A Comparative Study of Weighting Factors for WPCA Based on a Generalized Distance Measure

J.S. Castaño-Candamil, S. Garcia-Vega, D.H. Peluffo-Ordoñez, C.G. Castellanos-Domínguez

***Abstract*—Category 2. This work presents a comparative analysis between different weighting factors for PCA, obtained from a generalized distance measure. By employing such generalized distance, an optimization problem is established whose solution provides relevance values for each feature. Relevance values are used both for weighting data in a WPCA scheme to carry out a feature extraction and also for selecting relevant features. Their performance and suitability are assessed in terms of dimensionality reduction and classification performance achieved by applying both supervised and unsupervised classification techniques over real datasets.**

***Index Terms*— Feature selection, Feature extraction, WPCA, Eigen-decomposition, Data projection**

I. Introduction

In the area of pattern recognition, the problems of selecting relevant features from a data set or finding the appropriate data projection with the purpose of performing a classification task are typical and are usually found in many areas of science [1], namely, face recognition, analysis of biomedical data, motion analysis in computer vision systems, among others. To choose a method for feature selection or projection, many different factors must be taken into account: computational cost, classification performance, nature of database and specific problem to be solved. In order to perform this task, the Principal Component Analysis (PCA) represents a good alternative because of its non-parametric nature, easiness for implementation and versatility [2]. Nevertheless, when data are not statistically independent or cannot be obtained from a linear combination, PCA does not apply. Additionally, traditional PCA is not advisable when data contains outliers and noise [3].

For this reason, some variants of PCA have been introduced, focused on data separability, that are based on a weighted projection of data (known as WPCA), in such way that performed projection or features identified as relevant, achieve a good classification performance even when such projection or such selected features are not the best representation of the dataset.

In this work, it is proposed a comparative analysis of some weighting methods for unsupervised WPCA, based on a generalized distance measure, which leads to some approaches, such as, the MSE based-method, the M-inner product based-method regarding to a weighted affinity matrix (introduced in [1] as *Power-Embedded $Q-\alpha$ Method*). The suitability of each method in both feature selection processes and projection is suggested. In addition, it is used an approach as an alternative of $Q-\alpha$ method using a sub-optimal pre-established rotation matrix Q which avoids the iterative tuning procedure.

The remainder of this paper is structured as follows. In Section II, it is described the experimental set used, also the theoretical context and development of the studied methods is presented. In Section III, the results are presented and discussed in terms of classification performance, computational cost, data separability and dimensionality reduction. Finally In Section IV, the conclusions of this paper and future work are presented.

J.S. Castaño-Candamil (jscastanoc@unal.edu.co), S. Garcia-Vega (segarciave@unal.edu.co), D.H. Peluffo-Ordoñez (dhpeluffoo@unal.edu.co), C.G. Castellanos-Domínguez (cgcastellanosd@unal.edu.co) are with Department of Electrical and Electronic Engineering, Universidad Nacional de Colombia sede Manizales, Colombia.

II. Materials and Methods

A. Used Datasets

Four datasets of the *UCI Machine learning repository* [4] are used. Each dataset is normalized regarding its mean and standard deviation and consist of only 2 classes in all cases. The number of features and observations of each dataset can be seen in Table I[1].

TABLE I: Datasets description.

Name	Features	Observations
Breast	9	699
P.I Diabetes	8	768
Heart	13	303
Ionosphere	34	351
E.coli	8	336

The performance of each weighting method is assessed in terms of the quality of classification achieved by applying the following classification methods on weighted relevant data (feature selection) and on projected data (feature extraction by means of WPCA) .

- Unsupervised methods:
  - K-means.
  - Multi-spectral clustering (as introduced in [5]).
- Supervised methods (70% of data for training, 30% for validation):
  - K-nearest neighbours (15 folds were used in all cases).
  - Linear Bayes Normal Classifier.

The performance appraisal of classification is conducted through the estimation of both supervised and unsupervised

[1]The E.coli dataset, only two classes are used, only for graphical analysis, not for classification performance indices

performance indices (only supervised indices on the supervised classification methods were used): sensitivity (S_e), specificity(S_p), classification percentage (CP), and clusters coherence (ϵ_M). The latter as introduced in [6].
The selection of the features is based on a normalized accumulative variance measure. Experiments were carry out such a way that 85% of variance was represented by the selected or extracted features.
As references, two weighting methods are used, on the one hand, an observations weighting method[6] described by

$$\boldsymbol{w}_i = \frac{1}{\sqrt{\frac{1}{q}\sum_{j=1}^{q} \boldsymbol{X}_{ij}^2}} \tag{1}$$

where $\boldsymbol{X}$ correspond to dataset and $\boldsymbol{w}$ correspond to the weighting vector.
On the other hand, an eigenvalues based weighting method [7] is used as well, in which the weighting vector is defined as follows

$$\boldsymbol{w}_i = \lambda_i^{-1/2} \tag{2}$$

This weighting vector is applied on the projected dataset, therefore, it can only be applied on feature extraction processes.

B. Weighted PCA

Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be the data matrix being centred i.e $\boldsymbol{\mu} = 0$, then the respective linear projection is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{V} \in \mathbb{R}^{n\times p}$, being $\boldsymbol{V}$ an orthogonal matrix. Generally, the projection is performed to a q-dimensional space i.e. $\widehat{\boldsymbol{X}} = \widehat{\boldsymbol{Y}}\widehat{\boldsymbol{V}}^{\mathsf{T}} \in \mathbb{R}^{n\times p}$ where the representation quality of $\boldsymbol{X}$ is given by an error function ϵ which can be expressed as a distance measure $\epsilon = d(\boldsymbol{X}, \widehat{\boldsymbol{X}})$. The main aim is to perform a linear projection based on a generalized weighting covariance matrix (weighted relevance matrix) $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{X}}$ which is defined as follows [8]:

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

where $\boldsymbol{W} = \text{ diag}(\sqrt{\boldsymbol{v}})$ is a weighting matrix and $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{W}$. To that end, a distance-based method has been considered, which employs M-norm as distance measure. This approach combines dimensionality reduction with variable selection.

1) Generalized case

The objective function that M-inner product based method minimizes is given by

$$\min_{\widehat{\boldsymbol{V}}} ||\boldsymbol{X} - \widehat{\boldsymbol{Y}}\widehat{\boldsymbol{V}}^{\mathsf{T}}||_{\boldsymbol{A}}^2 = \min_{\widehat{\boldsymbol{V}}} ||\boldsymbol{X} - \widehat{\boldsymbol{X}}||_{\boldsymbol{A}}^2$$
$$\text{s. t.} \quad \widehat{\boldsymbol{V}}^{\mathsf{T}}\widehat{\boldsymbol{V}} = \boldsymbol{I}_q$$

where $A \in \mathbb{R}^{n\times n}$ is any matrix.
In order to understand better this optimization problem, following are presented some important considerations and demonstrations. First, consider

$$||\boldsymbol{X} - \widehat{\boldsymbol{X}}||_{\boldsymbol{A}}^2 = \text{ tr}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{X} - \widehat{\boldsymbol{X}}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{X} - \boldsymbol{X}^{\mathsf{T}}\boldsymbol{A}\widehat{\boldsymbol{X}} + \widehat{\boldsymbol{X}}^{\mathsf{T}}\boldsymbol{A}\widehat{\boldsymbol{X}})$$

where $\langle\cdot,\cdot\rangle_{\boldsymbol{A}}$ denotes the M-inner product regards matrix $\boldsymbol{A}$. Also, other term of interest is $||\boldsymbol{X}||_{\boldsymbol{A}}^2$ that can be expressed as

$$||\boldsymbol{X}||_{\boldsymbol{A}}^2 = \text{ tr}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{X}) = \text{ tr}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^T) = \beta\sum_{j=1}^{p}\lambda_j$$

where λ_j denotes the j-th eigenvalue of $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ and $\beta = \text{ tr}(\boldsymbol{A})$. In the previous expression is evident that β must be greater than 0 (due to the norm conditions) and one way to guarantee this is choosing matrix $\boldsymbol{A}$ as a positive semi-definite matrix. Also, it is possible to write that

$$||\boldsymbol{X} - \widehat{\boldsymbol{X}}||_{\boldsymbol{A}}^2 = \beta\sum_{j=1}^{p}\lambda_j - 2\beta\sum_{j=1}^{q}\lambda_j + \beta\sum_{j=1}^{q}\lambda_j = \beta\sum_{j=q+1}^{p}\lambda_j$$

Having this in mind, it can be concluded that

$$||\boldsymbol{X}||_{\boldsymbol{A}}^2 = \beta\sum_{j=1}^{q}\lambda_j + ||\boldsymbol{X} - \widehat{\boldsymbol{X}}||_{\boldsymbol{A}}^2 = \text{ tr}(\widehat{\boldsymbol{X}}^T\boldsymbol{A}\boldsymbol{X}) + ||\boldsymbol{X} - \widehat{\boldsymbol{X}}||_{\boldsymbol{A}}^2$$

Because $||\boldsymbol{X}||_{\boldsymbol{A}}^2$ is constant, this is a dual optimization problem where minimizing the error function $||\boldsymbol{X} - \widehat{\boldsymbol{X}}||_{\boldsymbol{A}}^2$ is the same as maximizing its complement $\text{ tr}(\widehat{\boldsymbol{X}}^T\boldsymbol{A}\boldsymbol{X})$. Therefore, the optimization problem can be re-written as

$$\max_{\widehat{\boldsymbol{V}},\, \boldsymbol{A}} \text{ tr}(\widehat{\boldsymbol{V}}^T\boldsymbol{X}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{X}\widehat{\boldsymbol{V}}) \tag{3}$$

2) MSE-based Approach

If $\boldsymbol{A} = \boldsymbol{I}_n$ the initial problem is reduced to be:

$$\min_{\widehat{\boldsymbol{V}}} ||\boldsymbol{X} - \widehat{\boldsymbol{X}}||_{\boldsymbol{I}_n}^2 = \min_{\widehat{\boldsymbol{V}}} ||\boldsymbol{X} - \widehat{\boldsymbol{X}}||_2^2$$

where $||\cdot||_2$ represents the euclidean norm. In fact, when $\boldsymbol{A} = \boldsymbol{I}_n$, the term $||\boldsymbol{X} - \widehat{\boldsymbol{X}}||_{\boldsymbol{A}}^2$ can be expressed as:

$$||\boldsymbol{X} - \widehat{\boldsymbol{X}}||_{\boldsymbol{I}_n}^2 = ||\boldsymbol{X} - \widehat{\boldsymbol{X}}||_2^2$$

By applying an expected value operator and re-written the problem as an objective function to be maximized, it can be reached the following optimization problem

$$\max_{\widehat{\boldsymbol{V}}}\{\text{ tr}(\widehat{\boldsymbol{V}}^T\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\widehat{\boldsymbol{V}})\} = \max_{\widehat{\boldsymbol{V}}}\mathcal{E}\{\text{ diag}(\widehat{\boldsymbol{V}}^T\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\widehat{\boldsymbol{V}})\}$$

where $\mathcal{E}\{\cdot\}$ is the expected value which is considered as an arithmetic average. Previous optimization problem is called mean square error (MSE)-based approach. Which can be developed as follows

$$\mathcal{E}\{\text{ diag}(\widehat{\boldsymbol{V}}^T\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\widehat{\boldsymbol{V}})\} = \frac{1}{q}\sum_{j=1}^{q}\lambda_j\text{ tr}(\boldsymbol{v}_j\boldsymbol{v}_j^T)$$

Since $\text{ tr}(\boldsymbol{v}_{i\cdot 2}) = \boldsymbol{v}_j.^2$ then, it can be defined a resulting vector:

$$\boldsymbol{\rho} = \frac{1}{q}\sum_{j=1}^{q}\lambda_j\boldsymbol{v}_j.^2$$

where $\boldsymbol{v}_j.^2$ represents each element of $\boldsymbol{v}_j$ squared. In this case, the values of $\boldsymbol{\rho}$ correspond to a relevance index which express the accumulated variance of eigenvalues and eigenvectors.
Thus, a weighting matrix $\boldsymbol{W}$ can be calculated as $\boldsymbol{W} = \text{diag}(\sqrt{\boldsymbol{\rho}})$. An approximated weighting matrix $\widehat{\boldsymbol{W}}$ can be found by using only the most relevant eigenvalue and its respective eigenvector, i.e. $\widehat{\boldsymbol{\rho}} = \lambda_1\boldsymbol{v}_1.^2$ and $\widehat{\boldsymbol{W}} = \sqrt{\widehat{\boldsymbol{\rho}}}$.

3) $Q-\alpha$ Method

Other particular case arises when $\boldsymbol{A}=\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}$: Because of the given conditions of matrix $\boldsymbol{A}$ must satisfy, it can be chosen as the inner product between observations so that $\boldsymbol{A}=\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}$. In spectral clustering context, this matrix represents the trivial affinity matrix [9]. By replacing $\boldsymbol{A}$ in equation (3):

$$\text{tr}(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{A}\boldsymbol{Q})=\sum_{j=1}^{q}\lambda_j^2$$

where $\boldsymbol{Q}\in\mathbb{R}^{n\times q}$ is an arbitrary orthonormal matrix.
So, it can be introduced optimization problem given by

$$\max_{\boldsymbol{Q}}\ \text{tr}(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{A}\boldsymbol{Q})=\sum_{j=1}^{q}\lambda_j^2 \tag{4}$$

Now, redefining $\boldsymbol{A}$ as $\boldsymbol{A}_{\boldsymbol{\alpha}}=\sum_{i=1}^{p}\boldsymbol{\alpha}_i\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}_i=\boldsymbol{X}\ \text{diag}(\boldsymbol{\alpha})\boldsymbol{X}^{\mathsf{T}}$, where $\boldsymbol{\alpha}\in\mathbb{R}^p$ and $\boldsymbol{x}_i$ corresponds to i-th column of $\boldsymbol{X}$. In order to satisfy the conditions given by equation (4), it is necessary that $\text{tr}(\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{A}_{\boldsymbol{\alpha}})=\sum_{j=1}^{p}\lambda_j^2$ and therefore $\boldsymbol{\alpha}$ must be unitary, i.e., $||\boldsymbol{\alpha}||_2^2=\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\alpha}=1$. This method was introduced in [1] and is called $Q-\alpha$ method. Then, $Q-\alpha$ objective function can be written as:

$$\max_{\boldsymbol{Q},\,\boldsymbol{\alpha}}\ \text{tr}(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{Q})=\sum_{j=1}^{q}\lambda_j^2$$

The weight vector $\boldsymbol{\alpha}$ and the orthonormal matrix $\boldsymbol{Q}$ are determined at the maximal point of the optimization problem. Finally, the objective function can be rewriting as the following quadratic form:

$$\max_{\boldsymbol{\alpha}}\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{\alpha} \tag{5}$$

where $\boldsymbol{G}\in\mathbb{R}^{p\times p}$ is a matrix with elements $g_{ij}=(\boldsymbol{x}_i\ \boldsymbol{x}_j^T)\ \boldsymbol{x}_i\ \boldsymbol{Q}\ \boldsymbol{Q}^{\mathsf{T}}\ \boldsymbol{x}_j^T,\ \ i,j=1,\ldots,p$. As consequence, the previous equation becomes the objective function to be used in the unsupervised $Q-\alpha$ algorithm, as described in [1]. The matrix $\boldsymbol{G}$ is obtained from an arbitrary orthonormal transformation, it is necessary to apply an iterative method to tune the matrix $\boldsymbol{Q}$ and the weighting vector $\boldsymbol{\alpha}$. In this procedure, the whole data set is used, where the orthonormal matrix is updated per iteration to get the subset of relevant features. As a result, the computational load may increase. Nonetheless, based on variance criterion, it can be inferred that the first q components of $\widehat{\boldsymbol{x}}^{(l)}$ hold the most informative directions of weighting data, thus, the l $(q+1\leq l\leq p)$ directions do not contribute significantly to the explained variance. Then, time calculation when computing the vector $\boldsymbol{\alpha}$ can be reduced just to one iteration with no significant decrease of accuracy [1]. With this in mind, the feature relevance may be preserved optimizing the p original variables or the first q variables. Indeed, maximizing $\text{tr}(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{Q})$ is equivalent to maximize $\text{tr}(\boldsymbol{A}_{\boldsymbol{\alpha}}\boldsymbol{A}_{\boldsymbol{\alpha}})=\text{tr}(\boldsymbol{X}\ \text{diag}(\boldsymbol{\alpha})\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\ \text{diag}(\boldsymbol{\alpha})\boldsymbol{X}^{\mathsf{T}})$. Since this expression is bilinear regarding $\boldsymbol{\alpha}$, the objective function can be re-written as $\boldsymbol{\alpha}^T\boldsymbol{H}\boldsymbol{\alpha}$, where $\boldsymbol{H}_{ij}=\text{tr}(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}_i\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{x}_j)=\boldsymbol{x}_i\boldsymbol{x}_j^{\mathsf{T}}\ \text{tr}(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}_j)=(\boldsymbol{x}_i\boldsymbol{x}_j^{\mathsf{T}})^2$.

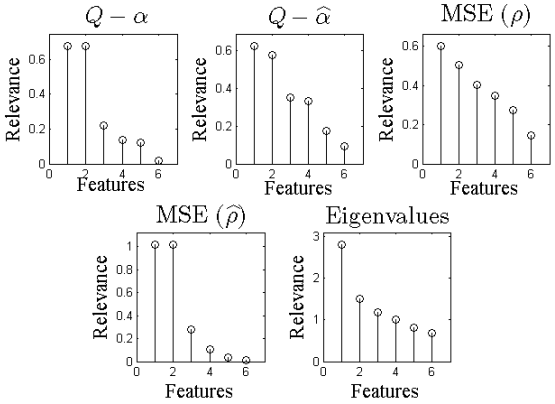

Fig. 1: Relevance values for 'E.coli' dataset.

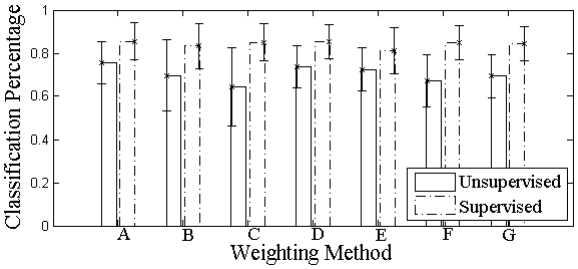

Fig. 3: Classification percentage achieved with each weighting method performing feature extraction (applying PCA). A)No Weighting. B)PCA pre-normalization. C) Eigenvalues weighting. D)$Q-\alpha$ method. E)$Q-\widehat{\alpha}$ method. F)ρ method. G)$\widehat{\rho}$ method.

Accordingly, it can be inferred that the approximate vector of relevance $\widehat{\boldsymbol{\alpha}}$ is the eigenvector corresponding to the largest eigenvalue of $(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{\cdot 2}$ (where notation $(\boldsymbol{\chi})^{\cdot 2}$ stands for the square of each one of the elements of the involved matrix $\boldsymbol{\chi}$). In conclusion, the weighting factor is related to either vectors: $\boldsymbol{\alpha}$ (complete case) and $\widehat{\boldsymbol{\alpha}}$ (approximate case).

III. Results and Discussion

Figure 1 shows an example of features relevance (e.coli dataset). It can be seen that α, $\widehat{\alpha}$ and $\widehat{\rho}$ present a more noticeable difference between features selected as relevant and features selected as non-relevant, this suggest a better performance in terms of data separability when these three methods are used. This hypothesis is supported by Figure 2, which shows the three first principal components of e.coli dataset using the studied weighting methods. It is noticeable the improvement of performance, in terms of data separability, when α, $\widehat{\alpha}$ and $\widehat{\rho}$ are used as weighting factors, indicating that if the same number of dimensions are used on a classification task, the performance of these three weighing methods would be superior than the result obtained when the other weighting methods are used. Although, the performance in terms of data separability cannot be generalized, it has been seen a similar behaviour on the other used datasets.

A. Suitability on feature extraction tasks

The performance of the different classification methods used on weighted and projected data (a feature extraction

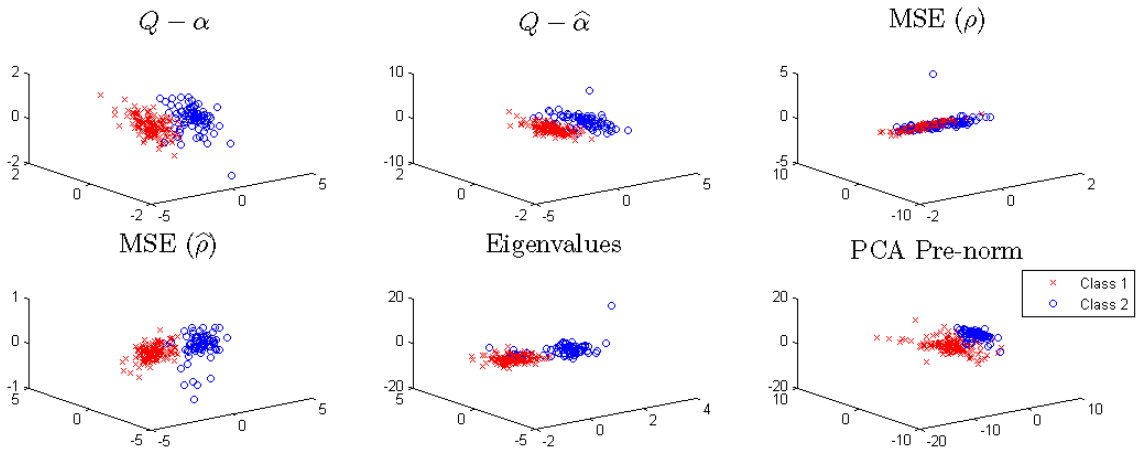

Fig. 2: Three first principal components of E.coli dataset, applying each weighting method.

process by means of PCA), can be seen on Figure 3 and Table II. Firstly, it can be noticed that the performance of supervised techniques is fairly uniform, while unsupervised techniques show a decrease in certain methods, mainly the eigenvalues and MSE based weighting methods. In spite of the uniform performance measures, it can be observed that the size of dataset used on classification tasks on weighted data (using α, $\widehat{\alpha}$, ρ, $\widehat{\rho}$ as weighting factors) is much smaller than the obtained by means of the reference techniques (including traditional PCA). Therefore, the computational cost of the classification tasks is less due to the reduction of dimensionality obtained with the studied weighting methods. The $Q-\alpha$ method did not represent a meaningful improvement with regard to dimensionality reduction compared to weighting methods used in this paper as a reference. Furthermore, it did not show a better performance even compared to traditional PCA (without weighting). Nevertheless, in some datasets, its performance was slightly superior compared to other methods. The results obtained on classification tasks by means of $Q-\alpha$ weighting, could suggest that this method is not very suitable in feature extraction, at least compared to the other methods with a fixed variance representation, because of the fact that the dimensionality reduction and the classification performance achieved is not significantly superior compared to the other weighting methods which are less computationally expensive. However, the approximate version of $Q-\alpha$ showed a somewhat better dimensionality reduction without a significant loss of classification performance and avoiding the iterative nature of standard $Q-\alpha$.
Regarding to the methods based on MSE, these techniques present a greater dimensionality reduction (down to 11.1% of the original dataset size). This dimensionality reduction is reflected in an improvement of the computational cost on classification tasks, but also in the loss of classification performance mainly in clustering techniques.

B. Suitability on feature selection tasks

The performance of the different classification methods used on weighted data (this means without applying PCA) can be seen on Figure 4 and Table III. Using the weighting factors

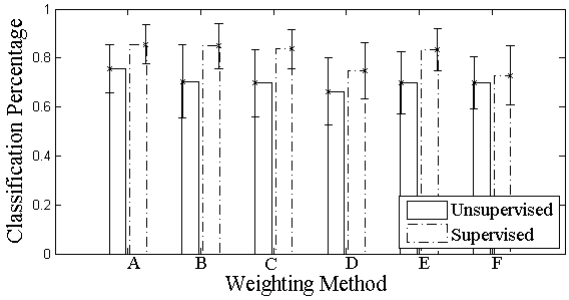

Fig. 4: Classification percentage achieved with each weighting method performing feature selection. A)No Weighting. B)PCA pre-normalization. C)$Q-\alpha$ method. D)$Q-\widehat{\alpha}$ method. E)ρ method. F)$\widehat{\rho}$ method.

as a criterion for selecting the most relevant features, it was found that the greater dimensionality reduction was achieved by means of $Q-\widehat{\alpha}$. Additionally, this method presents a very low computational cost at the expense of loss classification performance (compared to the other weighting techniques), mostly on supervised classification processes.
Regarding $\widehat{\rho}$, it is the best weighting method in terms of computational cost but also is the weighting factor that causes the greatest loss of classification percentage (mainly on supervised classification tasks) among the used techniques.
On the other hand, α and ρ are the factors whose performance (in terms of dimensionality reduction) is not very good but its performance in terms of classification percentage is evidently superior, mainly on supervised classification techniques. Furthermore, the classification performance of the weighting factor ρ is slightly inferior than the obtained by means of α, however, the computational cost of the weighting factor ρ is much smaller than the iterative α.

IV. Conclusions and Future Work

The best data separability results were obtained by means of α, $\widehat{\alpha}$ and $\widehat{\rho}$ as weighting factors. This behaviour is explained by the differences between relevant and non relevant features, which are much more noticeable in these

factors, in comparison to the other methods studied, where such difference is more uniform between all features. The latter causes a relatively bad data separability. $\widehat{\alpha}$ vector (obtained by means of a non iterative method) showed a great performance, with a much smaller computing time in comparison to standard $Q-\alpha$ but at the same time, keeping almost intact the attributes that cause a remarkable data separability. From the point of view of the classifiers performance, it can be seen that when a feature selection task was carried out, ρ and α were the weighting factors that indicated a better performance in terms of dimensionality reduction and classification quality, although, it is noteworthy that $\widehat{\alpha}$ classification performance was not bad and additionally achieved a remarkable dimensionality reduction (being 17% of the original data size the average reduction among the tested datasets). Furthermore, when a feature extraction task was carried out, the studied methods showed an homogeneous classification performance, nevertheless, $\widehat{\rho}$ presented the greatest dimensionality reduction in comparison to the other studied methods.

As future work, other weighting factors based on a distance measure should be studied. Additionally, multi-class datasets with a more complex nature (especially, a more complex separability) will be used.

REFERENCES

[1] L. Wolf and A. Shashua, "Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach," *Journal of machine learning*, vol. 6, pp. 1855 – 1887, 2005.

[2] V. Perlibakas, "Distance measures for pca-based face recognition," *Pattern Recognition Letters*, vol. 25, no. 6, pp. 711 – 724, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865504000248

[3] J. Rodriguez-Sotelo, E. Delgado-Trejos, D. Peluffo-Ordonez, D. Cuesta-Frau, and G. Castellanos-Dominguez, "Weighted-pca for unsupervised classification of cardiac arrhythmias," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, 31 2010-sept. 4 2010, pp. 1906 –1909.

[4] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[5] S. Y. Jianbo, S. X. Yu, and J. Shi, "Multiclass spectral clustering," in *In International Conference on Computer Vision*, 2003, pp. 313–319.

[6] D. Peluffo-Ordonez, "Estudio comparativo de metodos de agrupamiento no supervisado de latidos de senales ecg," Master's thesis, Universidad Nacional de Colombia Sede Manizales, 2009.

[7] H. Y. Wang and X. J. Wu, "Weighted pca space and its application in face recognition," in *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, vol. 7, 2005. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1527735

[8] H. H. Yue and M. Tomoyasu, "Weighted principal component analysis and its applications to improve fdc performance," in *Conference on decision and control*, 2004.

[9] S. X. Yu and J. Shi, "Multiclass spectral clustering," in *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*. Washington, DC, USA: IEEE Computer Society, 2003, p. 313.

Juan Sebastián Castaño Candamil He is currently student of Electronical Engineering at the Universidad Nacional de Colombia Sede Manizales. His main research interest are machine learning and its application in biosignals processing

Sergio Garcia Vega He is currently student of Electronical Engineering and Mathematics at the Universidad Nacional de Colombia Sede Manizales. His main research interest are digital signal processing and applied maths

Diego Hernan Peluffo-Ordoñez He received his degree in electronic engineering and the M.Eng. degree in industrial automation from the Universidad Nacional de Colombia, Manizales, Colombia, in 2008 and 2010, respectively. Currently, he is PhD student in the same university. His main research interests are applied maths and unsupervised learning and their applications in biosignals analysis.

César Germán Castellanos-Domínguez He received his undergraduate degree in radiotechnical systems and his Ph.D. in processing devices and systems from the Moscow Technical University of Communications and Informatics, in 1985 and 1990 respectively. Currently, he is a Full Professor in the Department of Electrical, Electronic and Computer Engineering at Universidad Nacional de Colombia Sede Manizales. He is also the Leader of the Signal Processing and Recognition Group at the same university. His teaching and research interests include information and signal theory, digital signal processing and bioengineering.

APPENDIX

A. Other results

Following are presented another experimental results in tables II and III.

TABLE II: Classification performance applying feature extraction. Time correspond to the time needed to calculate the weighting vector, R% is the dimensionality reduction being 100% the original dataset

	Breast		Diabetes		Heart		Ionosphere		Breast		Diabetes		Heart		Ionosphere	
	k-means	Spectral	k-means	Spectral	k-means	Spectral	k-means	Spectral	KNN	LCD	KNN	LCD	KNN	LCD	KNN	LCD
Without Weighting																
CP	0.96	0.701	0.706	0.646	0.832	0.76	0.706	0.741	0.96732	0.961	0.732	0.760	0.821	0.831	0.884	0.883
S_p	0.957	0.746	0.562	0.444	0.885	0.694	0.572	0.594	0.974	0.936	0.45 6	0.567	0.74	0.784	0.719	0.721
S_e	0.962	0.696	0.82	0.656	0.799	0.85	0.818	0.909	0.96391	0.97494	0.88178	0.86356	0.89028	0.87083	0.97512	0.97214
ϵ_M	0.9741	0.7465	0.8333	0.9518	0.758	0.757	0.8916	0.9206	–	–	–	–	–	–	–	–
Time	–															
R %	58.4625 ± 15.8283															
PCA pre-norm																
CP	0.957	0.636	0.648	0.649	0.825	0.778	0.692	0.396	0.97026	0.95098	0.71942	0.69333	0.82547	0.8367	0.9	0.769
S_p	0.941	0.378	0	0.474	0.89	0.729	0.553	0.052	0.97559	0.90798	0.48167	0.19167	0.73496	0.79837	0.827	0.466
S_e	0.966	0.652	0.65	0.654	0.787	0.83	0.823	0.525	0.96742	0.97393	0.84622	0.96089	0.90278	0.86944	0.9403	0.936
ϵ_M	0.9807	0.9394	0.919	0.818	0.7533	0.7476	0.8902	0.911	–	–	–	–	–	–	–	–
Time	0.0649 ± 0.0038															
R %	–															
Eigenvalues																
CP	0.965	0.386	0.667	0.646	0.529	0.727	0.746	0.4814	0.95229	0.96111	0.73652	0.75652	0.79176	0.83446	0.902	0.865
S_p	0.961	0.216	0.523	0.447	0.465	0.7	100	0	0.93427	0.93052	0.4575	0.56917	0.64065	0.77724	0.753	0.679
S_e	0.967	0.533	0.738	0.656	0.539	0.752	0.717	0.573	0.9619	0.97744	0.88533	0.85644	0.92083	0.88333	0.985	0.968
ϵ_M	0.9123	0.9775	0.82	0.933	0.7253	0.6892	0.707	0.916	–	–	–	–	–	–	–	–
Time	0.0638 ± 0.0143															
R %	58.5200±15.8797															
$Q - \alpha$																
CP	0.953	0.689	0.664	0.643	0.794	0.731	0.712	0.723	0.97157	0.96536	0.7487	0.77971	0.82547	0.83895	0.858	0.840
S_p	0.955	0.642	0.519	0.421	0.811	0.652	0.724	0.586	0.97559	0.93427	0.51	0.60083	0.75122	0.80163	0.666	0.632
S_e	0.952	0.697	0.735	0.655	0.783	0.864	0.852	0.852	0.96942	0.98195	0.876	0.87511	0.88889	0.87083	0.965	0.955
ϵ_M	0.9665	0.9849	0.891	0.938	0.8447	0.8979	0.9541	0.9575	–	–	–	–	–	–	–	–
Time	1.6375±0.9358															
R %	51.6375±17.6951															
$Q - \widehat{\alpha}$																
CP	0.957	0.684	0.672	0.640	0.737	0.673	0.712	0.721	0.96699	0.95948	0.68957	0.67855	0.76779	0.79176	0.829	0.815
S_p	0.964	0.725	0.531	0.488	0.7	0.607	0.577	0.583	0.9615	0.91174	0.35333	0.32417	0.64715	0.68455	0.6108	0.594
S_e	0.954	0.68	0.743	0.7511	0.773	0.784	0.826	0.847	0.96992	0.98496	0.86889	0.86756	0.87083	0.88333	0.950	0.937
ϵ_M	0.9719	0.9791	0.9405	0.943	0.833	0.853	0.946	0.949	–	–	–	–	–	–	–	–
Time	0.0761±0.0048															
R %	48.0750±10.5792															
MSE ρ																
CP	0.915	0.594	0.671	0.645	0.656	0.656	0.749	0.493	0.962	0.95327	0.75101	0.7713	0.78951	0.82097	.881	0.8602
S_p	0.95	0.372	0.528	0.444	0.597	0.595	100	0	0.95023	0.90141	0.46667	0.57417	0.73496	0.7626	0.731	0.661
S_e	0.9	0.656	0.747	0.655	0.746	0.759	0.7188	0.579	0.96792	0.98095	0.90267	0.87644	0.83611	0.87083	0.964	0.9701
ϵ_M	0.9915	0.8126	0.8949	0.9725	0.8042	0.8102	0.9825	0.9631	–	–	–	–	–	–	–	–
Time	0.0837 ±0.0037															
R %	42.62±22.4201															
MSE $\widehat{\rho}$																
CP	0.91	0.65	0.605	0.583	0.694	0.69	0.709	0.709	0.96699	0.95784	0.74261	0.77942	0.82172	0.81873	0.855	0.815
S_p	0.978	0	0.448	0.387	0.644	0.642	0.579	0.571	0.95775	0.91362	0.47167	0.55917	0.73496	0.76585	0.691	0.616
S_e	0.885	0.65	0.731	0.6672	0.752	0.746	0.809	0.84	0.97193	0.98145	0.88711	0.89689	0.89583	0.86389	0.945	0.925
ϵ_M	0.9995	0.8402	0.9536	0.9928	0.9983	0.9983	0.9604	0.9647	–	–	–	–	–	–	–	–
Time	0.06± 0.0037															
R %	22.465±11.1647															

TABLE III: Classification performance applying feature selection.

	Breast		Diabetes		Heart		Ionosphere		Breast		Diabetes		Heart		Ionosphere	
	k-means	Spectral	k-means	Spectral	k-means	Spectral	k-means	Spectral	KNN	LCD	KNN	LCD	KNN	LCD	KNN	LCD
Without Weighting																
CP	0.958	0.685	0.715	0.645	0.832	0.768	0.707	0.744	0.96634	0.96209	0.74493	0.7687	0.83521	0.83371	0.861	0.872
S_p	0.957	0.594	0.574	0.432	0.885	0.695	0.572	0.597	0.95681	0.92958	0.44583	0.55333	0.76748	0.78862	0.655	0.672
S_e	0.958	0.706	0.82	0.655	0.799	0.87	0.818	0.909	0.97143	0.97945	0.90444	0.88356	0.89306	0.87222	0.975	0.983
ϵ_M	0.9428	0.9667	0.808	0.916	0.71	0.714	0.863	0.9	–	–	–	–	–	–	–	–
Time	–															
R%	–															
PCA pre-norm																
CP	0.965	0.668	0.648	0.647	0.825	0.744	0.692	0.447	0.971	0.946	0.738	0.713	0.828	0.838	0.892	0.862
S_p	0.942	0.667	0	0.444	0.89	0.674	0.553	0.014	0.984	0.931	0.495	0.482	0.743	0.770	0.785	0.659
S_e	0.977	0.668	0.65	0.6545	0.787	0.844	0.823	0.549	0.964	0.954	0.868	0.836	0.901	0.897	0.952	0.974
ϵ_M	0.965	0.9477	0.91	0.793	0.706	0.71	0.82	0.894	–	–	–	–	–	–	–	–
Time	0.2296±0.3177															
R%	–															
$Q - \alpha$																
CP	0.954	0.668	0.673	0.596	0.785	0.475	0.706	0.718	0.959	0.95	0.746	0.751	0.807	0.809	0.812	0.805
S_p	0.952	0.625	0.531	0.441	0.766	0.087	0.571	0.569	0.937	0.899	0.503	0.540	0.733	0.773	0.580	0.553
S_e	0.956	0.671	0.754	0.729	0.8	0.507	0.824	0.904	0.971	0.976	0.876	0.864	0.870	0.840	0.940	0.944
ϵ_M	0.9692	0.9562	0.876	0.896	0.883	0.96	0.952	0.9531	–	–	–	–	–	–	–	–
Time	1.7260±0.8691															
R%	58.8200±20.7033															
$Q - \widehat{\alpha}$																
CP	0.849	0.81	0.669	0.452	0.694	0.495	0.678	0.658	0.916	0.903	0.647	0.658	0.701	0.709	0.825	0.636
S_p	0.966	0.74	0.528	0.2177	0.644	0.441	0.547	0.571	0.822	0.745	0.299	0.245	0.523	0.590	0.625	0.185
S_e	0.818	0.844	0.739	0.579	0.752	0.527	0.762	0.67	0.965	0.988	0.832	0.879	0.854	0.811	0.935	0.885
ϵ_M	0.992	0.9795	0.989	0.9934	0.9772	0.9994	0.9691	0.9371	–	–	–	–	–	–	–	–
Time	0.0887±0.0092															
R%	17.9175±7.1383															
MSE ρ																
CP	0.944	0.703	0.641	0.642	0.73	0.734	0.712	0.49	0.951	0.95	0.736	0.768	0.756	0.772	0.874	0.862
S_p	0.967	0.634	0.488	0.41	0.702	0.73	0.586	0	0.916	0.885	0.461	0.565	0.656	0.715	0.697	0.648
S_e	0.934	0.719	0.757	0.654	0.756	0.737	0.8	0.577	0.969	0.984	0.883	0.877	0.841	0.820	0.972	0.9801
ϵ_M	0.9746	0.9553	0.853	0.963	0.754	0.756	0.868	0.945	–	–	–	–	–	–	–	–
Time	0.2647±0.3410															
R%	58.53±15.8678															
MSE $\widehat{\rho}$																
CP	0.906	0.693	0.635	0.642	0.751	0.535	0.715	0.718	0.921	0.9049	0.654	0.673	0.672	0.668	0.738	0.5948
S_p	0.978	0.611	0.481	0.182	0.748	0.496	0.582	0.576	0.802	0.738	0.195	0.21	0.465	0.534	0.409	0.1387
S_e	0.88	0.712	0.739	0.649	0.753	0.562	0.824	0.862	0.984	0.993	0.899	0.916	0.85	0.783	0.9204	0.846
ϵ_M	0.993	0.9679	0.919	0.999	0.9772	0.9994	0.933	0.94	–	–	–	–	–	–	–	–
Time	0.0733±0.0109															
R%	32.1575±12.9214															