Unsupervised relevance analysis for feature extraction and selection
A distance-based approach for feature relevance

Diego H. Peluffo [1], John A. Lee [1,2], Michel Verleysen [1], José L. Rodríguez [3] and Germán Castellanos-Domínguez [4]

[1] *Université catholique de Louvain, Machine Learning Group - ICTEAM, Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium*

[2] *Université catholique de Louvain, Molecular Imaging Radiotherapy and Oncology - IREC, Avenue Hippocrate 55, B-1200 Bruxelles, Belgium*

[3] *Grupo de automática, Universidad Autónoma de Manizales, Antigua estación del ferrocarril, Manizales, Colombia*

[4] *Signal Processing and Recognition Group, Universidad Nacional de Colombia, Manizales, Km 7 via al Magdalena, Manizales, Colombia*

{*diego.peluffo,john.lee, michel.verleysen*}*@uclouvain.be, jlrodriguez@autonoma.edu.co, cgcastellanosd@unal.edu.co*

Keywords: Feature extraction, Feature relevance, Feature selection, M-norm, PCA

Abstract: The aim of this paper is to propose a new generalized formulation for feature extraction based on distances from a feature relevance point of view. This is done within an unsupervised framework. To do so, it is first outlined the formal concept of feature relevance. Then, a novel feature extraction approach is introduced. Such an approach employs the M-norm as a distance measure. It is demonstrated that under some conditions, this method can readily explain literature methods. As another contribution of this paper, we propose an elegant feature ranking approach for feature selection followed from the spectral analysis of the data variability. Also, we provide a weighted PCA scheme revealing the relationship between feature extraction and feature selection. To assess the behavior of the studied methods within a pattern recognition system, a clustering stage is carried out. Normalized mutual information is used to quantify the quality of resultant clusters. Proposed methods reach comparable results with respect to literature methods.

1 Introduction

Feature extraction and selection methods are mostly aimed at obtaining a new representation space by reducing the dimension of input data following a certain criterion. The former transforms the input data into a lower dimensional data. The latter chooses variables from input data keeping those that best fulfill a selection criterion. Both methods are often associated to data mining tasks of classification or clustering (Cai et al., 2010). Then, they provide a lower dimensionality space while preserving the relevant discriminatory information. Among the feature extraction methods, those based on linear transformation - specially principal component analysis (PCA) - are probably the most popular. Even though many new complex methods for dimensionality reduction and data representation have been recently introduced (Lee and Verleysen, 2007), PCA and its variants still remain appealing and suitable techniques due to their non-parametric nature, and easiness for both implementation and outcome interpretation. For instance, some remarkable applications are change detection (Kuncheva and Faithfull, 2012), image segmentation (Zhang et al., 2010), and biomedical signal classification (Rodríguez-Sotelo et al., 2012). Also, PCA is very versatile allowing for weighted versions – WPCA (Wolf and Bileschi, 2005) and kernel extensions – KPCA (Liu Fan and Tong, 2012). Some recent approaches have been focused on new aspects of interest such as sparsity (Journée et al., 2010) and rank robustness (Candès et al., 2011).

This work outlines a formal definition of feature relevance within a distance-based framework. Following from this definition, a generalized feature extraction approach is introduced. Such an approach transforms linearly the data by taking advantage of the spectral information of a quadratic form in terms of the data matrix. Here, the relevance concept is referred to ranking features regarding a specific optimality criterion aiming to determine either a subset of features or establish how much

each feature contributes to optimizing such criterion. In practice, optimality criterion is set for improving the performance of a pattern recognition system. The method proposed here is a generalized distance-based feature extraction approach (GDFE), which employs the M-norm as a distance measure (Rassias, 1997). It is demonstrated that under some conditions, this method naturally yields conventional Euclidean-distance-based approaches (Principal Component Analysis - PCA) as well as a quadratic formulation ($Q-\alpha$ method (Wolf and Bileschi, 2005)). Also, as an important contribution of this paper, a relevance ranking approach based on the covariance matrix spectrum is presented. In addition, this work describes an approach to combine feature extraction with feature selection to yield improved weighted appraches (Weighted PCA - WPCA).

This paper is organized as follows: Section 2 outlines the basics of the definition of relevance based on distances. Section 3 introduces the generalized framework and a novel feature selection approach. Also, it shows the links with other methods. Finally, experimental results and conclusions are presented in Sections 4 and 5, respectively.

2 Definition of feature relevance

According to a specific criterion, relevance analysis (RA) distinguishes those features that best represent determined input data and/or most contribute to effectively discriminate among the disjoint data subsets into the whole input data set. Such features are named *relevant features*. Thus, RA also recognizes features whose representation or discrimination capability is lower, named *irrelevant features*; as well as those having repeated information (*redundant features*). In general, relevant variables are determined as those having the maximal ranking values. In the following it is outlined the formal definition of relevance and how it can be estimated by a distance based approach. Let us consider a data matrix $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ comprising N data points or samples described by a d-dimensional feature set such that $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_N^\top]^\top = [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_d]$, where $\boldsymbol{\xi}_\ell \in \mathbb{R}^N$ is the column ℓ of data matrix representing the ℓ-th feature, and $\boldsymbol{x}_i$ is the row i standing for the i-th data point. Then, for each one of the features a relevance function g can be defined as follows: $g : \mathbb{R}^{N\times d} \times \mathbb{R}^N \to \mathbb{R}^+ \quad (\boldsymbol{X}, \boldsymbol{\xi}_\ell) \mapsto g(\boldsymbol{X}, \boldsymbol{\xi}_\ell)$, such that the value of $g(\boldsymbol{X}, \xi_\ell)$ increases when the relevance of ℓ-th variable is greater, otherwise it should decrease. Notice that function g matches each variable $\boldsymbol{\xi}_\ell$ with an unique non-negative value $g(\boldsymbol{X}, \boldsymbol{\xi}_\ell)$, here called relevance or ranking value. The ℓ-th relevance value has a significant meaning to determine if the corresponding feature is relevant for either representation or classification purposes. To qualify g as a relevance function, once a proper criterion is established, the following axioms must be satisfied (Sepúlveda-Cano et al., 2011):

- *Nonnegativity*, i.e. $g(\boldsymbol{X}, \boldsymbol{\xi}_\ell) \geq 0$, $\forall \ell \in [1, d]$.
- *Nullity*, the function $g(\boldsymbol{X}, \boldsymbol{\xi}_\ell)$ is zero if the feature $\boldsymbol{\xi}_\ell$ is not relevant at all.
- Non-redundancy, if $\boldsymbol{\xi}_\ell = a\boldsymbol{\xi}_l + b\mathbf{1}_d$, where $a, b \in \mathbb{R}$, $\forall \ell \neq l$, and $\mathbf{1}_d$ is a d-dimensional all ones vector, then $|g(\boldsymbol{X}, \boldsymbol{\xi}_\ell) - g(\boldsymbol{X}, \boldsymbol{\xi}_l)| \to 0$.

Here, a relevance analysis via a dissimilarity-based cost function is presented. Suppose that a low-rank representation $\widehat{\boldsymbol{X}} \in \mathbb{R}^{N\times d}$ of data matrix $\boldsymbol{X}$ is known, such that $\text{rank}(\widehat{\boldsymbol{X}}) = p$ being $p < d$. Note that this assumption is applicable when the number of samples is less than that of features ($N < d$) and $\boldsymbol{X}$ is a full-rank matrix ($\text{rank}(\boldsymbol{X}) = d$). Matrix $\widehat{\boldsymbol{X}}$ must be formed in such a way that the effect of the most relevant features is captured while some dissimilarity measure f between $\boldsymbol{X}$ and $\widehat{\boldsymbol{X}}$ is minimized.

3 Feature extraction and selection

Assume the feature or data matrix $\boldsymbol{X}$ being centered, that is to say with zero mean regarding its columns. This can be readily done by $\boldsymbol{X} \leftarrow \boldsymbol{X} - (1/N)\mathbf{1}_N\mathbf{1}_N^\top\boldsymbol{X}$. Then, we can extract features by means of a linear combination with a d-dimensional base arranged in a rotation matrix projection matrix $\boldsymbol{V} \in \mathbb{R}^{d\times d}$, such that $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d]$ and $\boldsymbol{v}_\ell \in \mathbb{R}^d$ represents the ℓ-th column. To guarantee the linear independency of column vectors, matrix $\boldsymbol{V}$ is assumed to be orthonormal; which implies that $||\boldsymbol{v}_\ell||_2 = 1$ for all ℓ, as well as $\boldsymbol{v}_\ell^\top\boldsymbol{v}_k = 0$ for $\ell \neq k$. Accordingly, projected data matrix $\boldsymbol{Y} \in \mathbb{R}^{N\times d}$ can be calculated as: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{V}$.

Dimensionality reduction over the new representation space: Generally, the projection is performed over a lower dimensional space, which means that data are projected with a low-rank representation of rotation matrix $\widehat{\boldsymbol{V}} \in \mathbb{R}^{d\times p}$ being $p < d$. Therefore, a truncated projected data matrix $\widehat{\boldsymbol{Y}} \in \mathbb{R}^{N\times p}$ can be written as: $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{V}}$. Likewise, a lower-rank data matrix $\widehat{\boldsymbol{X}} \in \mathbb{R}^{N\times d}$ can be obtained by reconstructing the data matrix using $\widehat{\boldsymbol{V}}$ instead of the whole orthonormal base. Then, we can also write that $\widehat{\boldsymbol{X}} = \widehat{\boldsymbol{Y}}\widehat{\boldsymbol{V}}^\top = \boldsymbol{X}\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top$. As

mentioned above, to quantify how accurately $\widehat{\boldsymbol{X}}$ represent $\boldsymbol{X}$, a distance-based error function $\mathrm{d}(\boldsymbol{X},\widehat{\boldsymbol{X}})$ is used.

3.1 Generalized-distance-based feature extraction (GDFE)

To quantify the dissimilarity between $\boldsymbol{X}$ and $\widehat{\boldsymbol{X}}$, we propose a generalized M-inner norm regarding to any positive semi-definite matrix $\boldsymbol{\Omega} \in \mathbb{R}^{N\times N}$ noted as $\mathrm{d}(\boldsymbol{X},\widehat{\boldsymbol{X}}) = ||\boldsymbol{X}-\widehat{\boldsymbol{X}}||^2_{\boldsymbol{\Omega}}$. For easiness, a squared version is considered. Then, aiming to determine the best truncated representation, we can pose the following optimization problem:

$$\min_{\widehat{\boldsymbol{V}}} ||\boldsymbol{X}-\widehat{\boldsymbol{X}}||^2_{\boldsymbol{\Omega}} \quad \text{s. t.} \quad \widehat{\boldsymbol{V}}^\top\widehat{\boldsymbol{V}} = \boldsymbol{I}_p, \tag{1}$$

where $\boldsymbol{I}_p$ denotes a p-dimensional identity matrix.

Theorem 3.1. ***(Optimal low-rank representation)*** *A feasible optimal solution of the problem*

$$\min_{\widehat{\boldsymbol{V}}} \|\boldsymbol{X}-\widehat{\boldsymbol{X}}\|^2_{\boldsymbol{\Omega}} \quad \textit{s. t.} \quad \widehat{\boldsymbol{V}}^\top\widehat{\boldsymbol{V}} = \boldsymbol{I}_p$$

is selecting $\widehat{\boldsymbol{V}}$ as the p largest eigenvectors of $\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}$.

Proof 3.1. *From the inner product definition and applying some trace properties, we can write a dual formulation as*

$$\max_{\widehat{\boldsymbol{V}}} \ \mathrm{tr}(\widehat{\boldsymbol{V}}^\top\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}\widehat{\boldsymbol{V}}) \quad \textit{s. t.} \quad \widehat{\boldsymbol{V}}^\top\widehat{\boldsymbol{V}} = \boldsymbol{I}_p. \tag{2}$$

Previous quadratic formulation can be readily solved by an eigenvector decomposition. □

Low-dimension parameter p can be set by a explained variance criterion so as to capture the n% of explained variance. Let $\widehat{\boldsymbol{\lambda}}$ be the normalized eigenvalues vector of matrix $\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}$ such that $\widehat{\lambda}_\ell = \lambda_\ell / \sum_{\ell=1}^{d}\lambda_\ell$. Then, value p is chosen as that satisfying the condition $\sum_{\ell=1}^{p}\widehat{\lambda}_\ell \approx n/100$.

According to the previous statements, feature extraction can be done by a linear projection $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{V}$ where $\boldsymbol{V}$ are the eigenvectors of $\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}$. Due that $\boldsymbol{V}$ arises from a maximization problem, its column vectors must be arranged in a decreasing order regarding the eigenvalues $\boldsymbol{\lambda}$, so if $\boldsymbol{V} = [\boldsymbol{v}_1,\ldots,\boldsymbol{v}_d]$, $\boldsymbol{\lambda} = [\lambda_1,\ldots,\lambda_d]$ and λ_ℓ corresponds to $\boldsymbol{v}_\ell$, then $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. Likewise, the dimensionality reduction process over the projected space is carried out with $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{V}}$ where $\widehat{\boldsymbol{V}}$ is formed by the first p columns of $\boldsymbol{V}$.

Euclidean norm-based Approach: Assuming the particular case $\boldsymbol{\Omega} = \boldsymbol{I}_n$, the original problem expressed in (1) is reduced to be:

$$\min_{\widehat{\boldsymbol{V}}} ||\boldsymbol{X}-\widehat{\boldsymbol{X}}||^2_2 \quad \text{s. t.} \quad \widehat{\boldsymbol{V}}^\top\widehat{\boldsymbol{V}} = \boldsymbol{I}_p,$$

where $||\cdot||_2$ represents the Euclidean norm. In fact, when $\boldsymbol{\Omega} = \boldsymbol{I}_N$, term $||\boldsymbol{X}-\widehat{\boldsymbol{X}}||^2_{\boldsymbol{\Omega}}$ becomes

$$||\boldsymbol{X}-\widehat{\boldsymbol{X}}||^2_{\boldsymbol{I}_N} = \ \mathrm{tr}\Big((\boldsymbol{X}-\widehat{\boldsymbol{X}})^\top(\boldsymbol{X}-\widehat{\boldsymbol{X}})\Big) = ||\boldsymbol{X}-\widehat{\boldsymbol{X}}||^2_2.$$

Then, the dual problem can be written as $\max_{\widehat{\boldsymbol{V}}} \mathrm{tr}(\widehat{\boldsymbol{V}}^\top\boldsymbol{X}^\top\boldsymbol{X}\widehat{\boldsymbol{V}}) \ \ \widehat{\boldsymbol{V}}^\top\widehat{\boldsymbol{V}} = \boldsymbol{I}_p$. Then, the feature extraction can be done by $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{V}}$ where $\widehat{\boldsymbol{V}}$ are the eigenvectors of $\boldsymbol{X}^\top\boldsymbol{X}$ (like in conventional principal component analysis – PCA).

Quadratic problem regarding the outer product: Another particular case arises when $\boldsymbol{\Omega} = \boldsymbol{X}\boldsymbol{X}^\top$. Because of the given conditions that affinity matrix must satisfy, $\boldsymbol{\Omega}$ can be chosen as the outer product between variables (inner product between data points, as well) so that $\boldsymbol{\Omega} = \boldsymbol{X}\boldsymbol{X}^\top$. In this case, $\boldsymbol{\Omega}$ as can be seen as a polynomial kernel. By replacing $\boldsymbol{\Omega}$ in equation (2), we have:

$$\mathrm{tr}(\widehat{\boldsymbol{V}}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}\widehat{\boldsymbol{V}}) = \ \mathrm{tr}(\boldsymbol{Q}^\top\boldsymbol{\Omega}\boldsymbol{\Omega}\boldsymbol{Q}) = \sum_{\ell=1}^{p}\lambda_\ell^2, \tag{3}$$

where $\boldsymbol{Q} \in \mathbb{R}^{N\times p}$ is an arbitrary orthonormal matrix.

3.2 Feature selection

Feature selection is aimed to determine a subset of features whose relevance value is greater than the remaining features. This is done according to a certain criterion, in this case the dissimilarity cost function established in (1).

PCA-based relevance analysis: To introduce the following theorem, we consider as a cost function the dissimilarity $||\boldsymbol{X}-\widehat{\boldsymbol{X}}||^2_2$ and establish as a goal determining how much each variable $\boldsymbol{\xi}_\ell$ contributes to minimize such cost function.

Theorem 3.2. ***(Ranking vector)*** *Let $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_1,\ldots,\lambda_d)$ and $\boldsymbol{V} = [\boldsymbol{v}_1,\ldots,\boldsymbol{v}_d]$ the eigendecomposition of $\boldsymbol{X}^\top\boldsymbol{X}$. The ranking of how much contribute each dimension of $\boldsymbol{X}$ to minimize $\|\boldsymbol{X}-\widehat{\boldsymbol{X}}\|^2_2$ when considering a p–dimensional orthonormal base can be calculated as:*

$$\boldsymbol{\eta} = \sum_{\ell=1}^{p}\lambda_\ell\boldsymbol{v}_\ell \circ \boldsymbol{v}_\ell, \tag{4}$$

being $\circ$ the Hadamard product $\boldsymbol{\eta} = [\eta_1,\ldots,\eta_d]^\top$ and η_ℓ the rank value for $\boldsymbol{\xi}_\ell$.

Proof 3.2. *Let us consider the singular value decomposition*

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\mathcal{V}^\top = \sum_{\ell=1}^{d} s_\ell\boldsymbol{u}_\ell\boldsymbol{v}_\ell^\top, \tag{5}$$

and $\boldsymbol{\xi}_\ell \in \mathbb{R}^N$ be the ℓ-th column of $\boldsymbol{X}$. Then, $\mathcal{V} = [\boldsymbol{v}_1 \cdots \boldsymbol{v}_d] \in \mathbb{R}^{d\times d}$ are the eigenvectors of $\boldsymbol{X}^\top\boldsymbol{X}$ as well as $\boldsymbol{U} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_N] \in \mathbb{R}^{N\times N}$ are the eigenvectors of $\boldsymbol{X}\boldsymbol{X}^\top$. Matrix $\boldsymbol{S} \in \mathbb{R}^{N\times d}$ is formed by the singular values $\{s_1,\ldots,s_d\}$. Eigenvectors must be normalized in such a way that $\|\boldsymbol{u}_i\| = 1$ and $\|\boldsymbol{v}_\ell\| = 1$ for all $\ell \in \{1,\ldots,d\}$ and $i \in \{1,\ldots,N\}$ in order to guarantee that the base becomes orthonormal. Consider any vector $\boldsymbol{\xi}_\ell$ and its lower-rank representation $\widehat{\boldsymbol{\xi}}_\ell$, which can be expressed as a linear combinations as follows: $\boldsymbol{\xi}_\ell = \sum_{i=1}^{d} c_{i\ell}\boldsymbol{u}_i$ and $\widehat{\boldsymbol{\xi}}_\ell = \sum_{i=1}^{p} c_{i\ell}\boldsymbol{u}_i$, where $c_{i\ell}$ is the i-th coefficient associated to ℓ-th column and $p < d < N$. Afterwards, since $\|\boldsymbol{\xi}_\ell - \widehat{\boldsymbol{\xi}}_\ell\|_2^2 = \sum_{i=p+1}^{d} c_{i\ell}^2$, and $\|\boldsymbol{\xi}_\ell\|_2^2 = \sum_{\ell=1}^{d} c_{i\ell}^2 = \|\boldsymbol{\xi}_\ell - \widehat{\boldsymbol{\xi}}_\ell\|_2^2 + \sum_{i=1}^{p} c_{i\ell}^2$, minimizing $\|\boldsymbol{\xi}_\ell - \widehat{\boldsymbol{\xi}}_\ell\|_2^2$ is the same as maximizing $\sum_{i=1}^{p} c_{\ell t}^2$. Then, recalling equation (5)*, we have that*

$$\mathrm{tr}(\widehat{\boldsymbol{V}}^\top\boldsymbol{\Omega}\widehat{\boldsymbol{V}}) = \sum_{i=1}^{p} s_i^2 \ \mathrm{tr}(\boldsymbol{v}_i\boldsymbol{v}_i^\top) = \sum_{i=1}^{p} \lambda_i \boldsymbol{v}_i^\top\boldsymbol{v}_i.$$

Thus, we can infer that the contribution of $\boldsymbol{\xi}_\ell$ to either maximize the quadratic form $\mathrm{tr}(\widehat{\boldsymbol{U}}^\top\boldsymbol{\Omega}\widehat{\boldsymbol{U}})$ or minimize $\|\boldsymbol{X} - \widehat{\boldsymbol{X}}\|_2^2$ is given by $\eta_\ell = \sum_{i=1}^{p} s_i^2 {v_{\ell i}}^2 = \sum_{i=1}^{p} \lambda_i v_{\ell i}^2$, where $v_{\ell i}$ is the entry ℓ of vector $\boldsymbol{v}_i$. Finally, gathering all the d dimensions in vector $\boldsymbol{\eta}$, the ranking vector becomes

$$\boldsymbol{\eta} = \sum_{\ell=1}^{p} \lambda_\ell \boldsymbol{v}_\ell \circ \boldsymbol{v}_\ell. \quad \square \tag{6}$$

Since the first principal components are to capture the most explained variance, η can be approximated by $\widehat{\boldsymbol{\eta}} = \lambda_1\boldsymbol{v}_1 \circ \boldsymbol{v}_1$, where λ_1 is the largest eigenvalue.

Quadratic formulation for feature selection: Recalling section 3.1 and redefining $\boldsymbol{\Omega}$ as $\boldsymbol{\Omega}_{\boldsymbol{\alpha}} = \sum_{\ell=1}^{d} \boldsymbol{\alpha}_\ell\boldsymbol{\xi}_\ell\boldsymbol{\xi}_\ell^\top = \boldsymbol{X}\ \mathrm{diag}(\boldsymbol{\alpha})\boldsymbol{X}^\top$, arises another interesting approach. This method is the so-called $Q-\alpha$ (Wolf and Bileschi, 2005). In order to satisfy the conditions given by equation (3), it is necessary that $\mathrm{tr}(\boldsymbol{\Omega}_{\boldsymbol{\alpha}}\boldsymbol{\Omega}_{\boldsymbol{\alpha}}) = \sum_{\ell=1}^{d} \lambda_\ell^2$, and therefore $\boldsymbol{\alpha}$ must be unit: $||\boldsymbol{\alpha}||_2^2 = \boldsymbol{\alpha}^\top\boldsymbol{\alpha} = 1$. Then, $Q-\alpha$ objective function can be written as:

$$\max_{\boldsymbol{Q},\,\boldsymbol{\alpha}} \mathrm{tr}(\boldsymbol{Q}^\top\boldsymbol{\Omega}_{\boldsymbol{\alpha}}\boldsymbol{\Omega}_{\boldsymbol{\alpha}}\boldsymbol{Q}) = \sum_{\ell=1}^{d} \lambda_\ell^2 \text{ s. t. } \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_p. \tag{7}$$

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

$$\max_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^\top\boldsymbol{G}\boldsymbol{\alpha} \quad \text{s. t.} \quad \boldsymbol{\alpha}^\top\boldsymbol{\alpha} = 1,$$

where $\boldsymbol{G} \in \mathbb{R}^{d\times d}$ is a matrix with entries $G_{k\ell} = (\boldsymbol{\xi}_k\ \boldsymbol{\xi}_\ell^\top)\ \boldsymbol{\xi}_k\ \boldsymbol{Q}\ \boldsymbol{Q}^\top\ \boldsymbol{\xi}_\ell^\top$, for $k,\ell \in \{1,\ldots,d\}$. Since 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 $\boldsymbol{\alpha}$. As a consequence, the previous equation becomes the objective function to be used in the unsupervised version of $Q-\alpha$ as described in (Wolf and Bileschi, 2005). The time calculation when computing the vector $\boldsymbol{\alpha}$ can be reduced just to one iteration with no significant decrease of accuracy (Wolf and Bileschi, 2005). To this end, the feature relevance may be preserved optimizing the d original variables or the first p variables. Indeed, maximizing $\mathrm{tr}(\boldsymbol{Q}^\top\boldsymbol{\Omega}_{\boldsymbol{\alpha}}\boldsymbol{\Omega}_{\boldsymbol{\alpha}}\boldsymbol{Q})$ is equivalent to maximize $\mathrm{tr}(\boldsymbol{\Omega}_{\boldsymbol{\alpha}}\boldsymbol{\Omega}_{\boldsymbol{\alpha}}) = \mathrm{tr}(\boldsymbol{X}\ \mathrm{Diag}(\boldsymbol{\alpha})\boldsymbol{X}^\top\boldsymbol{X}\ \mathrm{diag}(\boldsymbol{\alpha})\boldsymbol{X}^\top)$. Since this expression is bilinear regarding $\boldsymbol{\alpha}$, the objective function can be re-written as $\boldsymbol{\alpha}^\top\boldsymbol{H}\boldsymbol{\alpha}$, where $H_{k\ell} = \mathrm{tr}(\boldsymbol{x}_k^\top\boldsymbol{x}_k\boldsymbol{x}_\ell^\top\boldsymbol{x}_\ell) = \boldsymbol{x}_k\boldsymbol{x}_\ell^\top\ \mathrm{tr}(\boldsymbol{x}_k^\top\boldsymbol{x}_\ell) = (\boldsymbol{x}_k\boldsymbol{x}_\ell^\top)^2$. Accordingly, it can be inferred that the approximate vector of relevance $\widehat{\boldsymbol{\alpha}}$ is the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{H} = (\boldsymbol{X}^\top\boldsymbol{X}) \circ (\boldsymbol{X}^\top\boldsymbol{X})$.

3.3 Combining feature extraction and selection

Feature extraction can be enhanced using the relevance vectors as weighting factors as described in (Wolf and Bileschi, 2005). For instance, according to the above discussed, weighting vector can be chosen as the ranking or relevance vectors, namely $\boldsymbol{\eta}$, $\widehat{\boldsymbol{\eta}}$, $\boldsymbol{\alpha}$ and $\widehat{\boldsymbol{\alpha}}$. Let $\boldsymbol{w} \in \mathbb{R}^d$ a weighting vector, then the weighting matrix $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ is $\boldsymbol{W} = \mathrm{Diag}(\boldsymbol{w})$. Therefore, feature can be extracted by the modified projection: $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{W}\widehat{\boldsymbol{V}}$. To keep the the same spectral properties, we should ensure $\boldsymbol{W}$ to be orthonormal ($\boldsymbol{W}^2 = \boldsymbol{I}_d$). By normalizing weighting vector in such a way that $||\boldsymbol{w}||_2 = 1$, matrix $\boldsymbol{W}$ becomes orthonormal. In other words, we can write the functional of problem stated in (2) as $\mathrm{tr}(\widetilde{\boldsymbol{V}}^\top\boldsymbol{W}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{W}\widetilde{\boldsymbol{V}}) = \sum_{i=1}^{p} \lambda_i$, when $\boldsymbol{\Omega} = \boldsymbol{I}_p$.

Therefore, if assuming a weighted data matrix $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{N\times d}$ such that $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{W}$, matrix $\widetilde{\boldsymbol{V}}$ corresponds to the first p eigenvectors of $\widetilde{\boldsymbol{X}}^\top\widetilde{\boldsymbol{X}}$ arranged decreasingly regarding the corresponding eigenvalues. In conclusion, relevance vector can be used to extract features within a weighted PCA (WPCA) framework. From another point of view, considering the rotated eigenvectors $\widehat{\boldsymbol{V}} = \boldsymbol{W}\widetilde{\boldsymbol{V}}$, the projection can be done by $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}\widehat{\boldsymbol{V}}$. Doing so, the weighted principal components can be calculated directly from the eigen-space of the original data.

4 Results and discussion

For experiments, we use some numeric data sets from the UCI data repository (Bache and Lichman, 2013) as shown in Table 1. To test how the studied methods behave within a pattern recognition system, the representation spaces reached by each method are clustered by a standard grouping algorithm. Namely, K-means with centers initialized randomly and setting the number of clusters as that of classes c from each database. Clustering performance is assessed by a normalized-mutual-information-based index (NMI) as described in (Strehl and Ghosh, 2002), which is bounded within the interval $[0,1]$, reaching 1 when input cluster/class assignments are identical. Here, NMI is applied to compare the original class labels with each resulting cluster assignment. The clustering procedure is iterated 20 times for every representation spaces resulting from the studied methods as well as the row data, and then mean and standard deviation values of NMI are collected. Likewise, the explained variance parameter is set to be $n = 95\%$ for all the feature extraction methods. The same percentage is used to choose the relevance features by $Q-\alpha$, so that features contributing to the 95% of the area under the curve of α plotting are picked, once it is squared and decreasingly ordered. To perform $Q-\alpha$, we use the power embedded algorithm as (Wolf and Bileschi, 2005)setting the maximum number of iterations to 4. In (Wolf and Bileschi, 2005), author demonstrate that 4 iterations are enough to reach convergence. Tables 1 and 2 show respectively the performance reached by the feature extraction and selection methods. For the sake of shorthand notation, approach selecting variables using theorem 3.2 is denoted as *relPCA*, as well as the quadratic formulation as $Q-\alpha$ and its approximated version as $Q-\alpha$ App. Likewise, approaches extracting features by the WPCA scheme are named so: when $\boldsymbol{w} = \mathbf{1}_d$ as *PCA*, $\boldsymbol{w} = \boldsymbol{\eta}$ as *WPCA*($\boldsymbol{\eta}$), $\boldsymbol{w} = \boldsymbol{\alpha}$ as *WPCA*($\boldsymbol{\alpha}$), and $\boldsymbol{w} = \widehat{\boldsymbol{\alpha}}$ as *WPCA*($\widehat{\boldsymbol{\alpha}}$). Proposed approach GDFE is applied by setting $\Omega_{ij} = \exp(-0.5(\boldsymbol{x}_i - \boldsymbol{x}_j)/\sigma^2)$, where σ is chosen as the maximum one among the 10% of the minimum Euclidean distances between data points from $\boldsymbol{X}$. Term p stands for either the number of relevant features or the lower dimension parameter.

From the results, it can be noticed that either extracting or selecting features regarding the analysis of variability are able to improve the clustering performance. Since the clustering method is a simple one (clusters are naturally formed by a notion of similarity given by the Euclidean distance), the performance quality can be mostly attributed to the resultant representation space accomplished by the feature extraction/selection methods. The reference values for comparatively analyzing the considered methods are in the third and fifth columns from Table 1 wherein the values of dimensions d and NMI of the original input data are respectively shown. That said, the performance of feature selection methods should not only be measured by NMI but also by taking into consideration the number of selected features p. Those databases having at some extent classes linearly separability reach higher NMI, for instance, `Iris` and `Twonorm`. Nonetheless, feature selection may outperform other procedures when more more complex data are analyzed (less compactness in this case). It is worth mentioning that this occurs when the intrinsic data information is separable and concentrated in some dimensions (the relevant ones). PCA-based relevance approach (relPCA) outperforms conventional PCA and shows comparable results with $Q-\alpha$. Then, we can claim that variability measured directly from the spectrum of a generalized covariance matrix might provide suitable information to achieve new representation spaces when clusters are somewhat compact.

Regarding proposed GDFE, it is noticeable that this approach outperforms the remaining considered methods reaching better performance -or at least the same as that reached when using the row data- while using a low dimensional representation. Indeed, when choosing $\boldsymbol{\Omega}$ as any similarity matrix, not only the symmetric positive condition is fulfilled but an additional interesting property is also reached. Notice that the quadratic term $\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X}$ yields a projection matrix $\widehat{\boldsymbol{V}}$ involving simultaneously the variance (which can be seen as a global measure), and similarities capturing localized information. Then, a more flexible and versatile data projection is reached, which is able to deal with slightly complex data while fulfilled a certain variance or compactness criterion.

5 Conclusions and future work

This work presents a formal definition of feature relevance within an unsupervised framework based on distances. This framework yields a novel generalized feature extraction problem that linearly transforms data keeping improved components that most contribute to the variance, while analyzing simultaneously the local data information through similarities.

The generalized formulation allows to explain other approaches such as PCA and $Q-\alpha$. Also, based on the spectrum of a generalized covariance matrix, a new approach to estimate the feature relevance is introduced. We explain how to link feature extraction and feature selection, as well as, how to lead new

#	Data base	Row data				relPCA		$Q-\alpha$		$Q-\alpha$ App.	
		NMI	N	d	c	NMI	p	NMI	p	NMI	p
1	Iris	0.73±0.07	150	4	3	0.71±0.05	2	0.57±0.06	1	0.58±0.04	1
2	80x	0.64±0.06	45	8	3	0.65±0.06	6	0.38±0.07	5	0.38±0.07	5
3	Malaysia	0.38±0.01	291	8	20	0.47±0.02	4	0.57±0.01	1	0.57±0.01	1
4	Breast	0.75±0.00	683	9	2	0.75±0.00	7	0.77±0.00	6	0.77±0.00	6
5	Chromo	0.41±0.01	1143	8	24	0.42±0.01	6	0.43±0.01	5	0.42±0.01	5
6	Satellite	0.56±0.06	6435	36	6	0.57±0.05	32	0.61±0.02	5	0.60±0.02	5
7	Soybean1	0.68±0.03	266	35	15	0.69±0.02	18	0.65±0.02	15	0.65±0.01	15
8	Twonorm	0.85±0.00	7400	20	2	0.85±0.00	18	0.81±0.00	18	0.81±0.00	18

Table 1: Feature selection results. Notation N, d and c denotes the number of data points, the number of dimensions and the number of class, respectively. Term p is the number of relevant features.

#	PCA		WPCA(η)		GDFE		WPCA(α)		WPCA($\widehat{\alpha}$)	
	NMI	p	NMI	p	NMI	p	NMI	p	NMI	p
1	0.79±0.00	1	0.82±0.02	4	0.85±0.00	1	0.80±0.00	4	0.79±0.00	4
2	0.66±0.06	5	0.72±0.08	3	0.74±0.07	5	0.67±0.06	4	0.67±0.06	4
3	0.42±0.01	1	0.41±0.01	8	0.42±0.01	1	0.46±0.01	2	0.46±0.01	2
4	0.75±0.00	6	0.79±0.00	4	0.75±0.00	4	0.76±0.00	5	0.75±0.00	5
5	0.41±0.01	5	0.45±0.01	3	0.51±0.01	5	0.47±0.01	3	0.47±0.01	3
6	0.58±0.05	5	0.61±0.02	4	0.65±0.05	5	0.62±0.06	5	0.61±0.07	5
7	0.66±0.01	15	0.68±0.01	4	0.75±0.02	15	0.71±0.03	5	0.72±0.02	5
8	0.85±0	18	0.83±0.00	17	0.87±0.00	18	0.86±0.00	18	0.85±0.00	18

Table 2: Feature extraction results. Here, p denotes the number of considered components of the orthonormal base.

WPCA alternatives via relevance analysis.

As a future work, new distances as well as other similarity matrices are to be explored aiming at the design of a non-supervised system with optimal representation stages.

6 Acknowledgments

J.A. Lee is a Research Associate with the FRS-FNRS (Belgian National Scientific Research Fund). This work has been partially funded by FRS-FNRS project 7.0175.13 DRedVis.

REFERENCES

Bache, K. and Lichman, M. (2013). UCI machine learning repository.

Cai, D., Zhang, C., and He, X. (2010). Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 333–342. ACM.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11.

Journée, M., Nesterov, Y., Richtárik, P., and Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research*, 11:517–553.

Kuncheva, L. I. and Faithfull, W. J. (2012). Pca feature extraction for change detection in multidimensional unlabelled streaming data. In *21st International Conference on Pattern Recognition (ICPR)*, pages 1140–1143. IEEE.

Lee, J. A. and Verleysen, M. (2007). *Nonlinear dimensionality reduction*. Springer.

Liu Fan, X. T. and Tong, S. (2012). Kernel pca and nonlinear asm. *ICCS 2012*, 138:287.

Rassias, T. M. (1997). *Inner product spaces and applications*, volume 376. CRC Press.

Rodríguez-Sotelo, J. L., Peluffo-Ordoñez, D., Cuesta-Frau, D., and Castellanos-Domínguez, G. (2012). Unsupervised feature relevance analysis applied to improve ecg heartbeat clustering. *Computer methods and programs in biomedicine*, 108(1):250–261.

Sepúlveda-Cano, L. M., Acosta-Medina, C. D., and Castellanos-Domínguez, G. (2011). Relevance analysis of stochastic biosignals for identification of pathologies. *EURASIP Journal on Advances in Signal Processing*, 2011:3.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.

Wolf, L. and Bileschi, S. (2005). Combining variable selection with dimensionality reduction. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 801–806. IEEE.

Zhang, L., Dong, W., Zhang, D., and Shi, G. (2010). Two-stage image denoising by principal component analysis with local pixel grouping. *Pattern Recognition*, 43(4):1531–1549.