as feature selection procedure

Rec.		N	S	V	F	Q	Time (s)	f_1	f_2	f_1/f_2	p	k_{end}	ε_M
100	Beats	2237	33	1	0	0							
	Se	100	100	100			39.64	-1634.54	-1634.54	1	6	12	0.93
	Sp	100	100	100									
101	Beats	1858	3	0	0	2							
	Se	100	100			100	32.11	-1265.34	-1265.34	1	5	12	0.91
	Sp	100	100			100							
103	Beats	2080	2	0	0	0							
	Se	100	100				35.9	-1407.01	-1407.01	1	5	7	0.93
	Sp	100	100										
106	Beats	1506	0	520	0	0							
	Se	99.67		99.81			34.75	-962.07	-967.27	0.99	5	8	0.93
	Sp	99.81		99.67									
107	Beats	0	0	59	0	2076							
	Se			100		100	36.97	-967.79	-967.79	1	5	8	0.91
	Sp			100		100							
111	Beats	2121	0	1	0	0							
	Se	100		100			36.62	-1474.27	-1474.27	1	4	4	0.95
	Sp	100		100									
112	Beats	2535	2	0	0	0							
	Se	100	100				44.25	-1793.16	-1793.16	1	5	7	0.91
	Sp	100	100										
113	Beats	1787	6	0	0	0							
	Se	100	100				30.68	-1211.32	-1211.32	1	5	11	0.91
	Sp	100	100										
118	Beats	2164	96	16	0	0							
	Se	99.95	96.84	64.29			39.28	-1727.78	-1922.98	0.89	5	6	0.71
	Sp	98.17	99.72	99.96									
119	Beats	1541	0	444	0	0							
	Se	100		100			34.78	-754.8	-754.8	1	6	5	0.93
	Sp	100		100									
121	Beats	1859	1	1	0	0							
	Se	100	100	100			31.37	-1284.33	-1284.33	1	5	10	0.92
	Sp	100	100	100									
123	Beats	1513	0	3	0	0							
	Se	100		100			25.68	-979.14	-979.14	1	5	12	0.94
	Sp	100		100									
124	Beats	1534	31	47	5	0							
	Se	99.93	80	89.36	75		27.76	-1213.84	-1445.01	0.84	6	8	0.71
	Sp	85.19	99.94	100	100								
201	Beats	1633	128	198	2	0							
	Se	99.51	88	96.97	100		33.85	-1474.16	-1638.04	0.9	7	8	0.8
	Sp	94.44	99.4	100	100								
207	Beats	1542	106	210	0	0							
	Se	99.81	100	99.52			21.7	-1777.18	-1777.18	0.99	6	12	0.9
	Sp	99.68	100	99.82									
210	Beats	2421	22	195	10	0							
	Se	99.83	47.62	92.78	80		46.78	-1666.93	-1742.51	0.79	6	9	0.62
	Sp	93.78	99.92	99.51	99.89								
213	Beats	2639	28	220	362	0							
	Se	99.54	96.3	81.82	57.73		57.95	-1916.98	-2081.31	0.92	7	13	0.79
	Sp	78.49	99.97	99.21	98.27								
214	Beats	2000	0	256	1	2							
	Se	100		99.22	100	100	39.39	-1215.36	-1215.36	0.99	6	10	0.9
	Sp	99.22		100	100	100							
217	Beats	244	0	162	260	1540							
	Se	89.34		94.9	84.23	99.48	39.41	-1340.65	-1341.88	0.94	7	10	0.84
	Sp	98.31		99.61	98.19	98.94							
220	Beats	1951	94	0	0	0							
	Se	100	97.83				35.66	-1331.81	-1331.81	0.99	5	8	0.88
	Sp	97.83	100										
221	Beats	2029	0	396	0	0							
	Se	99.85		99.49			42.69	-1534.08	-1534.08	0.99	6	9	0.87
	Sp	99.49		99.85									
222	Beats	2272	209	0	0	0							
	Se	98.72	53.92				44.56	-1961.96	-2746.74	0.71	7	5	0.52
	Sp	53.92	98.72										
230	Beats	2253	0	1	0	0							
	Se	100		100			41.2	-1658.06	-1658.06	1	6	8	0.93
	Sp	100		100									
231	Beats	1566	1	2	0	0							
	Se	100	100	100			34.25	-1399.04	-1399.04	1	6	9	0.9
	Sp	100	100	100									
232	Beats	397	1381	0	0	0							
	Se	90.43	100				31.17	-1575.12	-1.7698	0.9	7	7	0.66
	Sp	100	90.43										
234	Beats	2698	50	3	0	0							
	Se	100	90	100			48.71	-1873.5	-1943.55	0.93	7	6	0.69
	Sp	90.57	100	100									
Σ	$\mu(Se)$	99.06	91.21	95.7	82.83	99.87	$\mu(\cdot)$			$\mu(\cdot)$	$\mu(\cdot)$	$\mu(\cdot)$	$\mu(\cdot)$
	$\mu(Sp)$	95.56	99.3	99.88	99.39	99.74	37.2			0.95	6	9	0.84

Table 9.11: Results for full segment methodology using NPDBC as clustering method and *WPCA-MSE* approach as feature selection procedure

Rec.		N	S	V	F	Q	Time (s)	f_1	f_2	f_1/f_2	p	k_{end}	ε_M
100	Beats	2237	33	1	0	0							
	Se	94	94	94			39.53	-1791.07	-1791.07	1	10	15	0.9
	Sp	94	94	94									
101	Beats	1858	3	0	0	2							
	Se	94.3	94.3			94.3	32.4	-1268.76	-1268.76	1	9	14	0.89
	Sp	94.3	94.3			94.30							
103	Beats	2080	2	0	0	0							
	Se	100	100				36.3	-1443.28	-1443.28	1	10	9	0.9
	Sp	100	100										
106	Beats	1506	0	520	0	0							
	Se	99.73		100			35.14	-951.91	-9547.05	0.99	10	9	0.87
	Sp	100		99.73									
107	Beats	0	0	59	0	2076							
	Se			100		100	37.6	-1046.24	-1046.24	1	9	10	0.9
	Sp			100		100							
111	Beats	2121	0	1	0	0							
	Se	100		100			36.81	-1474.3	-1474.3	1	8	9	0.9
	Sp	100		100									
112	Beats	2535	2	0	0	0							
	Se	100	100				44.65	-1793.13	-1793.13	1	8	10	0.92
	Sp	100	100										
113	Beats	1787	6	0	0	0							
	Se	100	100				30.82	-1211.39	-1211.39	1	9	10	0.91
	Sp	100	100										
118	Beats	2164	96	16	0	0							
	Se	99.91	97.89	64.29			39.92	-1709.49	-1997.6	0.86	10	14	0.69
	Sp	98.17	99.72	99.96									
119	Beats	1541	0	444	0	0							
	Se	100		100			34.88	-783.16	-783.16	1	9	9	0.89
	Sp	100		100									
121	Beats	1859	1	1	0	0							
	Se	93.6	93.6	93.6			31.5	-1280.23	-1280.23	1	10	16	0.88
	Sp	93.6	93.6	93.6									
123	Beats	1513	0	3	0	0							
	Se	100		100			25.68	-1014.66	-1014.66	1	8	9	0.92
	Sp	100		100									
124	Beats	1534	31	47	5	0							
	Se	99.93	80	89.36	75		28.16	-1213.01	-1461.5	0.83	9	11	0.66
	Sp	85.19	99.94	100	100								
201	Beats	1633	128	198	2	0							
	Se	99.51	80.83	94.44	0		33.88	-1471.78	-1839.72	0.8	10	11	0.63
	Sp	89.69	99.4	100	100								
207	Beats	1542	106	210	0	0							
	Se	94	94	94			21.85	-196.2	-196.2	1	10	15	0.9
	Sp	94	94	99									
210	Beats	2421	22	195	10	0							
	Se	99.75	33.33	91.24	44.44		46.83	-1773.79	-2073.34	0.85	8	16	0.58
	Sp	91.07	99.96	99.27	99.89								
213	Beats	2639	28	220	362	0							
	Se	99.47	100	82.27	57.43		57.81	-1885.29	-2049.23	0.92	9	12	0.83
	Sp	79.83	99.97	99.07	98.23								
214	Beats	2000	0	256	1	2							
	Se	93.5		93.5	93.5	93.5	39.32	-1396.67	-1396.67	1	10	16	0.9
	Sp	93.5		93.5	93.5	93.5							
217	Beats	244	0	162	260	1540							
	Se	84.84		91.98	87.69	99.28	39.29	-1372.59	-1373.85	0.99	9	12	0.9
	Sp	98.32		99.66	97.63	98.95							
220	Beats	1951	94	0	0	0							
	Se	99.9	96.81				36.17	-1327.62	-1327.62	0.99	10	9	0.9
	Sp	96.81	99.9										
221	Beats	2029	0	396	0	0							
	Se	100		99.49			43.42	-1534.98	-1534.98	0.99	10	9	0.91
	Sp	99.49		100									
222	Beats	2272	209	0	0	0							
	Se	98.72	50.75				44.84	-1956.25	-2934.38	0.67	9	7	0.49
	Sp	50.75	98.72										
230	Beats	2253	0	1	0	0							
	Se	100		0			42.35	-1745.38	-2618.07	0.67	9	10	0.5
	Sp	0		100									
231	Beats	1566	1	2	0	0							
	Se	95	95	95			34.82	-1402.27	-1402.27	1	9	13	0.92
	Sp	95	95	95									
232	Beats	397	1381	0	0	0							
	Se	94.71	100				31.25	-674.52	-691.669	0.98	8	10	0.91
	Sp	100	94.71										
234	Beats	2698	50	3	0	0							
	Se	100	90	100			49	-1999.52	-1999.52	1	1	12	0.93
	Sp	90.57	100	100									
Σ	$\mu(Se)$	97.64	88.27	88.59	59.68	96.77	$\mu(\cdot)$			$\mu(\cdot)$	$\mu(\cdot)$	$\mu(\cdot)$	$\mu(\cdot)$
	$\mu(Sp)$	89.77	97.84	98.57	98.21	96.68	37.47			0.94	9	11	0.83

because reduces the probability of wrongly grouping. Indeed, if estimating the number of groups k by considering whole data set, it would exist k groups that could not enough to separate the classes into data set. In other hand, by estimating the number of groups per each segment general procedure is less sensitive to minority classes.

Regarding feature selection stage, it can be noted that two used methods shows similar results, without taking into account penalization factor for number of groups. However, algorithm $Q-\alpha$ is preferable because generates a less number of relevant features and then computational cost can be decreased. Also, it is noted that estimation of the number of groups and measure $\varepsilon_{_M}$ becomes better values when employing resultant parameters from $Q-\alpha$ method. Affinity matrix resultant from algorithm $Q-\alpha$ is obtained before an iterative refining, that is to say, taking into account intrinsically information about data. For this reason, it could operate with better performance than other affinity matrices.

Non-supervised measure $\varepsilon_{_M}$ proved to be a good index to quantify the quality of resultant grouping because taking into consideration the number of groups and achieves a value near to 1 when grouping is correct, but is sensitive to affinity matrix.

In Table 9.11, recording 100 is one of recordings that the fact of penalization factor is observed. Note that $f_1/f_2 = 1$, then, it should be expected values of Sp and Se closely near to 100%, but this does not occur because the resultant number of groups is bigger than admisible and therefore measures is penalized.

Finally, in Figures 9.40 and 9.41, an example of clustered heartbeats, is presented by using the recording 217 as test that in accordance with the Table 9.4, contains heartbeats of type: N, V, f and P (in the Figures, the symbol P is represented as /, which corresponds to the original tag for paced beats in the MIT/BIH database).

Figures show the QRS complexes associated to each cluster and in addition it is shown both the heartbeat class and the number of heartbeats per each cluster.

Table 9.12 depicts the confusion matrix of the clustering procedure and Table 9.13 depicts the clustering performance. Although, the sensitivity is less than 90% to classes V and P, all heartbeat types are identified with the proposed methodology. The low value of sensitivity is associated with the low number of heartbeats in the recording regarding majority classes (for this case, types N and f).

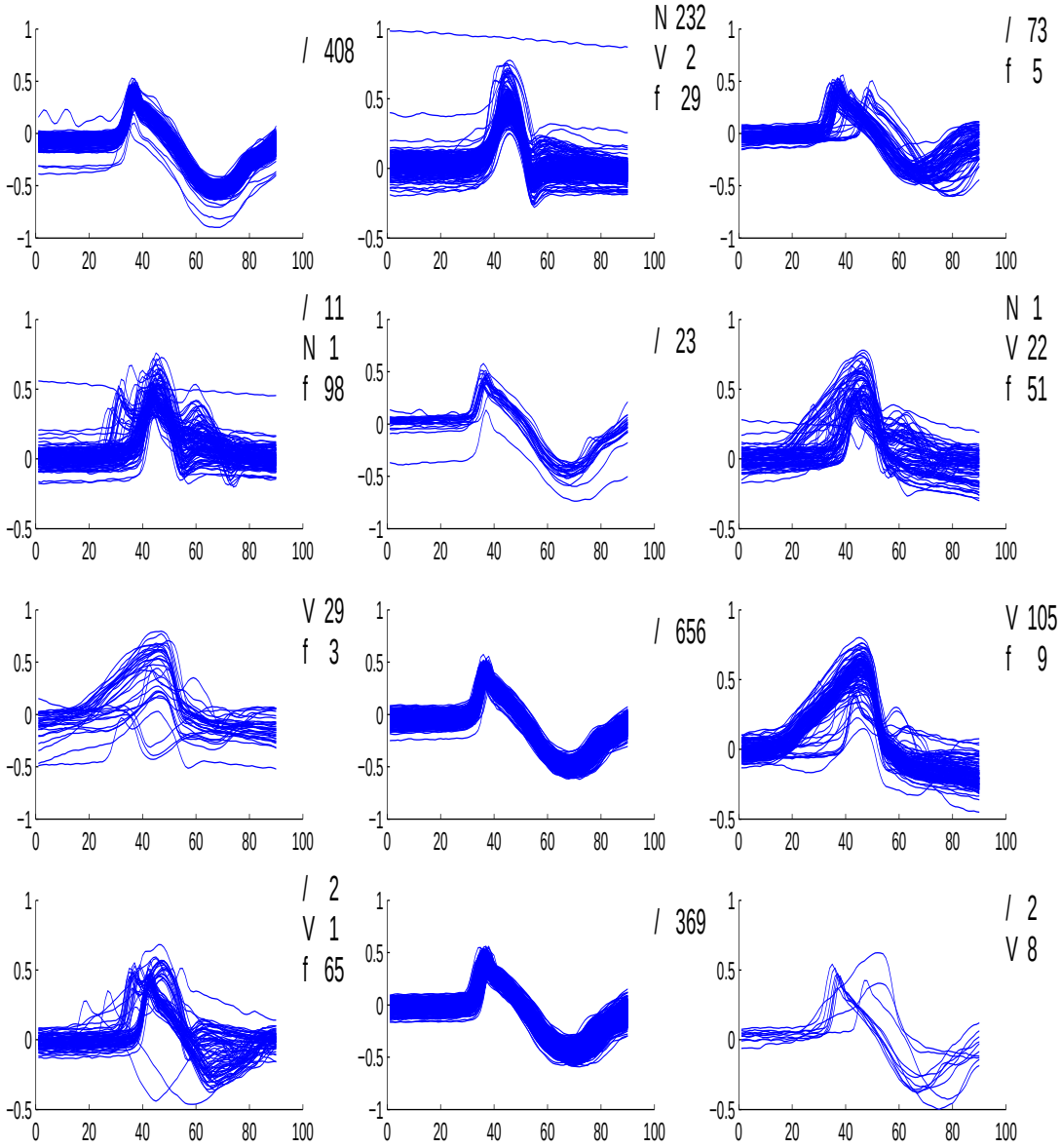


Figure 9.40: General methodology applied over recording 217.

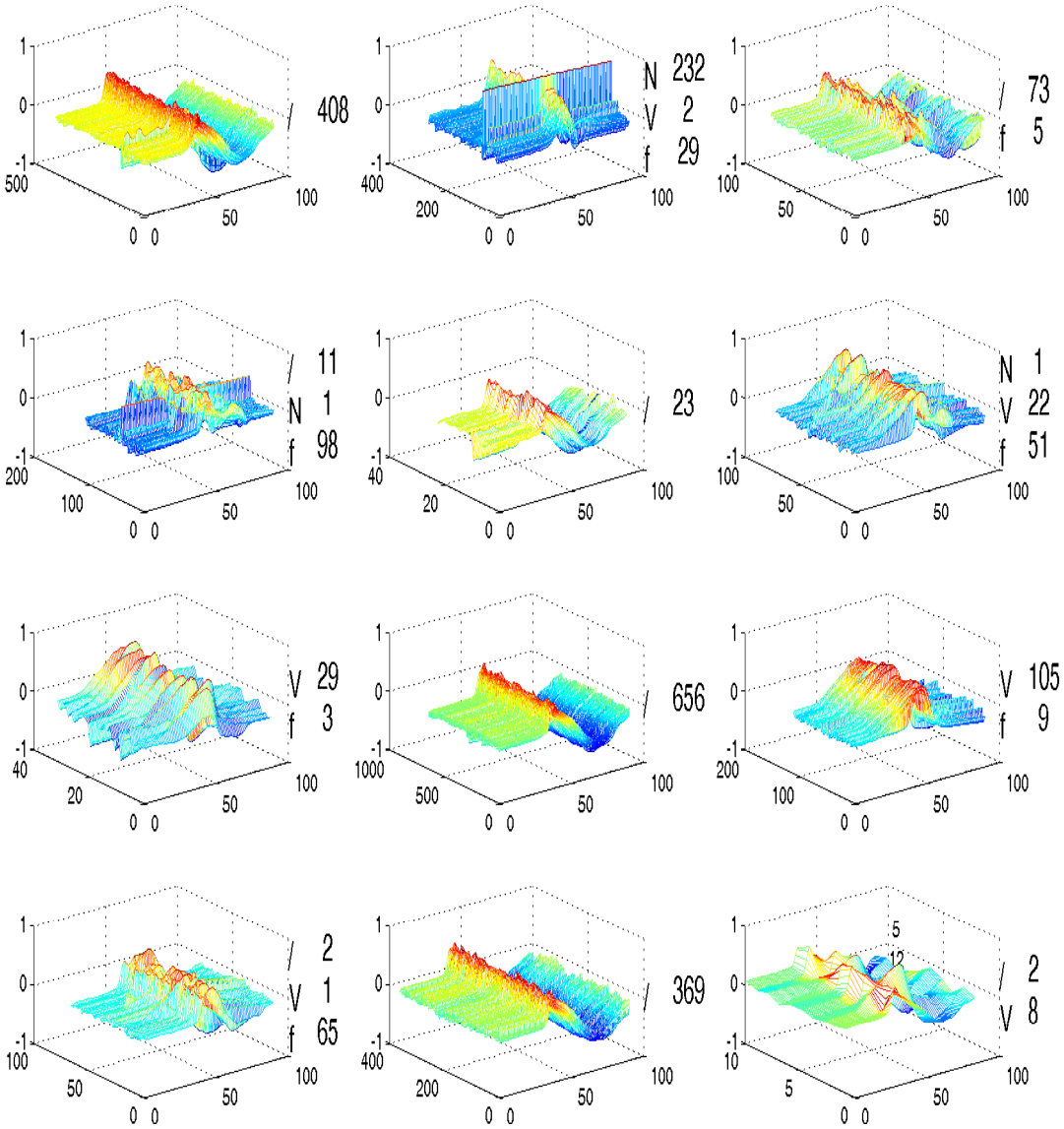


Figure 9.41: General methodology applied over recording 217.

Table 9.12: Confusion matrix of the clustered heartbeats

	N	V	f	P(/)
N	**232**	2	0	29
V	0	**134**	0	12
f	0	2	**1537**	5
P(/)	12	24	2	**214**

Table 9.13: clustering performance (sensitivity and specificity) of the recording 217

Measure	N	V	f	P(/)
Se	95.0820	82.7160	99.8700	82.3077
Sp	98.4192	99.4126	98.9489	98.0463

Part IV

Conclusions and Contributions

Chapter 10

Conclusions and Future Work

10.1 Conclusions

The proposed methodology for unsupervised Holter monitoring of cardiac arrhythmias that is based on variable–wise relevance analysis and CBC-based clustering leads to an improvement in determining those heartbeat types recommended by AAMI. Because of strong asymmetry among class observations, to provide convenient separability among certain types of cardiac arrhythmias, the heartbeat–derived features are properly selected by their weighted projection, allowing the assessment of the relative importance of each feature immersed in the original data representation. In addition, due to restrictions for reducing computational load, the proposed methodology is carried out by successive division analysis throughout time, where each recording is separately processed, thus significantly reducing the processing time.

It must be noted that the relevance analysis provides enough generalization capability, mainly because most informative features are weighted and projected. In general, in this study, the M-inner product-based approach performs better than MSE-based approach, and although its iterative nature leads to high computational cost, the segment analysis compensates for this effect, permitting its implementation in real time applications. Moreover assumed grouping that includes initial parameter estimation, i.e. the estimation of number of groups and center initialization, which is based on spectral techniques and soft partitional clustering, generates a proper final partition.

In summary, the proposed methodology provides an useful tool for analyzing cardiac arrhythmias with suitable quality since it is based on non-supervised training, that is, there is no need for labelling recordings that are generally not feasible for Holter

monitoring. In addition, testing the developed methodology by using introduced cluster validity measures shows a comparable performance in comparison to other referenced studies, based on either supervised or unsupervised training on the MIT/BIH database.

10.2 Future work

As future work, several aspects of this methodology can be studied in order to improve and expand the capabilities proposed in this study.

In the filtering stage, the possibility of filtering both interferences and low/high frequency noise with multiresolution denoising can be explored, by studying the proper selection of wavelet mother, level of decomposition, level of noise, thresholding types, among others.

Although a good performance was obtained in the R-peak detection, the segmentation of ECG waves was not performed, mainly due to the fact that arrhythmia morphologies concentrate the information in the QRS complex, however in order to detect other pathologies that present changes in P or T waves of ECG signals, it is necessary to develop a segmentation methodology.

Regarding relevance procedures, there is the possibility of studying the behavior of "non relevant" features if the "relevant" are suppressed in the feature set. This is intended to analyze features that can contain hidden and relevant information to detect some type of biosignal pathology.

In the clustering stage, additional spectral clustering stages should be explored with the possibility of unifying the stages of feature selection and clustering, in order to further improve the accuracy and the computational load for the Holter monitoring system.

Chapter 11

Contributions of this Thesis

It is important to highlight the main contributions of this work, which are:

- In the R-peak detection stage, the Shannon energy envelope, is proposed as nonlinear transformation, with the aim of improving the detection of peaks with low amplitude, obtaining better results in the sensitivity measure.

- The formalization and application of the MSE-based relevance measure for the one-dimensional case in biosignal analysis, in order to select relevant features for detecting cardiac arrhythmias.

- The ability to select a subset of relevant features and its projection into a new space in order to cluster the data set in a proper final partition by means of the M-inner product-based method. This method is non-supervised and does not require the setup of parameters that fit in new models of ECG signals.

- There is a connection between the relevance analysis with Q-α algorithm and the clustering stage, if the objective function is evaluated after the convergence of the non supervised feature selection algorithm, estimating directly the number of groups in segments associated with the analyzed recording and reducing the total processing time.

- The non-supervised clustering scheme, that incorporates estimation of the number of groups, initialization criteria and partitional soft-grouping stages, is improved by including an additional stage that corresponds to the segment analysis based on the study of dissimilarity measures between two contiguous segments.

- The inclusion of a soft membership function in the grouping stage, improves the final partition after a proper center initialization, converging into a better-sited local optimal of the clustering objective function.
- The inclusion of non-supervised indices to assess the final partition quality based on spectral information of transformed data.
- The proposed methodology is non-supervised in all stages and fulfil with the AAMI standards in arrhythmia analysis. The procedure has the advantage of not requiring parameter fitting. Finally, with the segment analysis, it is possible to implement the methodology in real time applications.

Part V

Appendixes

Appendix A

Academic Discussion

During the course of this work, the results of the thesis were presented in both national and international conferences as well as in both national and international journals. Also, a Master Thesis and a project were developed as result of this research process.

A.1 Papers in International Conference Proceedings

1. Segment Clustering For Holter Recording Analysis, ***JL. Rodriguez-Sotelo****, D. Peluffo-Ordoñez, D. Cuesta-Frau, G. Castellanos-Dominguez*, accepted for presentation at the Computers in Cardiology (CINC) 2010, Belfast, Ireland.

2. Weighted-PCA for Unsupervised Analysis of Cardiac Arrhythmias, ***JL. Rodriguez-Sotelo****, D. Peluffo-Ordoñez, D. Cuesta-Frau, G. Castellanos-Dominguez*, accepted for presentation at the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'10)), Buenos Aires, Argentina, September 2010.

3. Nonparametric Density-Based Clustering for Cardiac Arrhythmia Analysis, ***JL. Rodriguez-Sotelo****, D. Peluffo-Ordoñez, D. Cuesta-Frau, G. Castellanos-Domínguez.* Computers in Cardiology (CINC) 2009. Park-City Utah, USA.

4. Unsupervised feature selection in cardiac arrhythmias analysis. ***JL. Rodríguez-Sotelo****, D. Cuesta-Frau, D. Peluffo-Ordóñez, G. Castellanos-Domínguez.* Engineering in Medicine and Biology Conference. EMBC, 2009. Minneapolis, USA.

5. Clasificación de arritmias en registros Holter empleando análisis no supervisado de tipo particional y espectral. *D. Peluffo-Ordóñez,* ***JL. Rodríguez-Sotelo****, G. Castellanos-Domínguez*. V Seminario Internacional de Ingeniería Biomédica (SIB). Universidad de los Andes, Bogotá, 2009.

6. Atrial premature beats detection using heart rate variability analysis and Fuzzy C-means clustering. ***JL. Rodríguez-Sotelo****, D. Cuesta-Frau, D. Peluffo-Ordóñez, G. Castellanos-Domínguez*. 30th IEEE Annual International Conference (EMBC'08), Canada, August 2008.

7. Agrupamiento no supervisado de latidos ECG usando características WT, Dynamic Time Warping y k-means modificado. ***J.L Rodríguez****, D. Cuesta, G. Castellanos*. Memorias del IV Congreso Latinoamericano de Medicina Biomédica, Isla Margarita, Venezuela, Septiembre 2007.

8. An improved method for unsupervised analysis of ecg beats based on WT features and J-means. ***J.L. Rodríguez****, D. Cuesta, G. Castellanos*. Memorias del congreso Computers In Cardiology 2007, ISSN: 0276-6547, ISBN: 978-1-4244-2533-4, Digital Object Identifier: 10.1109/CIC.2007.4745552, Durham, North Carolina, 30 Sep - 3 Oct. 2007.

9. Recognition of Caridiac Arrhythmias by Means of Beat Clustering on ECG-Holter Records. *E. Delgado,* ***J.L. Rodríguez****, F. Jiménez, D. Cuesta, G. Castellanos.* Memorias del congreso Computers In Cardiology 2007, ISSN: 0276-6547, ISBN: 978-1-4244-2533-4, INSPEC Accession Number: 10454600, Digital Object Identifier: 10.1109/CIC.2007.4745446, Durham, North Carolina, 30 Sep. - 3 Oct. 2007.

10. Estimación de tasa de compresión para transmisión de Fonocardiogramas. ***J.L. Rodríguez****, J.C. García*, Memorias del congreso de la sociedad española de ingeniería Biomédi-ca (CASEIB), Noviembre, 2007.

11. Remoción de interferencia AC en ECG, usando Denoising y la medida Weighted Diagnostic Distortion (WDD). *L.E. Avendaño,* ***JL. Rodríguez****, G. Castellanos.* Memorias del congreso anual, CASEIB. España, Noviembre 2006.

A.2 Articles in both International and National Journals

1. Analysis of relevance for unsupervised holter monitoring of cardiac arrhythmias, ***JL Rodriguez-Sotelo****, D Peluffo-Ordoñez, D Cuesta-Frau, G Castellanos-Dominguez.* Physiological Measurement IOP, (in revision), 2010.

2. Unsupervised classification of atrial premature hearbeats using a prematurity index and wave morphology features. ***J.L. Rodríguez-Sotelo****, D. Cuesta-Frau, G. Castellanos-Domínguez.* Medical & Biological Engineering & Computing. ISSN: 0140-0118, Volume 47, Number 7, p.p. 731-741 / julio de 2009

3. Estudio comparativo de métodos de selección de caracaterísticas de inferencia supervisada y no supervisada Revista Tecno Lógicas. *D. Peluffo-Ordóñez,* ***JL. Rodríguez-Sotelo****, G. Castellanos-Domínguez*, ISSN: 0123-7799, 2009.

4. Remoción y evaluación de la interferencia AC en ECG, usando Denoising y la medida Weighted Diagnostic Distortion (WDD). *L.E. Avendaño, G. Castellanos-Domínguez,* ***JL. Rodríguez-Sotelo***. Revista Nacional Indexada, Scientia et Technica, Año XII, No. 31, Agosto 2006 UTP. ISSN 0122-1701.

A.3 Master Thesis and Project

- Estudio comparativo de métodos de agrupamiento no supervisado de latidos en señales ECG. *Diego Hernán Peluffo-Ordoñez.* Advisor: Germán Castellanos-Domínguez. Master Thesis in Industrial Automation. Universidad Nacional de Colombia, sede Manizales, 2010.

- Análisis de bioseñales en la identificación de arritmias mediante técnicas no supervisadas. Convocatoria Nacional de Investigación 2009. Apoyo a Tesis de Posgrado. Cod. 10242. Main Researcher: *G. Castellanos.* 2009-2010.

Appendix B

Sparsity and Positivity of α

Before technical issues, it is worthwhile to make qualitative arguments (which are the basis of developing this approach to begin with) as to the underlying reason for sparsity. Consider rewriting the optimization criterion (6.16) by an equivalent criterion:

$$\min_{\boldsymbol{\alpha},\boldsymbol{Q}} \left\{ \|\boldsymbol{A}_{\boldsymbol{\alpha}} - \boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{A}_{\boldsymbol{\alpha}}\|_F^2 - \|\boldsymbol{A}_{\boldsymbol{\alpha}}\|_F^2 \right\}, \tag{B.1}$$

where $\|\cdot\|_F^2$ is the square Frobenius norm of a matrix defined as the sum of squares of all entries of the matrix. The first term of (B.1) measures the distance between the columns of $\boldsymbol{A}_{\boldsymbol{\alpha}}$ and the projection of those columns onto a q-dimensional subspace (note that $\boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}}$ is a projection matrix). This term receives a low value if indeed $\boldsymbol{A}_{\boldsymbol{\alpha}}$ has a small q number of dominant eigenvectors, i.e., the spectral properties of the feature subset represented by $\boldsymbol{A}_{\boldsymbol{\alpha}}$ are indicative to a good clustering score.

Since $\boldsymbol{M} = \boldsymbol{X}^{\mathsf{T}}$ and $\boldsymbol{A}_{\boldsymbol{\alpha}}$ can be expressed as $\boldsymbol{A}_{\boldsymbol{\alpha}} = \sum_i^p \alpha_i \boldsymbol{m}_i \boldsymbol{m}_i^{\mathsf{T}}$ is represented by the sum of rank-1 matrices, only a *small* number of them can be combined if the first term is desired to be small. The second term (which may be viewed also as a regularization term) encourages addition of more rank-1 matrices to the sum provided they are redundant, i.e., are already spanned by the previously selected rank-1 matrices. This makes the point that the feature selection scheme looks for relevant features but not necessarily the minimal set of relevant features. To summarize, from a qualitative point of view the selection of values for the weights α_i is directly related to the rank of the affinity matrix $\boldsymbol{A}_{\boldsymbol{\alpha}}$ which should be small if indeed $\boldsymbol{A}_{\boldsymbol{\alpha}}$ arises from a clustered configuration of data points. A uniform spread of values α_i would result in a high rank for $\boldsymbol{A}_{\boldsymbol{\alpha}}$, thus the criteria function encourages a non-uniform (i.e., sparse) spread of weight values.

The argument presented above is in order to facilitate clarity of the approach and should not be taken as a proof for sparsity. The positivity and sparsity issues are approached in the sequel from a different angle which provides a more analytic handle to the underlying search process than the qualitative argument above.

B.1 Positivity of α

The key for the emergence of a sparse and positive $\boldsymbol{\alpha}$ has to do with the way the entries of the matrix $\boldsymbol{G}$ are defined. Recall that $g_{ij} = (\boldsymbol{m}_i^\mathsf{T}\boldsymbol{m}_j)\boldsymbol{m}_i^\mathsf{T}\boldsymbol{Q}\boldsymbol{Q}^\mathsf{T}\boldsymbol{m}_j$ and that α comes out as the leading eigenvector of $\boldsymbol{G}$ (at each iteration). If $\boldsymbol{G}$ were to be non-negative (and irreducible), then from the Perron-Frobenius theorem the leading eigenvector is guaranteed to be non-negative (or same-sign). However, this is not the case and $\boldsymbol{G}$ in general has negative entries as well as positive ones. However, from a probabilistic point of view the probability that the leading eigenvector of $\boldsymbol{G}$ will come out positive rapidly approaches 1 with the growth of the number of features — this under a couple of simplifying assumptions.

The simplifying assumptions are made in order to derive a probabilistic argument, is first that the entries of the upper triangular part of $\boldsymbol{G}$ are independent. The second simplifying approximation is that the columns of $\boldsymbol{Q}$ are sampled uniformly over the unit hypersphere. Although the independence and uniformity assumptions are indeed an idealization of the true nature of $\boldsymbol{G}$ and $\boldsymbol{Q}$, they nevertheless allow us to derive a powerful probabilistic argument which shows in a rigorous manner that the weights α_i are non-negative with probability 1. A statement which agrees with practice over extensive experimentations which have been performed.

The probabilistic approach follows from the observation that each entry of $\boldsymbol{G}$ consists of a sum of products of three inner-products:

$$g_{ij} = \sum_{l=1}^{q} (\boldsymbol{m}_i^\mathsf{T}\boldsymbol{q}_l)(\boldsymbol{m}_j^\mathsf{T}\boldsymbol{q}_l)(\boldsymbol{m}_i^\mathsf{T}\boldsymbol{m}_j). \tag{B.2}$$

In general, a product of the form $f = (\boldsymbol{a}^\mathsf{T}\boldsymbol{b})(\boldsymbol{a}^\mathsf{T}\boldsymbol{c})(\boldsymbol{b}^\mathsf{T}\boldsymbol{c})$, where $\|\boldsymbol{a}\| = \|\boldsymbol{b}\| = \|\boldsymbol{c}\| = 1$ satisfies $-1/8 \leq f \leq 1$ where $f = 1$ when $\boldsymbol{a} = \boldsymbol{b} = \boldsymbol{c}$. Since $f \geq -1/8$ (will be proven below) there is an asymmetry on the expected value of f, i.e., the expected values of the entries of $\boldsymbol{G}$ are biased towards a positive value and should expect a bias towards a positive leading eigenvector of $\boldsymbol{G}$. Below, the derivation of the expectation on the

entries of $\boldsymbol{G}$ (assuming independence and uniformity) is described, and it is proved the main theorem showing that a random matrix whose entries are sampled i.i.d. form some distribution with positive mean and bounded variance has a positive leading eigenvector with probability 1 when the number of features p is sufficiently large.

Proposition B.1.1. *The minimal value of* $f = (\boldsymbol{a}^T\boldsymbol{b})(\boldsymbol{a}^T\boldsymbol{c})(\boldsymbol{b}^T\boldsymbol{c})$ *where* $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^n$ *are defined over the unit hypersphere is* $-1/8$.

Proof B.1.1. *The QR decomposition of 3 points on the unit hypersphere takes the from:*

$$[\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}] = [\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3] \begin{bmatrix} 1 & \cos(\beta) & \cos(\gamma_1) \\ 0 & \sin(\beta) & \sin(\gamma_1)\cos(\gamma_2) \\ 0 & 0 & \sin(\gamma_1)\sin(\gamma_2) \end{bmatrix} \tag{B.3}$$

where $[\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3] \in \mathbb{R}^p$ *are three orthogonal vectors.*

The problem, therefore, becomes the problem of minimizing

$$f = \cos(\beta)\cos(\gamma_1)\left\{\cos(\beta)\cos(\gamma_1) + \sin(\beta)\sin(\gamma_1)\cos(\gamma_2)\right\} \tag{B.4}$$

with respect to $\beta, \gamma_1, \gamma_2$. *Since* γ_2 *appears only in the* $\cos(\gamma_2)$ *expression, it can take only the values of* 1 *or* -1 *at the minimum energy point. By symmetry it is possible to assume it to be* -1, *and the problem reduces to the problem of minimizing* $(1/2)\cos(\beta+\gamma_1)\left\{\cos(\beta+\gamma_1)+\cos(\beta-\gamma_1)\right\}$. *The minimum occurs when* $\cos(\beta-\gamma_1)$ *is either* 1 *or* -1. *Both problems* $(1/2)\cos(u)(\cos(u)-1)$ *and* $(1/2)\cos(u)(\cos(u)+1)$ *have a minimum of* $-1/8$. □

Proposition B.1.2. *The expected value of* $f = (\boldsymbol{a}^T\boldsymbol{b})(\boldsymbol{a}^T\boldsymbol{c})(\boldsymbol{b}^T\boldsymbol{c})$ *where* $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^n$ *and* $\boldsymbol{c}$ *is uniformly sampled over the unit hypersphere, is* $(1/n)(\boldsymbol{a}^T\boldsymbol{b})^2$.

Proof B.1.2. *This expectation is given by the following integral*

$$\int (\boldsymbol{a}^T\boldsymbol{b})(\boldsymbol{a}^T\boldsymbol{c})(\boldsymbol{c}^T\boldsymbol{b})d\sigma(\boldsymbol{c}) = (\boldsymbol{a}^T\boldsymbol{b})\boldsymbol{a}^T(\int \boldsymbol{c}\boldsymbol{c}^T d\sigma(\boldsymbol{c}))\boldsymbol{b}. \tag{B.5}$$

The vector $\boldsymbol{c}$ *is taken from a uniform probability and in particular from a symmetric probability, i.e., where the probability of* $\boldsymbol{c}$ *and of remains the same under sign flipping of any subset of its entries (e.g.,* $p(\sqrt{2}[0.5, 0.5, 0, 0]) = p(\sqrt{2}[-0.5, 0.5, 0, 0])$*). Therefore,* $\int \boldsymbol{c}\boldsymbol{c}^T d\sigma(\boldsymbol{c})$ *is a multiplication of the identity matrix. From linearity of the trace and from the equality* $tr(\boldsymbol{c}\boldsymbol{c}^T) = \boldsymbol{c}^T\boldsymbol{c}$ *the trace of this matrix is* $\int \boldsymbol{c}^T\boldsymbol{c} d\sigma(\boldsymbol{c}) = 1$. *The*

matrix $\int \boldsymbol{c}\boldsymbol{c}^\mathsf{T} d\sigma(\boldsymbol{c})$*, therefore, is* $1/n$ *times the identity matrix in* $\mathbb{R}^n$*. The expectation* $\int (\boldsymbol{a}^\mathsf{T}\boldsymbol{b})(\boldsymbol{a}^\mathsf{T}\boldsymbol{c})(\boldsymbol{c}^\mathsf{T}\boldsymbol{b}) d\sigma(\boldsymbol{c})$ *then equals to* $(1/n)(\boldsymbol{a}^\mathsf{T}\boldsymbol{b})^2$*.* □

To get a rough estimate on the values in the matrix $\boldsymbol{G}$, it is further assumed that $\boldsymbol{a}$ and $\boldsymbol{b}$ are also evenly distributed on the n-dimensional sphere. In this case the expectation of $(\boldsymbol{a}^\mathsf{T}\boldsymbol{b})^2$ is $1/n$. To see this, it can be observed that the expectation $\mathbf{E}\left\{(\boldsymbol{a}^\mathsf{T}\boldsymbol{b})^2\right\} = \int\int (\boldsymbol{a}^\mathsf{T}\boldsymbol{b})^2 d\sigma(\boldsymbol{a}) d\sigma(\boldsymbol{b}) = \int \boldsymbol{a}^\mathsf{T}(\int \boldsymbol{b}\boldsymbol{b}^\mathsf{T} d\sigma(\boldsymbol{b}))\boldsymbol{a} d\sigma(\boldsymbol{a}) = \int \boldsymbol{a}^\mathsf{T}((1/n)\boldsymbol{I})\boldsymbol{a} d\sigma(\boldsymbol{a})$ where $\boldsymbol{I}$ is the identity matrix in $\mathbb{R}^n$.

Each entry g_{ij} is a sum of q such terms, $g_{ij} = \sum_{l=1}^{q} (\boldsymbol{m}_i^\mathsf{T}\boldsymbol{q}_l)(\boldsymbol{m}_j^\mathsf{T}\boldsymbol{q}_l)(\boldsymbol{m}_i^\mathsf{T}\boldsymbol{m}_j)$. If the features are irrelevant, it can be expected the correlation with the vector $\boldsymbol{q}_1$ to be similar to correlation with a "typical" random vector. In this case the above proposition applies. However, when $q > 1$ there are interrelations between the elements in the sum resulting from the orthogonality of the columns of $\boldsymbol{Q}$. The following proposition shows that the expectation is still larger than zero.

Proposition B.1.3. *The expected value of* $f = \sum_{i=1}^{q} (\boldsymbol{a}^\mathsf{T}\boldsymbol{b})(\boldsymbol{a}^\mathsf{T}\boldsymbol{c}_i)(\boldsymbol{b}^\mathsf{T}\boldsymbol{c}_i)$ *where* $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ *and* $\{\boldsymbol{c}_i\}$ *are orthonormal vectors uniformly sampled over the unit hypersphere in* $\mathbb{R}^n$ *is* $(q/n)(\boldsymbol{a}^\mathsf{T}\boldsymbol{b})^2$*.*

Proof B.1.3. *This expectation is given by the following integral*

$$(\boldsymbol{a}^\mathsf{T}\boldsymbol{b})\boldsymbol{a}^\mathsf{T}(\sum_{i=1}^{q} \int \boldsymbol{c}_i\boldsymbol{c}_i^\mathsf{T} d\sigma(\boldsymbol{c}_i \mid \boldsymbol{c}_1 \dots \boldsymbol{c}_{i-1}))\boldsymbol{b}, \tag{B.6}$$

where the main difference from the proof of Prop. B.1.2 is that now the probability distribution of $\boldsymbol{c}_i$ *is dependent on all the previous* $\boldsymbol{c}_1 \dots \boldsymbol{c}_{i-1}$*. Nevertheless, if* $\boldsymbol{c}_i$ *are uniformly sampled subject to the orthogonality constraint, the sum of integrals* $\boldsymbol{J} = \sum_{i=1}^{q} \int \boldsymbol{c}_i\boldsymbol{c}_i^\mathsf{T} d\sigma(\boldsymbol{c}_i \mid \boldsymbol{c}_1 \dots \boldsymbol{c}_{i-1})$ *is a product over the identity matrix in* $\mathbb{R}^n$*. To see this, consider products of the form* $\boldsymbol{v}^\mathsf{T}\boldsymbol{J}\boldsymbol{v}$*. From symmetry this product must be the same for every* $\boldsymbol{v} \in \mathbb{R}^n$*. i.e, since* $\boldsymbol{v}^\mathsf{T}\boldsymbol{J}\boldsymbol{v}$ *depends only on dot products (the distribution* $d\sigma(\boldsymbol{c}_i \mid \boldsymbol{c}_1 \dots \boldsymbol{c}_{i-1})$ *is a uniform distribution subject to constraints on dot products), it is invariant to an unitary transformation; in particular since any vector can be rotated to any other vector, it can be seen that it is not dependent on* $\boldsymbol{v}$*. It is known that* $tr(\boldsymbol{J}) = q$ *satisfying the proposition, as* $tr(\sum_{i=1}^{q} \int \boldsymbol{c}_i\boldsymbol{c}_i^\mathsf{T} d\sigma(\boldsymbol{c}_i \mid \boldsymbol{c}_1 \dots \boldsymbol{c}_{i-1})) = \sum_{i=1}^{q} \int \boldsymbol{c}_i^\mathsf{T}\boldsymbol{c}_i d\sigma(\boldsymbol{c}_i \mid \boldsymbol{c}_1 \dots \boldsymbol{c}_{i-1}) = k$*.* □

The body of results on spectral properties of random matrices deals with de distribution of eigenvalues [161]. For example, the corner-stone theorem known as Wigner's

semicircle theorem [162] is about the asymptotic distribution of eigenvalues with the following result: Given a symmetric $p \times p$ matrix whose entries are bounded independent random variables with mean μ and variance σ^2, then for any constant $c > 2\sigma$, with probability $1 - o(1)$ all eigenvalues except for at most $o(p)$ belong to $\Theta(c\sqrt{p})$, i.e., lie in the interval $I = (-c\sqrt{p}, c\sqrt{p})$.

It is also known that when $\mu = 0$ all the eigenvalues belong to the interval I (with probability $1 - o(1)$), while for the case $\mu > 0$ only the leading eigenvalue λ_1 is outside of I and

$$\lambda_1 = \frac{1}{p}\sum_{i,j} g_{ij} + \frac{\sigma^2}{\mu} + o(\frac{1}{\sqrt{p}}),$$

i.e., λ_1 asymptotically has a normal distribution with mean $\mu p + \sigma^2/\mu$ [163].

Then the procedure is to derive the asymptotic behavior of the leading eigenvector when $\mu > 0$ under the assumption that the entries of $\boldsymbol{G}$ are i.i.d. random variables. First, it is proved the theorem below, which deals with Gaussian random variables, and then extend it to bounded random variables:

Theorem B.1.1. Probabilistic Perron-Frobenius *Let $G = \{g_{ij}\}$ be a real symmetric $p \times p$ matrix whose entries for $i \geq j$ are independent identically and normally distributed random variables with mean $\mu > 0$ and variance σ^2. Then, for any $\varepsilon > 0$ there exist p_0 such that for all $p > p_0$ the leading eigenvector $\boldsymbol{v}$ of $\boldsymbol{G}$ is positive with probability of at least $1 - \varepsilon$.*

The scheme of the proof of the theorem is as follows. Let $\boldsymbol{G} = \mu\boldsymbol{J} + \sigma\boldsymbol{S}$ where $\boldsymbol{J} = \mathbf{1}\mathbf{1}^{\mathsf{T}}$ and s_{ij} are i.i.d. sampled according to $N(0, 1)$. Let $\boldsymbol{e} = (1/\sqrt{p})\mathbf{1}$, and let $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ be the spectrum of $\boldsymbol{G}$. From the semicircle law [162] and from [163] it is known that $\lambda_i = \Theta(\sqrt{p})$ for $i = 2, 3, \ldots, p$. The following auxiliary claims would be useful for proving the main theorem.

Lemma B.1.1. Bounds on Leading Eigenvalue. *Under the conditions of Theorem B.1.1 above, with probability $1 - o(1)$ the leading eigenvalue λ of $\boldsymbol{G}$ falls into the following interval:*

$$\mu p - \Theta(1) \leq \lambda \leq \mu p + \Theta(\sqrt{p}).$$

Proof B.1.4. *From the definition of the leading eigenvalue can be written:*

$$\begin{aligned} \lambda &= \max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{G}\boldsymbol{x} = \mu(\textstyle\sum_i x_i)^2 + \sigma \max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{S}\boldsymbol{x} \\ &\leq \mu p + \Theta(\sqrt{p}) \end{aligned}$$

where from the semicircle law, $\max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{S}\boldsymbol{x} = \Theta(\sqrt{p})$ *and from Cauchy-Schwartz inequality* $(\sum_i x_i)^2 = p(\sum_i x_i^2) = p$. *The lower bound follows from:*

$$\begin{aligned}\lambda &\geq \boldsymbol{e}^T\boldsymbol{G}\boldsymbol{e} = \mu p + \sigma \boldsymbol{e}^T\boldsymbol{S}\boldsymbol{e} \\ &= \mu p + \textstyle\sum_{i,j} s_{ij}/p \geq \mu p \Theta(1)\end{aligned}$$

Lemma B.1.2. *Under the conditions of Theorem B.1.1 above, with probability* $1-o(1)$, *the next bound is obtained as:*

$$\sum_i \nu_i \geq \sqrt{p} - c \tag{B.7}$$

for some constant c *where* ν_i *are the entries of the leading eigenvector* $\boldsymbol{v}$ *of* $\boldsymbol{G}$.

Proof B.1.5. *Let* $\boldsymbol{e} = a_1\boldsymbol{v} + \sum_{i=2}^p a_i\boldsymbol{v}_i$. *Since the eigenvectors and* $\boldsymbol{e}$ *are of unit norm, it can be inferred* $a_1^2 + \sum_{i=2}^p a_i^2 = 1$ *and without lost of generality it can assumed to be* $a_1 > 0$. *Therefore, it is possible to write* $\boldsymbol{e}^T\boldsymbol{G}\boldsymbol{e} = a_1^2\lambda + \sum_i \lambda_i a_i^2$. *Since* $\lambda_i = \Theta(\sqrt{p})$ *for* $i = 2, \ldots, p$ *and* $a_1^2 + \sum_i a_i^2 = 1$, *it follows the expression:*

$$\boldsymbol{e}^T\boldsymbol{G}\boldsymbol{e} = a_1^2\lambda + \Theta(\sqrt{p}).$$

Using the bound derived above of $\boldsymbol{e}^T\boldsymbol{G}\boldsymbol{e}$ *and Lemma B.1.1, it is obtained:*

$$\begin{aligned}\mu p - o(1) &\leq \lambda a_1^2 + \Theta(\sqrt{p}) \\ \frac{\mu p - \Theta(\sqrt{p})}{\mu p + \Theta(\sqrt{p})} &\leq a_1^2 \leq a_1\end{aligned}$$

from which can be conclude (with further manipulation):

$$1 - \frac{2\Theta(\sqrt{p})}{\mu p} = 1 - \frac{1}{\mu\Theta(\sqrt{p})} \leq a_1.$$

Consider now that a_1 *is the angle between* $\boldsymbol{e}$ *and* $\boldsymbol{v}$:

$$\frac{1}{\sqrt{p}}\sum_i \nu_i = \boldsymbol{e}^T\boldsymbol{v} = a_1 \geq 1 - \frac{1}{\mu\Theta(\sqrt{p})},$$

from which it is obtained:

$$\sum_i \nu_i \geq \sqrt{p} - c,$$

for some constant c. □

Then, the result is

$$\begin{aligned} \lambda\nu_i &= (\boldsymbol{G}\boldsymbol{v})_i = \mu\textstyle\sum_i \nu_i + \sigma(\boldsymbol{S}\boldsymbol{v})_i \\ &\geq \mu\sqrt{p} - C + \sigma\boldsymbol{g}^{\mathsf{T}}\boldsymbol{v} \end{aligned}$$

where $C = \mu c$ is a constant, $\boldsymbol{g}$ is some p-dimensional normally distributed i.i.d random vector. Demonstration would be done if it could be shown that the probability of the event $\boldsymbol{g}^{\mathsf{T}}\boldsymbol{v} > (1/\sigma)\mu\sqrt{p}$ occurs with probability $o(1)$, i.e., decays with the growth of p. The problem is that since $\boldsymbol{g}$ stands for a row of $\boldsymbol{S}$ and because $\boldsymbol{v}$ depends on $\boldsymbol{S}$ it is not possible to make the assumption that $\boldsymbol{g}$ and $\boldsymbol{v}$ are independent. Thus a straightforward tail bound would not be appropriate. The remainder of the proof below was contributed by [3] where care is taken to decouple the dependency between $\boldsymbol{g}$ and $\boldsymbol{v}$.

Proof B.1.6. Proof of Probabilistic Perron-Frobenius Theorem B.1.1. *Let $D(c)$ be the set of vectors in $\mathbb{R}^p$ satisfying Lemma B.1.2:*

$$D(c) = \left\{ \boldsymbol{v} \in \mathbb{R}^p : \|\boldsymbol{v}\| = 1, \sum_i \nu_i \geq \sqrt{p} - c \right\},$$

and let $\boldsymbol{g} \in \mathbb{R}^p$ be a vector of i.i.d. standard Normal distribution $N(0,1)$. It is interest to analyze the probability of the event

$$F(g) = \left\{ \exists \boldsymbol{v} \in D(c) \quad s.t. \quad \boldsymbol{g}^{\mathsf{T}}\boldsymbol{v} \geq \frac{\mu}{\sigma}\sqrt{p} \right\}, \quad \boldsymbol{g} \in \mathbb{R}^p \text{ in the case where } g_i \sim N(0,1).$$

In particular, it is necessary to show that the probability $P_{g_i \sim N(0,1)}(F(g))$ belongs to $o(1)$, i.e., decays with the growth of p.

Let $\boldsymbol{v} = \boldsymbol{e} + \boldsymbol{f}$ where $\boldsymbol{e} = \frac{1}{\sqrt{p}}\boldsymbol{1}$ was defined above and $\boldsymbol{f}$ is the residual. From the constraint $\|\boldsymbol{v}\|^2 = 1$, it is obtained a constraint on f:

$$\frac{2}{\sqrt{p}} \sum_i f_i + \sum_i f_i^2 = 0 \tag{B.8}$$

Given that $\boldsymbol{v} \in D(c)$, the following is yielded:

$$\sum_i \nu_i = \sqrt{p}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{e} = \sqrt{p} + \sum_i f_i \geq \sqrt{p} - c,$$

from which it is obtained another constraint on f:

$$-\sum_i f_i \leq c \tag{B.9}$$

Combining both constraints (B.8) and (B.9), it can be concluded the follow condition:

$$\|\boldsymbol{f}\|^2 = \frac{2c}{\sqrt{p}} \tag{B.10}$$

The expression $\boldsymbol{g}^T\boldsymbol{v}$ can be broken down to a sum of two terms $\boldsymbol{g}^T\boldsymbol{e}$ and $\boldsymbol{g}^T\boldsymbol{f}$. The first of these two terms is $o(1)$ by the law of large numbers, and so:

$$\begin{aligned} \boldsymbol{g}^T\boldsymbol{v} &= \boldsymbol{g}^T\boldsymbol{e} + \boldsymbol{g}^T\boldsymbol{f} \leq o(1) + \|\boldsymbol{g}\|\|\boldsymbol{f}\| \\ &\leq o(1) + \|\boldsymbol{g}\| \left(\frac{\sqrt{2c}}{p^{1/4}}\right) \end{aligned}$$

Vector $\boldsymbol{g}$ distributes according to the χ distribution with p degrees of freedom, which concentrates around $\sqrt{p}$. Therefore, with probability $1-o(1)$, $\|\boldsymbol{g}\| = \Theta(\sqrt{p})$. The probability that $\boldsymbol{g}^T\boldsymbol{v} \geq \Theta(\sqrt{p})$ is proportional to the probability that $\|\boldsymbol{g}\| \geq p^{3/4}$, which by the Gaussian tail bound decays exponentially with the growth of p. Since the probability that each entry of ν_i is negative decays exponentially, i.e., $p(\nu_i < 0) < e^{-Cp}$, for some constant C, then by the union-bound the union of such events $p(\nu_1 < 0 \cup \ldots \cup \nu_p < 0)$ is bounded from above by pe^{-Cp} which decays exponentially with the growth of p. □

Theorem B.1.1 used independent Gaussian random variables as a model to the matrix $\boldsymbol{G}$. This might seem a bit artificial, since the variables of the matrix $\boldsymbol{G}$ are dependent and bounded. While the independence assumption between all the elements in the upper triangular part of $\boldsymbol{G}$ is hard to remove, the use of Gaussian variables is not essential; as stated above the semi circle law holds for matrices with elements that are not necessarily Gaussian. The only place where is actually used the "Gaussianity" property was in the assumed structure of the variable $\boldsymbol{g}$. Since $\boldsymbol{g}$ contains normal i.i.d distributions, it was deduced that $\|\boldsymbol{g}\| = \Theta(\sqrt{p})$ and that the probability of $\|\boldsymbol{g}\| \geq p^{3/4}$ decays exponentially. Instead of Gaussianity, it can be used the property that the elements of the matrix $\boldsymbol{G}$ are bounded, and instead of Gaussian tail bounds, Hoeffding's tail inequality can be employed [164]. Henceforth, the one sided inequality will be used[1].

[1]This is the inequality one gets while proving Hoeffding's inequality. It differs from the canonical inequality in that the one sided case has a factor of 2 improvement.

Lemma B.1.3. (Hoeffding's one-sided tail inequality). *Let $X_1, \ldots X_n$ be bounded independent random variables such that $X_i \in [a_i, b_i]$. Then for $S_n = X_1 + \ldots + X_n$ the following inequality holds*

$$P(S_n - \mathbf{E}\{S_n\} \geq t) \leq exp\left(-\frac{2t^2}{\sum_{i=1}^{p}(b_i - a_i)^2}\right)$$

Using Hoeffding's lemma, the following lemma could be used to bound the norm of $\boldsymbol{g}$.

Lemma B.1.4. *Let $\boldsymbol{g}$ be a random p-vector of i.i.d bounded variables, i.e., for each $i = 1 \ldots p$, $|\boldsymbol{g}_i| < M$. The following holds for some constant C:*

$$P(\|\boldsymbol{g}\|^2 \geq Dp^{1/2+\varepsilon}) \leq exp\left(-\frac{C^2D^2p^{2\varepsilon}}{M^2}\right)$$

Proof B.1.7. *Hoeffding's inequality will be applied to the random variable $\frac{1}{n}\|\boldsymbol{g}\|^2$, which has a mean μ that does not depend on p.*

Assume $\gamma \geq Dp^{-1/2+\varepsilon}$, where $\varepsilon > 1/2$. For some $p > \hat{p}$, and for some c, $\gamma - \mu \geq c\gamma$, the following is obtained:

$$P(\frac{\|\boldsymbol{g}\|^2}{p} \geq \gamma) = P(\frac{\|\boldsymbol{g}\|^2}{p} - \mu \geq \gamma - \mu) \leq P(\frac{\|\boldsymbol{g}\|^2}{p} \geq c\gamma).$$

Now, Hoeffding's one sided inequality is applied again:

$$P(\frac{\|\boldsymbol{g}\|^2}{p} - \mu \geq c\gamma) \leq exp\left(-\frac{c^2p\gamma^2}{M^2}\right) \leq exp\left(-\frac{c^2D^2p^{2\varepsilon}}{M^2}\right).$$

□

By letting $D = 1$ and $\varepsilon = 1$, one gets that the probability $P(\|\boldsymbol{g}\| \geq p^{3/4}) = P(\|\boldsymbol{g}\|^2 \geq p^{3/2}) = P(\frac{1}{p}\|\boldsymbol{g}\|^2 \geq p^{1/2})$ is smaller than $e^{-\frac{c^2p^2}{M^2}}$. This is similar to the Gaussian case, and is sufficient to prove Theorem B.1.1 in the case in which bounded variables are used instead of Gaussian variables. To summarize the positivity issue, the weight vector $\boldsymbol{\alpha}$ comes out positive due to the fact that it is the leading eigenvector of a matrix whose entries have a positive mean (Propositions B.1.1 and B.1.2). Theorem B.1.1 made the connection between matrices which have the property of a positive mean and the positivity of its leading eigenvector in a probabilistic setting.

B.2 Sparsity

Now, the issue of the sparsity of the weight vector $\boldsymbol{\alpha}$ will be studied. It has been observed in the past that the key for sparsity lies in the positive combination of terms [165] — therefore there is a strong, albeit anecdotal, relationship between the positivity of $\boldsymbol{\alpha}$ and the sparsity feature. Below, it will be established a relationship between the spectral properties of the relevant and the irrelevant feature sets, and the sparsity of the feature vector.

Let $\boldsymbol{M}$ be the (normalized) data matrix consisting of p rows. Assume that the rows of the matrix have been sorted such that the first p_1 rows are relevant features, and the next $p_2 = p - p_1$ features are irrelevant. Let the matrix containing the first p_1 rows be noted as $\boldsymbol{M}_1$, and let the matrix containing the rest of the rows be $\boldsymbol{M}_2$, i.e, $\boldsymbol{M} = [\frac{\boldsymbol{M}_1}{\boldsymbol{M}_2}]$.

The elements of the vector $\boldsymbol{\alpha}$ that correspond to the irrelevant features were studied to show that these elements have a small magnitude. If these p_2 weights are low, it can expected the effect of the associated features to be small. It will next be tied the average of these values to the spectral properties of $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$.

Recall the weight vector $\boldsymbol{\alpha}$ is the first eigenvector of the matrix $g_{ij} = (\boldsymbol{m}_i^\mathsf{T}\boldsymbol{m}_j)\boldsymbol{m}_i^\mathsf{T}\boldsymbol{Q}\boldsymbol{Q}^\mathsf{T}\boldsymbol{m}_j$, where $\boldsymbol{m}_i$ are the rows of the matrix $\boldsymbol{M}$, and $\boldsymbol{Q}$ is a matrix containing q orthonormal columns $\boldsymbol{q}_i$, $i = 1 \dots q$. Let λ be the largest eigenvalue of $\boldsymbol{G}$.

Lemma B.2.1. (Sum of irrelevant features' weight.) *Using the above definitions, let γ_i, $i = 1 \dots p_2$ be the eigenvalues of $\boldsymbol{M}_2\boldsymbol{M}_2^\mathsf{T}$.*

$$\sum_{i=p_1+1}^{p} \alpha_i \leq \sqrt{\frac{\sum_{i=1}^{p_2} \gamma_i^2}{\lambda}}$$

Proof B.2.1. *Note that if $\sum_{i=p_1+1}^{p} \alpha_i \leq 0$ the lemma holds trivially. Let $[\frac{0}{1}]$ be the vector with p_1 zeros and p_2 ones.*

$$\sum_{i=p_1+1}^{p} \alpha_i = \begin{bmatrix} 0 \\ \hline 1 \end{bmatrix}^\mathsf{T} \boldsymbol{\alpha} = \sqrt{\begin{bmatrix} 0 \\ \hline 1 \end{bmatrix}^\mathsf{T} \boldsymbol{\alpha}\boldsymbol{\alpha}^\mathsf{T} \begin{bmatrix} 0 \\ \hline 1 \end{bmatrix}} \leq \sqrt{\frac{\begin{bmatrix} 0 \\ \hline 1 \end{bmatrix}^\mathsf{T} \boldsymbol{G} \begin{bmatrix} 0 \\ \hline 1 \end{bmatrix}}{\lambda}}$$

where the last inequality follows from the spectral decomposition of the positive definite matrix $\boldsymbol{G}$, to which $\boldsymbol{\alpha}$ is an eigenvector with an eigenvalue of λ.

Let $\widehat{\boldsymbol{G}}$ be the matrix containing the point-wise squares of the elements of $\boldsymbol{M}\boldsymbol{M}^\mathsf{T}$, i.e., $\widehat{g}_{ij} = (\boldsymbol{m}_i^\mathsf{T}\boldsymbol{m}_j)^2$.

Let $\widehat{\boldsymbol{Q}}$ be a matrix containing $p-q$ orthonormal columns that span the space orthogonal to $\boldsymbol{Q}$.

The matrix $(\widehat{\boldsymbol{G}}-\boldsymbol{G})$ has a structure similar to $\boldsymbol{G}$, but with $\widehat{\boldsymbol{Q}}$ instead of $\boldsymbol{Q}$, and is also positive definite. To see this notice that $\boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}}+\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{Q}}^{\mathsf{T}}=\boldsymbol{I}_p$ and that the ij element of $(\widehat{\boldsymbol{G}}-\boldsymbol{G})$ is therefore given by

$$\widehat{g}_{ij}-g_{ij}=(\boldsymbol{m}_i^{\mathsf{T}}\boldsymbol{m}_j)^2-(\boldsymbol{m}_i^{\mathsf{T}}\boldsymbol{m}_j)\boldsymbol{m}_i^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{m}_j=(\boldsymbol{m}_i^{\mathsf{T}}\boldsymbol{m}_j)\boldsymbol{m}_i^{\mathsf{T}}\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{Q}}^{\mathsf{T}}\boldsymbol{m}_j$$

Then,

$$\begin{bmatrix}0\\1\end{bmatrix}^{\mathsf{T}}\boldsymbol{G}\begin{bmatrix}0\\1\end{bmatrix}=\begin{bmatrix}0\\1\end{bmatrix}^{\mathsf{T}}\widehat{\boldsymbol{G}}\begin{bmatrix}0\\1\end{bmatrix}-\begin{bmatrix}0\\1\end{bmatrix}^{\mathsf{T}}(\boldsymbol{G}-\widehat{\boldsymbol{G}})\begin{bmatrix}0\\1\end{bmatrix}\leq\begin{bmatrix}0\\1\end{bmatrix}^{\mathsf{T}}\widehat{\boldsymbol{G}}\begin{bmatrix}0\\1\end{bmatrix}=\|\boldsymbol{M}_2\boldsymbol{M}_2^{\mathsf{T}}\|_F^2=\sum_{i=1}^{p_2}{\gamma_i}^2.$$

The lemma follows. □

The denominator in the bound ($\sqrt{\lambda}$) is exactly the quantity that the algorithms maximize. The higher this value, the tighter the bound. In the ideal case, almost all of the energy in the features is contained in the space spanned by the columns of $\boldsymbol{Q}$. Let $[\frac{1}{0}]$ be the vector of p_1 ones, followed by p_2 zeros. It is possible to write:

$$\lambda=\boldsymbol{\alpha}\mathsf{T}\boldsymbol{G}\boldsymbol{\alpha}\geq\frac{[\frac{1}{0}]^{\mathsf{T}}\boldsymbol{G}[\frac{1}{0}]}{p_1}\sim\frac{[\frac{1}{0}]^{\mathsf{T}}\widehat{\boldsymbol{G}}[\frac{1}{0}]}{p_1}=\frac{\|\boldsymbol{M}_1\boldsymbol{M}_1^{\mathsf{T}}\|_F^2}{p_1}$$

The bound will be tightest if all relevant features have high correlations. In this case, it can be expected $\sqrt{\lambda}$ to be linear in p_1. Therefore the addition of more relevant features reduces the weights of the irrelevant features.

Without any assumption about the entries of the data matrix $\boldsymbol{M}$, it is not possible to say much on the numerator of the bound in Lemma B.2.1. However, by using random matrices, it is possible to qualitatively evaluate this bound.

The numerator of the bound contains the term $\sum_{i=1}^{p_2}\gamma_i^2$, which is just the squared Frobenius norm of $\boldsymbol{M}_2\boldsymbol{M}_2^{\mathsf{T}}$. Let $\boldsymbol{W}_2=\boldsymbol{M}_2\boldsymbol{M}_2^{\mathsf{T}}$ and $\|\boldsymbol{W}_2\|_F^2=\operatorname{tr}(\boldsymbol{W}_2\boldsymbol{W}_2^{\mathsf{T}})=\operatorname{tr}(\boldsymbol{W}_2^2)$. The expectation of this expression (where $\boldsymbol{M}_2$ is drawn from a random distribution), normalized by the number of rows in $\boldsymbol{M}_2$, is called the *second moment* of $\boldsymbol{W}_2$. More generally, if $\boldsymbol{A}$ is a random matrix of size $p\times p$, the k moment of it is defined as $m_k=\frac{1}{p}\operatorname{tr}(\boldsymbol{A}^k)$. For large p this definition coincides with the moments of the distribution of the eigenvalues of the matrix $\boldsymbol{A}$.

Consider now matrices $\boldsymbol{W}$ of the form $\boldsymbol{W}=\frac{1}{n}\boldsymbol{M}\boldsymbol{M}^{\mathsf{T}}$, where $\boldsymbol{M}$ is an $p\times n$

random matrix with zero mean (weakly) independent elements with a variance of 1. These matrices are called Wishart matrices. Note that the elements in the matrix $\boldsymbol{M}$ need not be Gaussian. The rows of the matrix $\boldsymbol{M}$ are not explicitly normalized. However, the scale of $\frac{1}{n}$ can be thought of as a scale of $\frac{1}{\sqrt{n}}$ for each element of $\boldsymbol{M}$, and due to the central limit theorem it can be expected for large enough values of n to have the mean of each row approximately zero and the norm of each row approximately 1. Hence, Wishart matrices well approximate the data matrices, if it is assumed that the elements of the data matrices are independent. For the bulk of irrelevant features, this may be a reasonable assumption.

For large p, the moments of the Wishart matrices are well approximated by the Narayana polynomials $m_k = \sum_{j=0}^{k-1} \frac{(p/j)^j}{j+1} \binom{k}{j} \binom{k-1}{j}$. In particular, the second moment is given by $1 + p/n$. Since the moment is the appropriate trace scaled by $\frac{1}{p}$, it can be expected $\sqrt{\sum_{i=1}^{p_2} \gamma_i^2}$ to behave similarly to $\sqrt{p_2(1 + p_2/n)}$.

Therefore, the rate in which the bound on the sum of squares of weights of irrelevant features grow is mostly linear. The implication is that the $Q - \alpha$ algorithm is robust to many irrelevant features: to a first approximation, the bound on the average squared weight of an irrelevant feature remains mostly the same, as the number of irrelevant features increases.

B.3 Sparsity and Generalization

The sparsity of the returned vector of weights does more than just directly ensure that the irrelevant features are left out; it also helps the generalization ability of the returned kernel by lowering the trace of the kernel matrix.

Recall that in this optimization scheme, the vector of weights $\boldsymbol{\alpha}$ has a norm of 1, and is expected to have all positive elements. For norm-1 vectors, the sum $\sum_i \alpha_i$ is highest when the elements of the vector are evenly distributed. Due to the sparsity of the returned vector of weights, it can be expected the above sum to be much lower than what can be obtained with a uniform weighting of the data.

Consider the matrix $\boldsymbol{A}_{\boldsymbol{\alpha}}$, the linear kernel matrix based on the weights returned by the $Q - \alpha$ algorithm. $\boldsymbol{A}_{\boldsymbol{\alpha}}$, which equals $\sum \alpha_i \boldsymbol{m}_i \boldsymbol{m}_i^{\mathsf{T}}$, is a weighted sum of rank-one matrices. Since the features are normalized to have norm-1, each such rank-one matrix $\boldsymbol{m}_i \boldsymbol{m}_i^{\mathsf{T}}$ has a trace of 1. Therefore, the trace of the kernel matrix $\boldsymbol{A}_{\boldsymbol{\alpha}}$ is exactly the sum of the elements of the vector $\boldsymbol{\alpha}$.

From the discussion above, the trace of the kernel matrix returned by the $Q - \alpha$

algorithm is expected to be low. This is exactly the criteria for a good kernel matrix expressed by "the trace bounds." The trace bounds are Rademacher complexity type of error bounds for classifiers that are linear combinations of kernel functions [166]. These bounds relate the trace of the kernel matrix used for training with the generalization error of the classifiers that were trained. The lower the trace, the lower the bound on the difference between the testing error and the training error.

Although there is no immediate analog to the concept of generalization error in the unsupervised case, it can be expected a similar criteria to hold for this case as well. A good kernel matrix for unsupervised learning should support the separation given by the set of true underlying labels (although unknown). It should not, however, support any random labeling. This is exactly what is measured by the Rademacher process: how well does the class of functions used for learning separate random partitions of the training set.

Appendix C

Dynamic Time Warping

Dynamic Time Warping (DTW) method was applied in some sections throughout this work, for this reason, this appendix gives a brief description of the procedure.

An important step in clustering is to select a distance measure, which determines the similarity calculated between two vector sequences. If an appropriate measure of the similarity of two vector sequences is available, clustering followed by prototype extraction is a suitable unsupervised learning method for this problem. Finding such measure of similarity is difficult because time series that are qualitatively the same but, may be, quantitatively different in at least two ways. First, they may be of different lengths, making difficult or impossible to embed the time series in a metric space in order to use, for example, Euclidean distance to determine similarity. Second, within a single time series, the rate at which progress is made can vary nonlinearly. The same pattern may evolve slowly at first and then speed up, or may begin quickly and then slow down. Such differences in rate make similarity measures, such as cross-correlation, unusable [167].

In the case of ECG signals analysis whose heart rate dynamics changes the R-R interval length between two consecutive heartbeats and because of the high variability of its main patterns, namely, the P wave, QRS complex, P-R segment or S-T segment, it is necessary to apply another class of similarity measure, specially when comparing waveforms of several number of heartbeats, which have to be grouped into k clusters, subsets, or heartbeats morphologies [168].

Dynamic Time Warping is a method commonly used to normalize the length between two vector sequences of different length, finding an optimal alignment function between them. The cumulative cost of the alignment path obtained can be used as a similarity o dissimilarity measure.

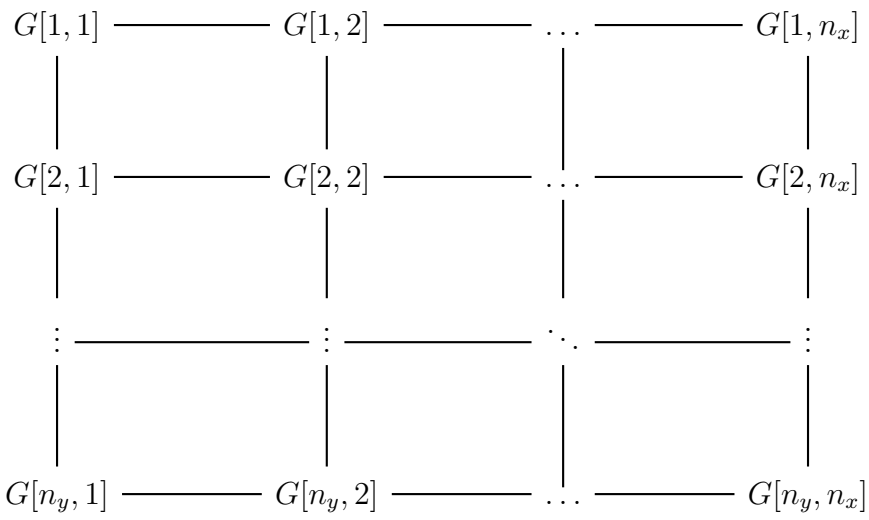


Figure C.1: Dynamic Programming Matrix $\boldsymbol{G}$, and associated multistage graph

It has been applied successfully on several contexts: speech recognition, economic, control process, ECG heartbeats clustering, among others [132], [10], [64].

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two sequences of length n_x and n_y, $(n_x \neq n_y)$ respectively, that are represented by,

$$\begin{aligned} \boldsymbol{x} &= (x_1 \ldots x_{n_x}) \\ \boldsymbol{y} &= (y_1 \ldots y_{n_y}), \end{aligned} \tag{C.1}$$

The alignment procedure allows us to compare each value x_i of the first sequence with a value y_j of the second sequence, in this way, it is obtained the whole set of possible comparisons represented as a matrix $\boldsymbol{G}$, or as a multistage graph, as shown in the Figure (C.1).

The matrix $\boldsymbol{G}$ is called Dynamic Programming Matrix, which has as many rows as the length of $\boldsymbol{x}$ and as many columns as the length of $\boldsymbol{y}$. The dissimilarity measure is obtained in the last element of the matrix G, i.e. $G[n_x, n_y]$. This value must be normalized with respect to the length of the alignment path in order to avoid the length effect.

The goal of algorithm is to find the node path with the minimum cost between two sequences, according to certain constraints. The node path is defined as the ordered set of nodes from the initial node (i_1, j_1) to the final node (i_{end}, j_{end}), in the next form, $[(i_1, j_1), (i_2, j_2) \ldots (i_{end}, j_{end})]$.

In order to establish the cost of the path, it is necessary to define the cost between each pair of nodes. The cost function associated with the $k-$th transition is denoted as

$d[(i_{k-1}, j_{k-1})|(i_k, j_k)]$. The cost must be a positive value and can be measured trough the norm L_1 or L_2.

The warping path has to satisfy several side conditions:

- *Monotonicity:* The scan of the matrix $\boldsymbol{G}$ is from left to right and from up to down. i.e. $i_k \leq i_{k+1}$ and $j_k \leq j_{k+1}$ respectively.
- *Continuity:* The transition between the previous and the actual nodes can not have intermediate nodes. i.e. $i_k - i_{k-1} \leq 1$ and $j_k - j_{k-1} \leq 1$.
- *Boundary:* With this constraint the first and the last element of the matrix $\boldsymbol{G}$ coincide with the first and the last element of both sequences. i.e. $i_1 = j_1 = 1$, $i_{end} = n_x$ and $j_{end} = n_y$.
- *Window:* $|i_k - j_k| = \Delta$, where Δ is a positive integer.
- *Slope:* Neither too steep nor too gentle a gradient should be allowed.

The final cost is equal to the cumulative sum of each one of the transitions:

$$D = \sum_{k=2}^{K} d[(i_{k-1}, j_{k-1})|(i_k, j_k)] \tag{C.2}$$

where K represents the number of nodes with respect to the path.

Additionally, a weighting function $w(a_k, b_k)$ can be applied regarding the transition from (i_k, j_k) to (i_{k-1}, j_{k-1}), in this way, the whole expression can be normalized in order to get a measure regardless the path length.

$$D = \frac{\sum_{k=2}^{K} d[(i_{k-1}, j_{k-1})|(i_k, j_k)] w(a_k, b_k)}{\sum_{k=2}^{K} w(a_k, b_k)} \tag{C.3}$$

Making $N = \sum_{k=2}^{K} w(a_k, b_k)$, it is obtained,

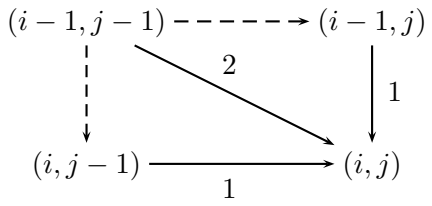
$$D = \frac{G[n_x, n_y]}{N} \tag{C.4}$$

Then, the recursive expression to each element of the matrix $\boldsymbol{G}$ is:

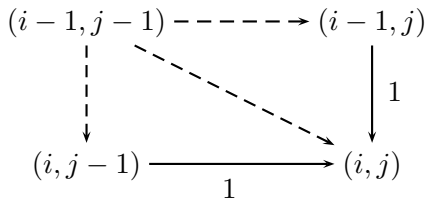
$$G[i,j] = \min_{a,b\ P} \sum_{k=2}^{K} d[(i_{k-1}, j_{k-1})|(i_k, j_k)]w(a_k, b_k) \tag{C.5}$$

where $G[i-a, j-b]$ represents the cumulative minimum cost associated with a set of productions P. The Figure C.2 shows a set of productions commonly used in dynamic programming, which are properly selected in order to satisfy properties of a metric-space structure [64].

(a) P1



(b) P2



(c) P3

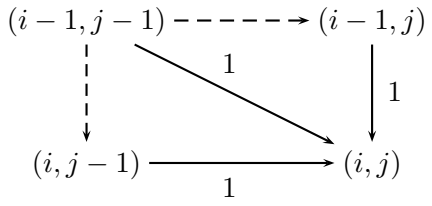


Figure C.2: Types of productions (The transition associated to dashed lines has weight 0)

By using the set of productions $P1$ and taking into account the respective values

for transition weights between consecutive nodes, matrix $\boldsymbol{G}$ can be obtained as:

$$G[i,j] = \min \begin{cases} G[i,j-1] + d[(i,j)|(i,j-1)] \\ G[i-1,j] + d[(i,j)|(i-1,j)] \\ G[i-1,j-1] + 2d[(i,j)|(i-1,j-1)] \end{cases} \tag{C.6}$$

where the first node is $G[1,1] = |x[1] - y[1]|$. In such way, the dissimilarity between the two sequences is:

$$D = \frac{G[n_x, n_y]}{n_x + n_y} \tag{C.7}$$

Global constraints can also be applied in order to keep the search path inside a region where the alignment is more meaningful from the point of view of the application (see Figure C.3). To avoid paths or path portions along the boundaries of the dynamic programming matrix, the search is reduced to a centered window, assigning a high cost to the nodes outside that region. Mathematically, $G[i,j] = \infty$, if $(i,j) \notin \mathfrak{R}$.

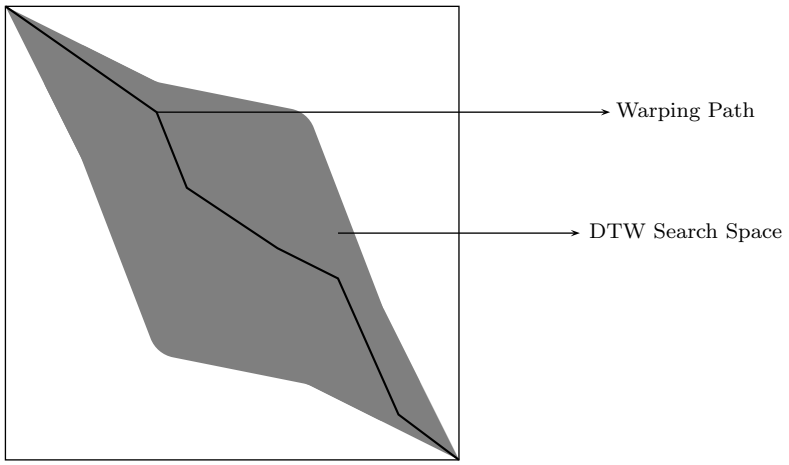


Figure C.3: Global constraint to optimize the calculation of the DTW

Appendix D

Spectral Clustering

In several sections throughout this work, the affinity matrix is discussed. In the section 6.3 the trivial affinity matrix is defined in order to establish the relationship of the data points each other. In the sections 7.6, 7.4 and 8.5.1, the estimation of number of groups was studied, which is performed by means of the properties of the affinity matrix.

This appendix describes some useful types of affinity matrix based on the properties of spectral clustering.

D.1 Introduction

In the area of pattern recognition and classification, clustering methods based on graphs and spectral analysis are relatively new but have been applied successfully on several science fields. Such as circuits design [169], computational load balance for parallel applications [170], image segmentation [171, 172], among others. These methods are discriminative, i.e., do not require prior information (assumptions about the nature of data) and the partitions are generated taking advantage of the information obtained from its own method heuristic. In this case, a global decision criterion that takes into consideration the estimated value of the probability of two elements are belonging to the same cluster [150]. Therefore, this kind of analysis can be easily studied from graph theory, where such probability corresponds to the affinity measure between a pair of nodes. Commonly, this criterion is applied on a new representation space whose dimension is less than that original data and then a dimensionality reduction procedure is accomplished, in that way the relation among elements are conserved as well as possible. Thus, eigenvectors and eigenvalues based analysis takes place.

Spectral analysis methods are of great interest for classification tasks, mainly, in case of classes are not linearly separable. In particular, spectral clustering uses the information given by eigenspace (i.e, space generated by eigenvectors) because of eigenvectors are directly associated with the quality of clustering. Typically, clustering methods, due to their discriminative nature, require some prior initial parameters. In case of spectral clustering, it is necessary to establish a prior number of groups and the affinity matrix. There are several alternatives to choose the affinity measure and the selection of one of them is not a trivial task, but represents an important issue in spectral analysis.

D.2 Clustering Method

A weighted graph can be represented as $\mathbb{G} = (\mathbb{V}, \mathbb{E}, \boldsymbol{W})$, where $\mathbb{V}$ is the set of nodes or vertices, $\mathbb{E}$ is the set of edges and $\boldsymbol{W}$ represents the relation among nodes, in other words, the affinity matrix. Given that w_{ij} represents the weight of the edge between i-th and j-th element, it must be a non-negative value. In addition, in a non-directed graph is evident that $w_{ij} = w_{ji}$. Therefore, matrix must be a symmetric and positive semidefinite matrix. Figure D.1 shows an example of weighted graph.

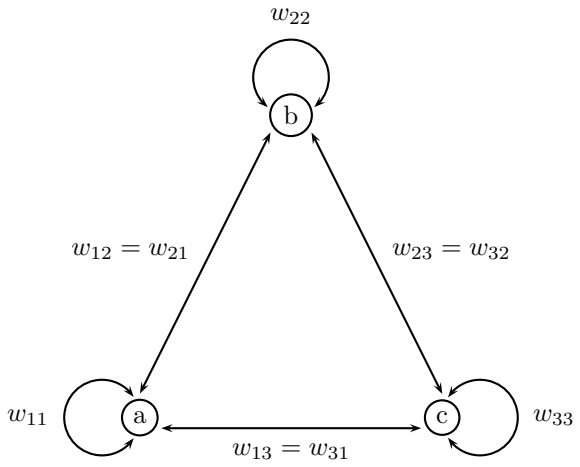


Figure D.1: Weighted graph with three nodes

Two measures are of interest: total weighted connections and degree.

By letting $\mathbb{G} = (\mathbb{V}, \mathbb{E}, \boldsymbol{W})$ be a weighted graph and $\mathbb{A}, \mathbb{B} \subset \mathbb{V}$, total weighted

connections between $\mathbb{A}$ and $\mathbb{B}$ can be computed as:

$$\text{links}(\mathbb{A}, \mathbb{B}) = \sum_{i \in \mathbb{A}, j \in \mathbb{B}} w_{ij} \tag{D.1}$$

The degree of a set is the total links to all the nodes:

$$\text{degree}(\mathbb{A}) = \text{links}(\mathbb{A}, \mathbb{V}) \tag{D.2}$$

Using the degree as a normalization term, it can be defined the normalized links as:

$$\text{linkratio}(\mathbb{A}, \mathbb{B}) = \frac{\text{links}(\mathbb{A}, \mathbb{B})}{\text{degree}(\mathbb{A})} \tag{D.3}$$

In spectral analysis, term $\mathbb{V} = \{1, \cdots, n\}$ represents the indices of data set to be grouped. The aim of spectral clustering is to decompose $\mathbb{V}$ into k disjoint subsets, then, $\mathbb{V} = \cup_{l=1}^{k} \mathbb{V}_l$ and $\mathbb{V}_l \cap \mathbb{V}_m = \emptyset$, $\forall l \neq m$, such decomposition is done, commonly, by using spectral information and orthonormal transformations.

Two special linkratios are considered for clustering: First, linkratio$(\mathbb{A}, \mathbb{A})$, which measures how many links stay within $\mathbb{A}$ itself [150]. The second one is linkratio$(\mathbb{A}, \mathbb{A} \backslash \mathbb{V})$, which is complement of the first one and measures how many links escape from $\mathbb{A}$. According to this, a good clustering can be achieved with both tight connections within partitions and loose connections between partitions. These two goals are captured in the k-way normalized associations (knassoc) and normalized cuts criteria (kncuts), as follows:

$$\text{knassoc}(\Gamma_{\mathbb{V}}^{k}) = \frac{1}{k} \sum_{l=1}^{k} \text{linkratio}(\mathbb{V}_l, \mathbb{V}_l) \tag{D.4}$$

$$\text{kncuts}(\Gamma_{\mathbb{V}}^{k}) = \frac{1}{k} \sum_{l=1}^{k} \text{linkratio}(\mathbb{V}_l, \mathbb{V}_l \backslash \mathbb{V}) \tag{D.5}$$

Because a normalization term was applied, it can be proved that,

$$\text{knassoc}(\Gamma_{\mathbb{V}}^{k}) + \text{kncuts}(\Gamma_{\mathbb{V}}^{k}) = 1 \tag{D.6}$$

Thus, maximizing the associations and minimizing the cuts are achieved simultaneously. Therefore, the objective function to be maximized in order to obtain a good

clustering is only:

$$\varepsilon(\Gamma_{\mathbb{V}}^k) = \text{knassoc}(\Gamma_{\mathbb{V}}^k) \tag{D.7}$$

This method is widely described in [150].

D.3 Affinity Matrix

The relation degree among nodes, also called affinity measure, can be represented as a symmetric and positive semidefinite matrix as is discussed above. Therefore, $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ being the data matrix, the affinity matrix can be chosen randomly as a matrix of the form $\boldsymbol{W}_r = \boldsymbol{X}_r\boldsymbol{X}_r^\mathsf{T}$, where $\boldsymbol{X}_r \in \mathbb{R}^{n\times d}$ is a random matrix.

Another form to select the affinity corresponds to the trivial case where affinity captures the inner products among rows of data matrix, i.e, $\boldsymbol{W}_t = \boldsymbol{X}\boldsymbol{X}^\mathsf{T}$. This form is the most frequently used.

By the other hand, since the affinity is a similarity measure (i. e., is opposite to distance or dissimilarity), intuitively, can be established alternatives as $w_{ij} = 1/\,\mathrm{d}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for $(i \neq j)$ with $w_{ii} = \text{const}$, where $\mathrm{d}(\cdot,\cdot)$ is a distance measure. This measure presents an inverse relation with respect to distance and this could not convenient, in all cases, because it can generate strong changes on affinity matrix $\boldsymbol{W}$. Then, soft versions have been proposed, such as the Exponential affinity matrix, which can be expressed as $w_{ij} = e^{-d^2(\boldsymbol{x}_i,\boldsymbol{x}_j)}$ with $w_{ii} = \text{const}$, where the value of w_{ii} is set according to certain constraints when the itself affinity does not be of interest, e.g. it can be chosen as $w_{ii} = 0$. In [156], it is presented an alternative to the previous affinity measure called scaled affinity matrix which can be written as $w_{ij} = e^{-d^2(\boldsymbol{x}_i,\boldsymbol{x}_j)/(\sigma_i\sigma_j)}$, with $w_{ii} = 0$, where $\sigma_i = \mathrm{d}(\boldsymbol{x}_i, \boldsymbol{x}_N)$ and $\boldsymbol{x}_N$ is the N-th nearest neighbor. The value of N is chosen regardless of the scale parameter and it is established according to data dimension [156].

In the Table D.1 some affinity measures are depicted.

Table D.1: Affinity measures

Affinity measure	**Expression**
Random	$\boldsymbol{W}_r = \boldsymbol{X}_r\boldsymbol{X}_r^\mathsf{T}$, where $\boldsymbol{X}_r \in \mathbb{R}^{n\times d}$ is a random matrix
Trivial	$\boldsymbol{W}_t = \boldsymbol{X}\boldsymbol{X}^\mathsf{T}$, where $\boldsymbol{X}$ corresponds to data matrix.
Exponential	$w_{ij} = e^{-d^2(\boldsymbol{x}_i,\boldsymbol{x}_j)}$ with $w_{ii} = \text{const}$
Scaled exponential	$w_{ij} = e^{-d^2(\boldsymbol{x}_i,\boldsymbol{x}_j)/(\sigma_i\sigma_j)}$, with $w_{ii} = 0$, where $\sigma_i = d(\boldsymbol{x}_i, \boldsymbol{x}_N)$ is the N-th nearest neighboor [156].

Part VI

Bibliography

Bibliography

[1] P. Laguna and L. Sörnmo. *Bioelectrical signal processing in cardiac and neurological applications*, volume 1. 2005.

[2] H. Liang, S. Lukkarinen, and I. Hartimo. Heart sound segmentation algorithm based on heart sound envelogram. In *Computers in Cardiology 1997*, pages 105 –108, 7-10 1997.

[3] Lior Wolf and Amnon Shashua. Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach. *Journal of machine learning*, 6:1855 – 1887, 2005.

[4] P. De-Chazal, M. O'Dwyer, and R. Reilly. Automatic classification of heartbeats using ecg morphology and heartbeat interval features. *IEEE transaction on biomedical engineering*, 51(7):1196–1206, 2004.

[5] C. Wen, M.F. Yeh, and K.C. Chang. Ecg beat classification using greyart network. *Signal Processing, IET*, 1(1):19–28, 2007.

[6] Y. Özbay, R. Ceylan, and B. Karlik. A fuzzy clustering neural network architecture for classification of ecg arrhythnmias. *Computers in Biology and Medicine*, 36:376–388, 2006.

[7] Rahime Ceylan and Yüksel Özbay. Comparison of fcm, pca and wt techniques for classification ecg arrhythmias using artificial neural network. *Expert Systems with Applications*, 33(2):286 – 295, 2007.

[8] M. Lagerholm, C. Peterson, G. Braccini, L. Edenbrandt, and L. Sörnmo. Clustering ecg complexes using hermite functions and self-organising maps. *IEEE trans. on. Biomed.*, 48:838–847, 2000.

[9] D. Cuesta, M. Biagetti, R. Quinteiro, P. Mico-Tormos, and M. Aboy. Unsupervised classification of ventricular extrasystoles using bounded clustering algorithms and morphology matching. *Medical and Biological Engineering and Computing*, 45(3):229–239, 2007.

[10] J.L. Rodríguez-Sotelo, D. Cuesta-Frau, and G. Castellanos-Domínguez. Unsupervised classification of atrial heartbeats using a prematurity index and wave morphology features. *Medical and Biological Engineering and Computing*, 47(7):731–741, 2009.

[11] Edilson Delgado-Trejos. *Generación y extracción/selección de características en la detección automática de isquemia y deficiencias valvulares sobre registros de la actividad cardíaca*. PhD thesis, Universidad Nacional de Colombia, Manizales, Colombia, 2008.

[12] J. J. Arango. *Manual de electrocardiografía*. Asociación Colombiana de Facultades de Medicina, 5 edition, 2003.

[13] A. C. Guyton and J. E. Hall. *Tratado de Fisiología médica*. McGraw-Hill, 9 edition, 1997.

[14] D. Cuesta J. Segura-Juárez, L. Samblas-Pena, and M. Aboy. A microcontroller-based portable electrocardiograph recorder. *IEEE transaction on biomedical engineering*, 51:1686–16906, 2004.

[15] G.B. Moody and R.G. Mark. Development and evaluation of a 2-lead ecg analysis program. *in Computers in Cardiology, IEEE Computer Society Press*, pages 39–44, 1982.

[16] E. Braunwald. *Tratado de Cardiología.*, volume 1. 4 edition, 1993.

[17] A. Guyton and J. E. Hall. *Tratado de fisiología médica*. 10 edition.

[18] J.M. Adriá Micó and P. Ibor. Medicina general. *Medicina General*, (66):426, 2004.

[19] J. Brugada, R. Brugada, and P. Brugada. Síndrome del bloqueo de rama derecha, elevación del segmento st de v1 a v3 y muerte súbita cardíaca. *Medicina General*, 1998.

[20] G.M. Pabón. Indicación de estudio electrofisiológico en pacientes con bloqueo intraventricular crónico. *Primer consenso sobre recomendaciones de estudio electrofisiológico y ablación con energía de radiofrecuencia*, 2001.

[21] G. ginsburg, N. Sunder, and P. Grace. Acute right bundle branch block as a presenting sign of acute pulmonary embolism. *Anesth Analg*, pages 789 – 791, 2006.

[22] I. Balaguer. Los estudios longitudinales en la prevención de las enfermedades cardiovasculares. *Revista Española de Salud Pública.*, pages 1135–5727.

[23] Z. Li, R. Devereux, S. Kjeldsen, S. Julius, K. Wachtell, H. Ibsen, M. Nieminen, S. Jern, P. Okin, and B. Dahlof. Association of left bundle branch block with cardiovascular morbidity and mortality in hypertensive patients with left ventricle hypertrophy: the life study. *Posters: Clinical Trials*.

[24] G. Engström, B. Hedblad, S. Juul-Möller, P. Tydén, and L. Janzon. Cardiac arrhythmias and stroke increased risk in men with high frequency of atrial ectopic beats. *Stroke, American Heart Association, Inc.*, pages 2925–2929, 2000.

[25] N.J. Holter. *New methods for heart studies*. Science.

[26] J. Dave, G. Gandhi, S. Nazeer, S. Jha, and A. Kocheril. Ventricular premature complexes. emedicine. *www.emedicine.com ([Online])*, 2005.

[27] H. Friedman. *Diagnostic electrocardiography and vectorcardiography*. Mc. Graw Hill, 3ra edition, 1989.

[28] Edilson Delgado-Trejos, César Germán Castellanos-Domínguez, and Montserrat Vallverdú-Ferrer. *Análisis de relevancia en espacios de representación orientado al soporte de diagnóstico*, volume 1 of *Notas de Automática Avanzada*. Facultad de Ingeniería y Arquitectura, Universidad Nacional de Colombia Sede Manizales, Manizales, Colombia, 2009.

[29] K. King. A model based toolkit for building medical diagnostic support systems in developing countries. In *IEE Colloquium on Intelligent Decision Support Systems and Medicine*, pages 911–916, 1992.

[30] Ethem Alpaydin. *Introduction to Machine Learning*. Massachusetts Institute of Technology, 2004.

[31] Zane Reynolds, Horst Bunke, Mark Last, and Abraham Kandel. Comparing representative selection strategies for dissimilarity representations. *International journal of intelligent systems*, 21:1093–1109, 2006.

[32] N. I. Moiseeva and V. V. Usov. Some medical and mathematical aspects of computer diagnosis. *IEEE Special Issue on Technology and Healt Services*, 57(11):1919– 1925, 1969.

[33] A. van Woerkom and K. Brodman. Statistics for a diagnostic mode. *Biometrics*, 17(2):299–318, 1961.

[34] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 2 edition, 1990.

[35] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley Sons, Inc., 2 edition, 2001.

[36] N. V. Takhor and D. Sherman. *The Biomedical Engineering Handbook*, chapter Wavelet (time-scale) analysis in biomedical signal processing, pages 886–906. CRC Press, 1995.

[37] Alfredo I. Hernández, Guy Carrault, Fernando Mora, and Alain Bardou. Model-based interpretation of cardiac beats by evolutionary algorithms: signal and model interaction. *Artificial Intelligence in Medicine. Elsevier*, 26(3):211–235, 2002.

[38] A. P. Nemirko, A. N. Kalinichenko, P. V. Murashov, N. N. Pogodin, D. V. Goncharov, S. E. Ivanova, C. M. Matous, and A. N. Sankin. Software complex for the recognition of diagnostically significant ECG changes. *Pattern Recognition and Image Analysis*, 16(1):9–11, 2006.

[39] M. El-Segaier, O. Lilja, S. Lukkarinen, L. Sörnmo, R. Sepponen, and E. Pesonen. Computer-based detection and analysis of heart sound and murmur. *Annals of Biomedical Engineering*, 33(7):937–942, 2005.

[40] K. Higuchi, K. Sato, H. Makuuchi, A. Furuse, S. Takamoto, and H. Takeda. Automated diagnosis of heart disease in patients with heart murmurs: application of a neural network technique. *Journal of Medical Engineering & Technology*, 30(2):61–68, 2006.

[41] Richard Ernest Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[42] Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.

[43] D. L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. In *Mathematical Challenges of the 21st Century*, 2000.

[44] Avrim L. Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.

[45] R. Rangayyan. *Biomedical signal analysis: A case-study approach*. IEEE press in Biomedical Engineering, 2002.

[46] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2 edition, 2005.

[47] M.L. Ahlstrom and J. Tompkins. Digital filters for real-time ecg signal processing using microprocessors. *IEEE Transactions on Biomedical Engineering*, 32(9):708–713, 1985.

[48] M. Okada. A digital filter for the qrs complex detection. *IEEE Transactions on Biomedical Engineering*, 26(12):700–703, 1979.

[49] L. Sörnmo. Time-varying digital filtering of ecg baseline wander. *Medical and Biological Engineering and Computers,*, 1993.

[50] J. Ramos. *Detección De Micropontenciales Auriculares dE Alta Frecuencia*. PhD thesis, Universidad Politécnica de Cataluña, 2001.

[51] N.J. Outram, E.C. Ifeachor, P.W. Van Eetvelt, and J.S.H. Curnow. Techniques for optimal enhancement and feature extraction of fetal electrocardiogram. *IEE Proceeding*, 142(6):482–489, Noviembre 1995.

[52] S. Olmos, J. García, R. Jané, and P. Laguna. Análisis de la señal ecg con expansiones ortogonales de reducido número de coeficientes. España, 1997. En Comunicaciones del XV Congreso Anual de la Sociedad Española de Ingeniería Biomédica.

[53] R. Limacher. Removal of power-line interference from the ecg signal by an adaptive digital filter. *Proc. of Europe Telemetry Conference*, 1996.

[54] H.C. So. A novel adaptive algorithm for sinusoidal interference cancellation. Technical report, University of Hong Kong, October 1997.

[55] D. Donoho. De-noising by soft-thresholding. *IEEE Trans.Information Theory*, 41(3):612–627, 1995.

[56] L.E. Avendaño and J.L.Rodríguez. Remoción y evaluación de la interferencia ac en ecg, usando denoising y la medida wdd. *Scientia et Technica*, 12(31):53–58, 2006.

[57] A.K. Ziarani and A. Konrad. A nonlinear adaptive method of elimination of power line interference in ecg signals. *IEEE transactions on biomedical engineering*, 49(6):540–547, 2002.

[58] George Takla, Kenneth Loparo, and Bala Nair. Matching a wavelet to ecg signal. In *28th IEEE EMBS Annual International Conference*, pages 1686–1689, New York city, New York, USA, August-September 2006. IEEE, The Printing House, Inc.

[59] M. Chapa and R. Rao. Algorithms for designing wavelets to match a specified signal. *IEEE Transaction on Signal Processing*, pages 3395–3406, 2000.

[60] T. Pawar, S. Chaudhuri, and S. Duttagupta. Analysis of ambulatory ecg signal. In *28th IEEE EMBS Annual International Conference*, pages 3094–3097, New York city, New York, USA, August-September 2006. IEEE, The Printing House, Inc.

[61] M. Milanesi, N. Martini, N. Vanello, V. Positano, M. F. Santarelli, R. Paradiso, D. De Rossi, and L. Landini. Multichannel techniques for motion artifacts removal from electrocardiographic signals. In *28th IEEE EMBS Annual International Conference*, pages 3391–3394, New York city, New York, USA, August-September 2006. IEEE, The Printing House, Inc.

[62] P. Laguna, R. Jané, O. Meste, P. W. Poon, P. Caminal, H. Rix, and N. V. Thakor. Adaptive filter for event-related bioelectric signals using an impulse correlated reference input: Comparison with signal averaging techniques. *IEEE trans. biomed. Eng.*, 39(10):1032–1044, October 1992.

[63] B-H. Tan, H. Shimizu, K. Hiromoto, Y. Furukawa, M. Ohyanagi, and T. Iwasaki. Wavelet transform analysis of hrv to assess the autonomic changes associated with spontaneous coronary spasm of variant angina. *Journal of Electrocardiography*, 36(2):117–124, April 2003.

[64] D. Cuesta. *Estudio de métodos para procesamiento y agrupación de señales electrocardiográficas*. PhD thesis, Universidad Politécnica de Valencia, 2001.

[65] J.S. Sahambi, S.N. Tandon, and R.K P. Bhatt. Using wavelet transform for ecg charecterization. *IEEE Transactions engineering in medice and biology*, pages 77–88, january 1997.

[66] J. Madeiro, P. Cortez, F. Oliveira, and R. Siqueira. A new approach to qrs segmentation based on wavelet bases and adaptive threshold technique. *Medical Engineering and Physics*, 2006.

[67] S. Kadambe, R. Murray, and G. F. Boudreaux-Bartels. Wavelet transform-based *QRS* complex detector. *IEEE Trans. on Biomed. Eng.*, 46(7):838–848, JULY 1999.

[68] R. Almeida, S. Gouveia, P. Rocha, P. Lago, and F. De Freitas. Wavelet transform based matlab system for the detection and delineation of qrs complexes in ambulatory ecg recordings. *in BioEng´2001 - 6th Portuguese Conference on Biomedical Engineering*, 2001.

[69] D. Benitez, P. Gaydecki, A. Zaidi, and A. Fitzpatrick. The use of the hilbert transform in ecg signal analysis. *Computers in Biology and Medicine*, pages 399–406, 2001.

[70] L. Wahab. Detection of qrs complexes in ecg signals based on poisson transform and root moments. Technical report, Imperial College London, South Kensington Campus, 2005.

[71] Natalia Arzeno, Chi-Sang Poon, and Zhi-De Deng. Quantitative analysis of qrs detection algorithms based on the first derivative of the ecg. In *28th IEEE EMBS Annual International Conference*, pages 1788–1791, New York city, New York, USA, August-September 2006. IEEE, The Printing House, Inc.

[72] Mirela Visinescu, Allen Bashour, Mohamed Bakri, and Bala Nair. Automatic detection of qrs complexes in ecg signals collected from patients after cardiac surgery. In *28th IEEE EMBS Annual International Conference*, pages 3724–3727, New York city, New York, USA, August-September 2006. IEEE, The Printing House, Inc.

[73] J. Martinez, R. Almeida, S. Olmos, A. Rocha, and P. Laguna. A wavelet-based ecg delineator: Evaluation standar databases. *IEEE Trans. on Biomed. Eng.*, 51(4):570–581, April 2004.

[74] R. Poli, S. Cagnoni, and G. Valli. Genetic design of optimum linear and nonlinear QRS detectors. *IEEE Transactions on Biomedical Engineering*, 42(11):1137–41, 1995.

[75] D. Brooks, H. Krim, J. Pesquet, and R. Mc Leod. Best basis segmentation of ecg signals using novel optimality criteria. pages 2750–2753, Atlanta, Georgia (USA), 1996. Proc. IEEE Conf. Acoust., Speech, Signal Processing.

[76] N. Hughes, S. Roberts, and L. Tarassenko. Semi-supervised learning of probabilistic models for ecg segmentation, 2005.

[77] C. Li, C. Zheng, and C. Tai. Detection of ecg characteristic points using wavelet transforms. *IEEE Trans.*, 42(1):21–28, January 1995.

[78] V. Afonso, W. Tompkins, and S. Luo. Ecg beat detection using filter banks. *IEEE Trans. on Biomedical Enginnering*, 46(2):192–202, February 1999.

[79] M. Bahura, M. Hassani, and M. Hubin. Dsp implementation of wavelet transform for real time ecg wave forms detection and heart rate analysis. *Comput. Meth. Programs Biomed.*, (52):35–44, February 1997.

[80] J. Lee, K. Jeong, J. Yoon, and M. Lee. A simple real-time qrs detection algorithm. In *in Proc. 18th Ann. Int. Conf.*, pages 1396–1398. IEEE Engineering in medicine and biology soc., August-September 1996.

[81] P.S. Hamilton and W.J. Tompkins. Quantitative investigation of qrs detection rules using the mit/bih arrhythmia database. *IEEE Trans. on Biom. Eng.*, 33(12):1157–1164, December 1986.

[82] J. Pan and W.J. Tompkins. A real-time detector algorithm. *IEEE Trans. on Biom. Eng.*, 32(3):230–236, March 1985.

[83] R. Poli, S. Cagnoni, and G. Valli. Genetic design of optimum linear and non linear qrs detectors. *IEEE Trans. on Biom. Eng.*, 42:1137–1141, November 1995.

[84] J Moraes, M. Freitas, and E. Costa G. A qrs complex detection algorithm using electrocardiogram leads. journal = IEEE Trans. on Biom. Eng., year = 2002, volume = 42, pages = 1137-1141, month = November,.

[85] F. Gritzali. Towards a generalized scheme for qrs detection in ecg waveforms. *Signal Processing*, 15:183==192, 1988.

[86] F. Jager. Feature extraction and shape representation of ambulatory electrocardiogram using the karhunen-loève transform. *Electrotechnical Review*, 69(2):83–89, march 2002.

[87] T. Stamkopoulos, K. Diamantaras, N. Maglaveras, and Michael Strintzis. Ecg analysis using nonlinear pca neural networks for ischemia detection. *IEEE Transactions On Signal Processing*, 46(11):3058–3066, Noviembre 1989.

[88] H. Haraldsson, L. Edenbrandt, and M. Ohlsson. Detecting acute myocardial infarction in the 12-lead ecg using hermite expansions and neural networks. Technical report, Department of theoretical Physics, Lund University, Sweden, 2002.

[89] A. Rosado, G. Camps, J. Guerrero, JV Francés, J. Muñoz, and AJ Serrano. Enhancing feature extraction for vf detection using data mining techniques. Technical report, Grup Processament Digital de Senyals (GPDS), Universitat de València, Spain, 2002.

[90] U. Kunzmann, G. Von Wagner, J. Schöchlin, and A. Bolz. Parameter extraction of ecg signals in real-time. *Biomedizinische Technik*, 47(1):875–878, 2002.

[91] H. Gholam Hosseini and H. Nazeran. Efficient features for ann-based ecg classifiers. Technical report, School of Engineering, The Flinders University of South Australia, 1999.

[92] L. Lhotská, M. Fejtová, J. Macek, and D. Novák. Biological data preprocessing: A case study. Technical report, Gerstner Laboratory, Departament of Cyber-

netics, Czech Technical University in Prague, Faculty of Electrical Engineering, 2003.

[93] R. Olszewski. *Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data*. PhD thesis, School of Computer Science Carnegie Mellon Univesity, Pittsburg, PA 15213, 2001.

[94] L. Sörnmo, P.O. Börjesson, M. Nygards, and O. Pahlm. A method for evaluation of qrs shape features using a mathematical model for the ecg. *IEEE Transactions on Biomedical Engineering*, 28(10), 1981.

[95] P. Aguirre, J.Cardelino, and N. Loeff. Sistema de detección, clasificación e identificación en línea decomplejos qrs. Technical report, Instituto de Ingeniería Eléctrica Facultad de Ingeniería Universidad de la República, Montevideo, Uruguay, 2002.

[96] J. García and P. Laguna. Diseño y evaluación de un detector de isquemia basado en la transformada de karhunen.loève. *Grupo de tecnología de las telecomunicaciones*.

[97] R.H. Rozo, editor. *Cardiología*, volume I. Sociedad Colombiana de Cardiología, first edition, 1999.

[98] R. Andreao, B. Dorizzi, and J. Boudy. Ecg signal analysis through hidden markov models. *IEEE Transactions on Biomedical Engineering*, 53(8):1541–1549, 2006.

[99] C. Torrence and G. Compo. A practical guide to wavelet analysis. *Bull. Am. Meteor. Soc.*, 79(1):61–78, 1998.

[100] R. Ceylan and Y. Özbay. Comparison of fcm, pca and wt techniques for classification ecg arrhythmias using artificial neural network. *Expert systems with applications*, 2006.

[101] A. Ahmad and L. Dey. A feature selection technique for classificatory analysis. 26(1):43–56, January 2005.

[102] Isabelle Guyon and Andr´e Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[103] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1:131–156, 1997.

[104] H. Liu and H. Motoda. Feature selection for knowledge discovery and data mining.

[105] Daphne Koller and Mehran Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.

[106] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273 – 324, 1997. Relevance.

[107] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46(1/3):389–422, 2002.

[108] Alain Rakotomamonjy. Variable selection using svm based criteria. *J. Mach. Learn. Res.*, 3:1357–1370, 2003.

[109] Manoranjan Dash A and Huanliu B. www.elsevier.com/locate/artint consistency-based search in feature selection. 2002.

[110] T. M. Cover and J. A. Thomas. *Elements of information theory*. 1991.

[111] Mark Last and Oded Maimon. A compact and accurate model for classification. *IEEE Trans. on Knowl. and Data Eng.*, 16(2):203–215, 2004.

[112] R. Battit. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 4(3):537–550, 1991.

[113] E. Grall-Maes and P. Beauseroy. Mutual information-based feature extraction on the time-frequency plane. *IEEE Trans. Sig. Proc.*, 50(4):779–790, 2002.

[114] D. Huang and Tommy W.S. Chow. Effective feature selection scheme using mutual information. *Neurocomputing*, 63:325 – 343, 2005. New Aspects in Neurocomputing: 11th European Symposium on Artificial Neural Networks.

[115] Mark Last, Abraham Kandel, and Oded Maimon. Information-theoretic algorithm for feature selection. *Pattern Recogn. Lett.*, 22(6-7):799–811, 2001.

[116] Edoardo Amaldi and Viggo Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theor. Comput. Sci.*, 209(1-2):237–260, 1998.

[117] Ch.M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, 1995.

[118] Rich Caruana and Dayne Freitag. Greedy attribute selection. In *Machine Learning, Proceedings of the Eleventh International Conference on Machine Learning, (ICML 1994)*, pages 28–36, New Brunswick, New Jersey), 1994. Morgan Kauffmann.

[119] P Somol, P Pudil, and J Kittler. Fast branch and bound algorithms for optimal feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7):900–912, July 2004.

[120] J. Doak. An evaluation of feature selection methods and their application to computer security. Technical Report CSE-92-18, Department of Computer Science, University of California, Davis, CA, 1992.

[121] B. Bhanu and Y.Q.A. Lin. Genetic algorithm based feature selection for target detection in sar images. 21(7):591–608, July 2003.

[122] M. Kudo. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33(1):25–41, January 2000.

[123] Shlomo Zilberstein. Using anytime algorithms in intelligent systems, 1996.

[124] Il-Seok Oh, Jin-Seon Lee, and Byung-Ro Moon. Hybrid genetic algorithms for feature selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(11):1424–1437, 2004.

[125] K. Minami, H. Nakajima, and T. Yoyoshima. Real time discrimination of the ventricular tachyarrhythmia with fourier-transform neural network. *IEEE trans. on. Biomed.*, 46:179–185, 1999.

[126] Ying-Hsiang Chen and Sung-Nien Yu. Comparison of different wavelet subband features in the classification of ecg beats using probabilistic neural network. In *28th IEEE EMBS Annual International Conference*, pages 1398–1401, New York city, New York, USA, August-September 2006. IEEE, The Printing House, Inc.

[127] Recomended practice for testing and reporting performance results of ventricular arrhythmia detection algorithms. Technical report, AAMI (Association for the Advancement of Medical Instrumentation, 1998.

[128] Y. Hu, S. Palreddy, and W. Tompkins. A patient adaptable ecg beat classfier using a mixture of experts approach. *IEEE trans. on. Biomed.*, 44:891–900, 1997.

[129] P. Chazal, M. O'Dwyer, and R. Reilly. Automatic classification of hearbeats using ecg morphology and heartbeat interval features. *IEEE trans. on. Biomed.*, 51:1196–1206, 2004.

[130] D. Cuesta, J. Pérez-Cortés, and G. Andreau-García. Clustering of ecg signals in computer-aided holter analysis. *Computer Methods and Programs in Biomedicine*, 72:179–196, 2003.

[131] B. Köhler, C. Henning, and R. Orglmeister. The principles of software qrs detection: reviewing and comparing algorithms for detecting this important ecg waveform. *IEEE tran. on Biomed.*, 21:42–57, 2002.

[132] D. Cuesta, M. Biagetti, R. Quintero P. Micó-Tormos, and M. Aboy. Unsupervised detection of ventricular extrasystoles using bounded clustering algorithms and morphology matching. *IEEE tran. on Biomed.*, 2006.

[133] M. Paoletti and C. Marchesi. Discovering dangerous patterns in long-term ambulatory ecg recordings using a fast qrs detection algorithm and explorative data analysis. *Computer Methods and programs in biomedicine*, 82:20–30, 2006.

[134] B. Zhang, M. Hsu, and U. Dayal. K-harmonic-means, a data clustering algorithm. Technical report, Hewlett Packard Labs, 1999.

[135] K. Chang, R. Lee, C. Wen, and M. Yeh. Comparison of similarity measures for clustering ecg complexes. In *32 annual international conference on computers in cardiology*, Lyon, France, September 2005. INSERM, UCBL, INSA, HCL.

[136] Dean Cvetkovic, Elif Derya Übeyli, and Irena Cosic. Wavelet transform feature extraction from human ppg, ecg, and eeg signal responses to elf pemf exposures: A pilot study. *Digital Signal Processing*, 18(5):861 – 874, 2008.

[137] Amir homayoon Najmi and John Sadowsky. The continuous wavelet transform and variable resolution time-frequency analysis.

[138] H. So. Adaptive cancellation of multiple interfering sinusoids. *Electronic Letters*, 34(24):2301–2303, November 1998.

[139] D. Cuesta-Frau et al. Reducción de ruido en señales electrocardiográficas mediante la transformada wavelet. In *Libro de Actas XVIII Congreso Anual de la Sociedad de Ingeniería Biomédica*, pages 103–106, Cartagena, España, September 2000. CASEIB.

[140] O. Pahlm and L. Sörno. Software QRS detection in ambulatory monitoring: a review. *Medical & biological engineering & computing*, 22(4):289–297, 1984.

[141] T.Allan Pryor, Richard Russell, Alberto Budkin, and W.Gary Price. Electrocardiographic interpretation by computer. *Computers and Biomedical Research*, 2(6):537 – 548, 1969.

[142] P. Laguna, N. V. Thakor, P. Caminal, and R. Jane. Low-pass differentiators for biological signals with known spectra : application to ECG signal processing. *IEEE transactions on biomedical engineering*, 37(4):420–425, 1990.

[143] Per Ola Börjesson, Olle Pahlm, Leif Sörnmo, and M.E. Nygards. Adaptive qrs detection based on maximum a posteriori estimation. *IEEE Transactions on Biomedical Engineering*, 29:341–351, 1982.

[144] W. A. H. Engelse and C. Zeelenberg. A single scan algorithm for qrs-detection and feature extraction. *Computers in Cardiology*, 6:37–42, 1979.

[145] Pierre Dutilleul. Multi-frequential periodogram analysis and the detection of periodic components in time series. *Communications in statistics*, 2001.

[146] Matthias Emmert-Streib, Frank; Dehmer. *Information Theory and Statistical Learning*. Springer, 2009.

[147] V. Perlibakas. Distance measures for pca-based face recognition. *Pattern Recognition Letters*, 25(16):711–724, April 2004.

[148] J. L. Rodríguez Sotelo, D. Cuesta Frau, D. Peluffo Ordónez, G. Castellanos Domínguez, and D. Novak. Unsupervised feature selection in cardiac arrhythmia analysis. *Engineering in medicine and conference. EMBC*, 2009.