An eigenanalysis of data centering in machine learning

Paul Honeine

An eigenanalysis of data centering in machine learning

Paul Honeine, *Member, IEEE*

Abstract—Many pattern recognition methods rely on statistical information from centered data, with the eigenanalysis of an empirical central moment, such as the covariance matrix in principal component analysis (PCA), as well as partial least squares regression, canonical-correlation analysis and Fisher discriminant analysis. Recently, many researchers advocate working on non-centered data. This is the case for instance with the singular value decomposition approach, with the (kernel) entropy component analysis, with the information-theoretic learning framework, and even with nonnegative matrix factorization. Moreover, one can also consider a non-centered PCA by using the second-order non-central moment.

The main purpose of this paper is to bridge the gap between these two viewpoints in designing machine learning methods. To provide a study at the cornerstone of kernel-based machines, we conduct an eigenanalysis of the inner product matrices from centered and non-centered data. We derive several results connecting their eigenvalues and their eigenvectors. Furthermore, we explore the outer product matrices, by providing several results connecting the largest eigenvectors of the covariance matrix and its non-centered counterpart. These results lay the groundwork to several extensions beyond conventional centering, with the weighted mean shift, the rank-one update, and the multidimensional scaling. Experiments conducted on simulated and real data illustrate the relevance of this work.

Index Terms—Kernel-based methods, Gram matrix, machine learning, pattern recognition, principal component analysis, kernel entropy component analysis, centering data.

1 Introduction

Most pattern recognition in machine learning can be explained by performing an eigenanalysis, with an eigendecomposition or a spectral decomposition (*i.e.*, singular value decomposition or SVD) [1]. These machines seek a set of relevant axes from a given dataset. The principal component analysis (PCA) [2] is the most prominent eigenanalysis problem for feature extraction and dimensionality reduction. In this case, the most relevant axes, obtained from the eigendecomposition of the covariance matrix, capture the largest amount of variance in the data. Other machines include multidimensional scaling [3], partial least squares regression (PLS) [4], canonical-correlation analysis (CCA) [5] and its classification-based version known as Fisher discriminant analysis (FDA) [6]. The latter two methods solve a generalized eigendecomposition problem.

Kernel-based machines provide an elegant framework to generalize these linear pattern recognition methods to the nonlinear domain. They rely on the concept of *kernel trick*, initially introduced by Aizerman *et al.* in [7]. The main breakthrough lies in two folds. On the one hand, most pattern recognition, classification and regression algorithms can be written in terms of inner products between data. On the other hand, by substituting each inner product by a (positive definite) kernel function, a nonlinear transformation is implicitly operated on the data without any significant computational cost. Therefore, the eigenanalysis as well as most of the operations are performed on the kernel matrix, which corresponds to an inner product "Gram" matrix in some feature space. This property is revealed in kernelized versions of PCA [8], PLS [9], CCA [10] and FDA [11]. See [12], [13], [14] for a survey of kernel-based machines.

In several kernel-based machines, as given for instance in PCA, CCA, and FDA, data should be centered in the feature space, by shifting the origin to the centroid of the data. From an algorithmic point of view, centering is performed easily with matrix algebra, either in batch mode by a subsequent column and row centering of the kernel matrix [8], or in a recursive way when dealing with online learning [15]. From a theoretical point of view, centering reveals central moments, *i.e.*, moments about the centroid/mean of the available data, as well as other related statistics. Well-known central moments include the second-order central moment, also called covariance, which is investigated by the PCA for estimating the maximum-variance directions. Furthermore, these directions minimize the reconstruction error.

Many researchers advocate the use of non-centered data in pattern recognition. Information is extracted directly, either with the data matrix and its Gram matrix, or with non-central moments. Several motivations were revealed in favor of working on non-centered data. The intuitive motivation is the application of the spectral decomposition without data-centering in many pattern recognition and machine learning problems, thus providing a sort of a non-centered PCA by using the second-order non-central moment. This is the case for instance in signal analysis and classification [16] and in designing dictionaries for sparse representation [17]. A key motivation towards keeping data non-centered is the

• *M. Honeine is with the Institut Charles Delaunay (CNRS), Université de Technologie de Troyes, Troyes, France.*

nonparametric density estimation with kernel functions [18], as revealed recently with exceptional performance in the (kernel) entropy component analysis (ECA) [19] and the information-theoretic learning framework [20]. See also [21]. A further motivation is that data often deviate from the origin, and the measure of such deviation may constitute an interesting feature, such as in hyperspectral unmixing [22]. It turns out that in many fields of computer science, signal processing and machine learning, acquired data are nonnegative, and even positive. This is the case with the study of gene expressions in bioinformatics [23], with eigenfaces in the computer vision problem of human face recognition [24], with online handwritten character recognition [25], and with numerous applications for nonnegative matrix factorization [26], [27]. As a consequence, nonnegative Gram matrices are pervasive. Such information is lost when centering the data, as well as several interesting properties[1].

The issue of centering the data versus keeping the data uncentering is an open question in pattern recognition: (Pearson) correlation versus congruence coefficients, (centered) PCA versus non-centered PCA (or SVD), covariance and centered Gram matrices versus their non-centered counterparts. In this paper, we study the impact of centering the data on the distribution of the eigenvalues and eigenvectors of both inner-product and outer-product matrices. By examining the Gram matrix and its centered counterpart, we show the interlacing property of their eigenvalues. We devise bounds connecting the eigenvalues of these two matrices, including a lower bound on the largest eigenvalue of the centered Gram matrix. Furthermore, we examine the eigenvectors of the inner product and outer product matrices. We provide connections between the most relevant eigenvector of the covariance matrix and that of the non-centered matrix, a result that corroborate the work in [30].

In our study, we focus on the eigenanalysis of the Gram matrices. This work opens the way to understanding the impact of centering the data in most kernel-based machine. This is shown by bridging the gap between the (centered) PCA and the (non-centered) ECA. Moreover, our work goes beyond the PCA and ECA, since it extends naturally to many kernel-based machines where the eigen-decomposition of the Gram matrix is crucial. To this end, we revisit the multidimensional scaling problem where centering is essential. Moreover, we provide extension beyond conventional mean-centering.

Related (and unrelated) work

In machine learning for classification and discrimination, the issue of centering the data has been addressed in few publications. In the Bayesian framework proposed in [31], the bias-free formulation of support vector machines (SVM) and least-squares SVM yields the eigendecomposition of the centered Gram matrix, while Gaussian processes yield similar expressions applied on the non-centered Gram matrix. In [32], the authors propose a modified FDA to take into account the fact that centering leads to a singular Gram matrix, even when the non-centered Gram matrix is non-singular. More recently, it is devised in [33] that data and label should be centered when dealing with the alignment criterion. In [34], the author consider the issue of optimizing the centering as well as the low-rank approximation problem.

In Bayesian statistics, the impact of centering has been extensively studied in multilevel models, when dealing with hierarchically nested models [35], [36]. In this case, centering is performed either with the grand mean, or "partially" with a group mean centering, at different levels [37], [38]. See [39, Chapter 5.2] for a comprehensive review on the centering issue in multilevel modeling. This problem is revisited within a Bayesian framework in [40], [41], with the issue of centered or non-centered parameterisations. This is beyond the scope of this paper, since we investigate non-parametric methods such as PCA and ECA.

To the best of our knowledge, only Cadima and Jolliffe investigated in [30] the relationship between the PCA and its non-centered variant. To this end, they confronted the eigendecomposition of the covariance matrix with the eigendecomposition of its non-centered counterpart. Such connection can be done thanks to the rank-one update that connects both matrices. In this paper, we study the eigendecompositions of the centered and the non-centered Gram matrices. It is easy to see that this is a much harder problem. By performing an eigenanalysis of the Gram matrices, we provide a framework that integrates the analysis of PCA, ECA, MDS, and beyond. This work opens the door to the study of centering or not in kernel-based machines.

Notation

The matrix $\mathbf{I}$ is the n-by-n identity matrix, and the vector $\mathbf{1}$ is the all-ones vector of n entries. Then, $\mathbf{1}\mathbf{1}^\top$ is the n-by-n matrix of all ones, and $\mathbf{1}^\top\mathbf{1} = n$. Moreover, $\mathbf{1}^\top\boldsymbol{M}\mathbf{1}$ is the grand sum of the matrix $\boldsymbol{M}$. The subscript $_\mathrm{c}$ allows to recognize the case of centered data, as opposed to the non-centered one: $\boldsymbol{X}_\mathrm{c}$ versus $\boldsymbol{X}$, $\boldsymbol{K}_\mathrm{c}$ versus $\boldsymbol{K}$, $\boldsymbol{C}_\mathrm{c}$ versus $\boldsymbol{C}$, $\lambda_{\mathrm{c}i}$ versus λ_i, etc.

The spectral decomposition (*i.e.*, singular value decomposition) of a matrix $\boldsymbol{M}$ defines its singular values $\sigma_1(\boldsymbol{M}), \sigma_2(\boldsymbol{M}), \ldots$, given in non-increasing order. The i-th largest eigenvalue of a square matrix $\boldsymbol{M}$ is denoted by $\lambda_i(\boldsymbol{M})$, and its trace by $\mathrm{tr}(\boldsymbol{M})$. A first analysis on the eigenvalues of a matrix is given by its trace, with:

$$\mathrm{tr}(\boldsymbol{M}) = \sum_i \lambda_i(\boldsymbol{M}). \tag{1}$$

This corresponds to the variance of the data when dealing with the data covariance matrix.

1. For instance, the Perron–Frobenius theorem [28], [29] states that, under some mild conditions, the non-negativity of a matrix is inherited by its unique largest eigenvalue and that the corresponding eigenvector has positive components. This result is no longer valid when the data are centered.

For the sake of clarity, the i-th largest eigenvalues of the inner product matrices $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$ are respectively λ_i and $\lambda_{\mathrm{c}i}$, namely $\lambda_i = \lambda_i(\boldsymbol{K})$ and $\lambda_{\mathrm{c}i} = \lambda_{\mathrm{c}i}(\boldsymbol{K}_{\mathrm{c}})$. The eigenpair $(\lambda_i, \boldsymbol{\alpha}_i)$ of $\boldsymbol{K}$ denotes its i-th largest eigenvalue λ_i and its corresponding eigenvector $\boldsymbol{\alpha}_i$. The *first* eigenvector is the one associated to the largest eigenvalue.

2 Introduction

In this section, we present the eigendecomposition problems of the inner product and outer product matrices, for both non-centered and centered data. Then, two kernel-based machines are presented, PCA and ECA, in order to contrast the paradigm of centering or not the data.

2.1 Non-centered data

Consider a set of n available samples, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, from a vector space of dimension d, with the conventional inner product $\boldsymbol{x}_i^\top \boldsymbol{x}_j$. Let $\boldsymbol{X} = [\boldsymbol{x}_1 \;\; \boldsymbol{x}_2 \;\; \cdots \;\; \boldsymbol{x}_n]$ be the data matrix, and $\boldsymbol{K} = \boldsymbol{X}^\top \boldsymbol{X}$ be the corresponding Gram matrix, *i.e.,* the inner product matrix. It turns out that this matrix encapsulates the essence of the information in the data, as illustrated with the kernel trick throughout the literature. It is therefore natural to focus on the Gram matrix in out study.

Let $(\lambda_i, \boldsymbol{\alpha}_i)$ be an eigenpair of the matrix $\boldsymbol{K}$, namely

$$\boldsymbol{K}\,\boldsymbol{\alpha}_i = \lambda_i\,\boldsymbol{\alpha}_i, \tag{2}$$

for $i = 1, 2, \ldots, n$, the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ being given in non-increasing order. These quantities describe the spectral decomposition of the matrix $\boldsymbol{K}$, with

$$\boldsymbol{K} = \boldsymbol{A}\,\boldsymbol{\Lambda}\,\boldsymbol{A}^\top. \tag{3}$$

The eigendecomposition of the Gram matrix $\boldsymbol{K}$ is related to the spectral decomposition of the data matrix $\boldsymbol{X}$, since the latter is given by

$$\boldsymbol{X} = \boldsymbol{W}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{A}^\top, \tag{4}$$

where $\boldsymbol{A}$ is the n-by-n matrix whose columns are the eigenvectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n$, and $\boldsymbol{\Lambda}^{\frac{1}{2}}$ is the d-by-n rectangular diagonal matrix whose i-th diagonal entry is

$$\sigma_i(\boldsymbol{X}) = \sqrt{\lambda_i}. \tag{5}$$

The d columns of $\boldsymbol{W}$ are called the left-singular vectors of $\boldsymbol{X}$. They define the spectral decomposition of the matrix $\boldsymbol{X}\boldsymbol{X}^\top$, which is known as the realized covariation matrix [42] in financial economics. For a coherent analysis, we consider in this paper the second-order *non-central* moment matrix $\boldsymbol{C} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\,\boldsymbol{x}_i^\top$, written in matrix form as $\boldsymbol{C} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$. Following the spectral decomposition of $\boldsymbol{X}$ in (4), we get

$$\boldsymbol{C}\,\boldsymbol{w}_i = \tfrac{1}{n}\lambda_i\,\boldsymbol{w}_i,$$

where $\boldsymbol{W} = [\boldsymbol{w}_1 \;\; \boldsymbol{w}_2 \;\; \cdots \;\; \boldsymbol{w}_d]$. The matrix $\boldsymbol{C}$ is less known than the second-order *central* moment matrix. The latter, called covariance matrix, is obtained after centering the data, as given in the following.

2.2 Centered data

Consider centering the data, $\boldsymbol{x}_{\mathrm{c}i} = \boldsymbol{x}_i - \boldsymbol{\mu}$ for $i = 1, 2, \ldots, n$, where $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ is the empirical mean (*i.e.,* centroid). We get in matrix form $\boldsymbol{X}_{\mathrm{c}} = \boldsymbol{X} - \boldsymbol{\mu}\boldsymbol{1}^\top$, where

$$\boldsymbol{\mu} = \tfrac{1}{n}\boldsymbol{X}\boldsymbol{1}.$$

From the spectral decomposition of the matrix $\boldsymbol{K}$ in (3), we can rewrite the norm of the mean as

$$\|\boldsymbol{\mu}\|^2 = \tfrac{1}{n^2}\boldsymbol{1}^\top\boldsymbol{K}\boldsymbol{1} = \tfrac{1}{n^2}\sum_{i=1}^n \lambda_i(\boldsymbol{\alpha}_i^\top\boldsymbol{1})^2, \tag{6}$$

which illustrates how each pair of eigenvalue and eigenvector contributes to the norm of the data mean. The impact of centering the data is clear on both the Gram and the covariance matrices, as illustrated next.

Let $\boldsymbol{K}_{\mathrm{c}} = \boldsymbol{X}_{\mathrm{c}}^\top\boldsymbol{X}_{\mathrm{c}}$ be the Gram matrix of the centered data, with entries $\boldsymbol{x}_{\mathrm{c}i}^\top\boldsymbol{x}_{\mathrm{c}j}$ for $i, j = 1, 2, \ldots, n$, namely

$$\boldsymbol{K}_{\mathrm{c}} = \boldsymbol{K} - \tfrac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top\boldsymbol{K} - \tfrac{1}{n}\boldsymbol{K}\boldsymbol{1}\boldsymbol{1}^\top + \tfrac{1}{n^2}\boldsymbol{1}\boldsymbol{1}^\top\boldsymbol{K}\boldsymbol{1}\boldsymbol{1}^\top, \tag{7}$$

Let $(\lambda_{\mathrm{c}i}, \boldsymbol{\alpha}_{\mathrm{c}i})$ be an eigenpair of this matrix, then:

$$\boldsymbol{K}_{\mathrm{c}}\,\boldsymbol{\alpha}_{\mathrm{c}i} = \lambda_{\mathrm{c}i}\,\boldsymbol{\alpha}_{\mathrm{c}i}. \tag{8}$$

Let $\boldsymbol{C}_{\mathrm{c}} = \frac{1}{n}\boldsymbol{X}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}^\top$ be the covariance matrix, namely the second-order central moment matrix defined by

$$\boldsymbol{C}_{\mathrm{c}} = \boldsymbol{C} - \boldsymbol{\mu}\boldsymbol{\mu}^\top. \tag{9}$$

The eigenpairs of this matrix define the eigenproblem

$$\boldsymbol{C}_{\mathrm{c}}\,\boldsymbol{w}_{\mathrm{c}i} = \tfrac{1}{n}\lambda_{\mathrm{c}i}\,\boldsymbol{w}_{\mathrm{c}i}. \tag{10}$$

Nonlinear extension using kernel functions

A symmetric kernel $\kappa(\cdot,\cdot)$ is called a positive definite kernel if it gives rise to a positive definite matrix, namely for all $n \in \mathbb{N}$ and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ we have

$$\boldsymbol{\beta}^\top\boldsymbol{K}\boldsymbol{\beta} \geq 0, \tag{11}$$

for all vectors $\boldsymbol{\beta}$, where the matrix $\boldsymbol{K}$ has entries $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Such kernel functions provide a nonlinear extension of the conventional inner product since, thanks to Mercer's theorem [43], [44], $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ corresponds to an inner product between transformed samples $\boldsymbol{x}_i^\phi$ and $\boldsymbol{x}_j^\phi$ in some feature space, namely

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = {\boldsymbol{x}_i^\phi}^\top\boldsymbol{x}_j^\phi. \tag{12}$$

Examples of nonlinear kernels include the polynomial kernel, of the form $(c + \boldsymbol{x}_i^\top\boldsymbol{x}_j)^p$, and the Gaussian kernel $\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$ where σ is the tunable bandwidth parameter.

It turns out that the derivations given in this paper can be easily generalized to kernel functions, since a kernel matrix corresponds to a Gram matrix in some feature space. Centering is performed in the feature space, with the centroid[2] $\boldsymbol{\mu}^\phi = \frac{1}{n}\boldsymbol{X}^\phi\boldsymbol{1}$. As a consequence, we get the same expression as in (6), with $\|\boldsymbol{\mu}^\phi\|^2 = \frac{1}{n^2}\boldsymbol{1}^\top\boldsymbol{K}\boldsymbol{1}$.

2. In most pattern recognition tasks, one does not need to quantify $\boldsymbol{\mu}^\phi$, but only its inner product with any $\boldsymbol{x}_i^\phi$. One can estimate its counterpart in the input space. This is the pre-image problem, which is clearly beyond the scope of this paper. See [45] for a recent survey.

2.3 The paradigm of centering or not the data

In order to confront centering the data with keeping the data non-centered, we present next two well-known machines for pattern recognition. On the one hand, advocating the use of central matrices, the PCA is presented in its two-folds, the conventional and the kernelized formulations. On the other hand, advocating the exploration of non-centered data, nonparametric density estimation using the ECA.

2.3.1 Case study 1: Principal component analysis

The PCA seeks the axes that capture most of the variance within the data [2]. It is well-known[3] that these axes are defined by the eigenvectors associated to the largest eigenvalues of the covariance matrix $\boldsymbol{C}_{\mathrm{c}}$, with

$$\boldsymbol{C}_{\mathrm{c}}\,\boldsymbol{w}_{\mathrm{c}i} = \tfrac{1}{n}\lambda_{\mathrm{c}i}\,\boldsymbol{w}_{\mathrm{c}i}.$$

Then, $\frac{1}{n}\lambda_{\mathrm{c}i}$ measures the variance along the axe $\boldsymbol{w}_{\mathrm{c}i}$, while the normalized i-th eigenvalue $\pi_{\mathrm{c}i} = \frac{\lambda_{\mathrm{c}i}}{\sum_j \lambda_{\mathrm{c}j}}$ accounts for the proportion of total variation. It is worth noting from (1) that the total variance of the data is given by the trace of $\boldsymbol{C}_{\mathrm{c}}$.

By substituting the definition of $\boldsymbol{C}_{\mathrm{c}}$ in the eigenproblem (10), we get

$$\boldsymbol{w}_{\mathrm{c}i} = \tfrac{n}{\lambda_{\mathrm{c}i}}\boldsymbol{C}_{\mathrm{c}}\,\boldsymbol{w}_{\mathrm{c}i} = \tfrac{1}{\lambda_{\mathrm{c}i}}\sum_{j=1}^{n}(\boldsymbol{x}_{\mathrm{c}j}^{\top}\boldsymbol{w}_{\mathrm{c}i})\,\boldsymbol{x}_{\mathrm{c}i}, \tag{13}$$

and therefore each $\boldsymbol{w}_{\mathrm{c}i}$ lies in the span of the data. This means that there exists a vector $\boldsymbol{\alpha}_{\mathrm{c}i}$ such that $\boldsymbol{w}_{\mathrm{c}i} = \boldsymbol{X}_{\mathrm{c}}\,\boldsymbol{\alpha}_{\mathrm{c}i}$. By injecting this relation in the above expression, we get $\boldsymbol{X}_{\mathrm{c}}\,\boldsymbol{\alpha}_{\mathrm{c}i} = \frac{1}{\lambda_{\mathrm{c}i}}\boldsymbol{X}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}^{\top}\boldsymbol{X}_{\mathrm{c}}\,\boldsymbol{\alpha}_{\mathrm{c}i}$. Equivalently, by multiplying each side by $\boldsymbol{X}_{\mathrm{c}}^{\top}$, we get $\boldsymbol{K}_{\mathrm{c}}\,\boldsymbol{\alpha}_{\mathrm{c}i} = \frac{1}{\lambda_{\mathrm{c}i}}\boldsymbol{K}_{\mathrm{c}}^{2}\boldsymbol{\alpha}_{\mathrm{c}i}$. Thus, we have the following eigenproblem

$$\boldsymbol{K}_{\mathrm{c}}\,\boldsymbol{\alpha}_{\mathrm{c}i} = \lambda_{\mathrm{c}i}\,\boldsymbol{\alpha}_{\mathrm{c}i}.$$

The eigenvectors of $\boldsymbol{K}_{\mathrm{c}}$ allow to identify the projection of any sample $\boldsymbol{x}$ onto the corresponding eigenvectors of $\boldsymbol{C}_{\mathrm{c}}$. Representing it on a set of eigenvectors is given by

$$\sum_{k}(\boldsymbol{x}^{\top}\boldsymbol{w}_{\mathrm{c}k})\,\boldsymbol{w}_{\mathrm{c}k} = \sum_{k}(\boldsymbol{x}^{\top}\boldsymbol{X}_{\mathrm{c}}\boldsymbol{\alpha}_{\mathrm{c}k})\boldsymbol{X}_{\mathrm{c}}\boldsymbol{\alpha}_{\mathrm{c}k}. \tag{14}$$

In order to satisfy the unit-norm of $\boldsymbol{w}_{\mathrm{c}i}$ for any i, the eigenvectors of $\boldsymbol{K}_{\mathrm{c}}$ should be normalized. Indeed, we have $\boldsymbol{w}_{\mathrm{c}i}^{\top}\boldsymbol{w}_{\mathrm{c}i} = \boldsymbol{\alpha}_{\mathrm{c}i}^{\top}\boldsymbol{X}_{\mathrm{c}}^{\top}\boldsymbol{X}_{\mathrm{c}}\boldsymbol{\alpha}_{\mathrm{c}i} = \boldsymbol{\alpha}_{\mathrm{c}i}^{\top}\boldsymbol{K}_{\mathrm{c}}\boldsymbol{\alpha}_{\mathrm{c}i} = \lambda_{\mathrm{c}i}\boldsymbol{\alpha}_{\mathrm{c}i}^{\top}\boldsymbol{\alpha}_{\mathrm{c}i}$. Therefore, one can define each feature $\boldsymbol{w}_{\mathrm{c}i}$ directly from the eigenvector $\boldsymbol{\alpha}_{\mathrm{c}i}$ of the Gram matrix $\boldsymbol{K}_{\mathrm{c}}$, after normalization such that

$$\|\boldsymbol{\alpha}_{\mathrm{c}i}\|^2 = \frac{1}{\lambda_{\mathrm{c}i}}. \tag{15}$$

This scaling allows to preserve the variance of the data along the respective axes [46]. When one needs to fix the scale along all these directions, the following normalization is considered:

$$\|\boldsymbol{\alpha}_{\mathrm{c}i}\|^2 = \frac{1}{\lambda_{\mathrm{c}i}^{\,2}}, \tag{16}$$

This scaling allows to set a unit variance within projected data on each axe [47].

3. By writing these vectors in a matrix $\boldsymbol{W}$, maximizing the variance of the projected data can be written as $\arg\max \mathrm{tr}(\boldsymbol{W}\boldsymbol{C}_{\mathrm{c}}\boldsymbol{W}^{\top})$, under the constraint $\boldsymbol{W}^{\top}\boldsymbol{W} = \mathbf{I}$. By using the Lagrangian multipliers and taking the derivative of the resulting cost function, we get the corresponding eigenproblem.

2.3.2 Case study 2: Nonparametric density estimation

Nonparametric density estimation is essential in many applied mathematical problems. Many machine learning techniques are based on density estimation, often with a Parzen window approach [48], [49]. This is the case of the information-theoretic methods [20], which are essentially based on the quadratic Rényi entropy, of the form $-\log\int p(\boldsymbol{x})^2 d\boldsymbol{x}$ for a probability density $p(\cdot)$. It is also the case of the (kernel) entropy component analysis (ECA) for data transformation and dimensionality reduction [19]. See also [18] for more details.

Often unknown, the probability density $p(\cdot)$ is estimated using a Parzen estimator of the form $\hat{p}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\kappa(\boldsymbol{x},\boldsymbol{x}_i)$ for a given kernel function centered at each available sample $\boldsymbol{x}_i$. By using the kernel matrix $\boldsymbol{K}$, the estimator is given by

$$\hat{p}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\kappa(\boldsymbol{x},\boldsymbol{x}_i) = \tfrac{1}{n}\mathbf{1}^{\top}\boldsymbol{\kappa}(\boldsymbol{x}),$$

where $\boldsymbol{\kappa}(\boldsymbol{x})$ is the vector of entries $\kappa(\boldsymbol{x},\boldsymbol{x}_i)$ for $i = 1,2,\ldots,n$. From the definition (12), it is easy to see that this expression corresponds to the inner product between the sample and the mean, with

$$\hat{p}(\boldsymbol{x}) = {\boldsymbol{x}^{\phi}}^{\top}\Big(\tfrac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^{\phi}\Big) = {\boldsymbol{x}^{\phi}}^{\top}\boldsymbol{\mu}^{\phi}$$

The quantity $\int p(\boldsymbol{x})^2 d\boldsymbol{x}$, related to the quadratic Rényi entropy, is therefore estimated by $\int \hat{p}(\boldsymbol{x})^2 d\boldsymbol{x} = \|\boldsymbol{\mu}^{\phi}\|^2$, which leads to

$$\int \hat{p}(\boldsymbol{x})^2 d\boldsymbol{x} = \tfrac{1}{n^2}\sum_{i=1}^{n}\lambda_i(\boldsymbol{\alpha}_i^{\top}\mathbf{1})^2, \tag{17}$$

where (6) is used. This expression uncovers the composition of the entropy in terms of eigenvectors of $\boldsymbol{K}$. This is the main motivation of the ECA, where the relevant eigenvectors are selected in order to maximize the estimated entropy, thus the smallest terms $\lambda_i(\boldsymbol{\alpha}_i^{\top}\mathbf{1})^2$. Therefore, any eigenvector $\boldsymbol{\alpha}_i$ for which $\boldsymbol{\alpha}_i^{\top}\mathbf{1} \neq 0$ and $\lambda_i \neq 0$ contributes to the entropy estimate, in contrast with the PCA where only eigenvectors associated to nonzero eigenvalues contribute to the variance.

Finally, we emphasize that non-centered data is used in the density estimation. Centering the data leads to a null density estimator, which yields an infinite quadratic Rényi entropy. This fact is also corroborated by several studies, including the (kernel) entropy component analysis. See [19] for more details.

2.4 Features from centered vs non-centered data

Even in conventional PCA, the issue of centering is still an open question[4]. To the best of our knowledge, only the work in [30] studied the link between the eigendecompositions of $\boldsymbol{C}$ and $\boldsymbol{C}_{\mathrm{c}}$. Expression (9) reveals the rank-one update nature between these two matrices. The analysis of the Gram matrices $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$ is obviously a much harder problem, as illustrated in expression (7).

In this paper, we take the initiative to study the of the inner product matrices, $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$, and carry on with the eigenanalysis of the outer product matrices, $\boldsymbol{C}$ and $\boldsymbol{C}_{\mathrm{c}}$. It turns out that the work [30] on $\boldsymbol{C}$ and $\boldsymbol{C}_{\mathrm{c}}$ can be derived easily from the proposed approach. Moreover, the study of the inner product matrices broadens the scope of the work to the analysis of all kernel-based techniques [12], [17], [19], beyond the PCA approach. Next section gives the main contributions of this paper, while this study is completed in Section 4 by several extensions, beyond centering.

3 Main results

The main results are given next, in two-folds: the (inner product) Gram matrix and the (outer product) covariance matrix. For each, we describe the relations between the eigenvectors of the centered matrix with those obtained from the non-centered one. The relations between the eigenvalues are studied by examining the inner product matrices, which allows the generalization of these results to nonlinear kernel functions. But before, we need to briefly introduce the orthogonal projection, and its link to the centering operation.

Background on orthogonal projections

The matrix of orthogonal projections onto the subspace spanned by the columns of a given matrix $\boldsymbol{M}$ is defined by $\boldsymbol{P}_{\boldsymbol{M}} = \boldsymbol{M}\big(\boldsymbol{M}^\top\boldsymbol{M}\big)^{-1}\boldsymbol{M}^\top$, while $\mathbf{I} - \boldsymbol{P}_{\boldsymbol{M}}$ denotes te projection onto its orthogonal complement. Projections are idempotent transformations, *i.e.*, $\boldsymbol{P}_{\boldsymbol{M}}\boldsymbol{P}_{\boldsymbol{M}} = \boldsymbol{P}_{\boldsymbol{M}}$. In particular, we are interested in this paper in the projection onto the all-ones vector $\mathbf{1}$ of n entries, with

$$\boldsymbol{P}_{\mathbf{1}} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top. \tag{18}$$

By considering the projection of the data onto the subspace spanned by $\mathbf{1}$ and its complement, we have[5]

$$\boldsymbol{X}\boldsymbol{P}_{\mathbf{1}} = \boldsymbol{\mu}\mathbf{1}^\top. \tag{19}$$

4. For instance in R (The R Project for Statistical Computing) [50], there are two ways to performs PCA: On the one hand, the *R-mode* by using the eigendecomposition of the covariance matrix as given in (10), with the function `princomp`; and on the other hand, the *Q-mode* by using the singular value decomposition of the non-centered data, as given in (4), with the function `svd`.

5. Since the data are given column-wise in the matrix $\boldsymbol{X}$, the mean $\boldsymbol{\mu}$ obtained from the operation $\boldsymbol{X}\boldsymbol{P}_{\mathbf{1}}$ is the vector of means of each row of $\boldsymbol{X}$. This is in opposition to the operation $\boldsymbol{P}_{\mathbf{1}}\boldsymbol{X}$ which provides the means of each of the columns, *i.e.*, each $\boldsymbol{x}_i$.

It is easy to see that the matrix of centered data is given by $\boldsymbol{X}_{\mathrm{c}} = \boldsymbol{X}(\mathbf{I} - \boldsymbol{P}_{\mathbf{1}})$ and verifies the identity $\boldsymbol{X}_{\mathrm{c}}\mathbf{1} = \mathbf{0}$. More generally, we have

$$(\mathbf{I} - \boldsymbol{P}_{\mathbf{1}})\mathbf{1} = \mathbf{1}^\top(\mathbf{I} - \boldsymbol{P}_{\mathbf{1}}) = \mathbf{0}. \tag{20}$$

Finally, for any square matrix $\boldsymbol{M}$ of appropriate size:

$$\boldsymbol{P}_{\mathbf{1}}\boldsymbol{M}\boldsymbol{P}_{\mathbf{1}} = \tfrac{1}{n^2}\mathbf{1}\mathbf{1}^\top\boldsymbol{M}\mathbf{1}\mathbf{1}^\top = \tfrac{\mathbf{1}^\top\boldsymbol{M}\mathbf{1}}{n}\boldsymbol{P}_{\mathbf{1}}. \tag{21}$$

This yields

$$\mathrm{tr}(\boldsymbol{P}_{\mathbf{1}}\boldsymbol{M}\boldsymbol{P}_{\mathbf{1}}) = \mathrm{tr}(\boldsymbol{P}_{\mathbf{1}}\boldsymbol{M}) = \mathrm{tr}(\boldsymbol{M}\boldsymbol{P}_{\mathbf{1}}) = \tfrac{1}{n}\mathbf{1}^\top\boldsymbol{M}\mathbf{1}, \tag{22}$$

where the cyclic property of the trace operator is used in the first two equalities. Therefore, we have

$$\mathrm{tr}\big((\mathbf{I} - \boldsymbol{P}_{\mathbf{1}})\boldsymbol{M}(\mathbf{I} - \boldsymbol{P}_{\mathbf{1}})\big) = \mathrm{tr}(\boldsymbol{M}) - \tfrac{1}{n}\mathbf{1}^\top\boldsymbol{M}\mathbf{1}. \tag{23}$$

3.1 Inner product – the Gram matrix $\boldsymbol{K}_{\mathrm{c}}$

Let $\boldsymbol{K} = \boldsymbol{X}^\top\boldsymbol{X}$ be the Gram matrix, then its centered counterpart is $\boldsymbol{K}_{\mathrm{c}} = \boldsymbol{X}_{\mathrm{c}}^\top\boldsymbol{X}_{\mathrm{c}}$, namely

$$\begin{aligned}
\boldsymbol{K}_{\mathrm{c}} &= (\mathbf{I} - \boldsymbol{P}_{\mathbf{1}})\boldsymbol{K}(\mathbf{I} - \boldsymbol{P}_{\mathbf{1}}) && (24)\\
&= \boldsymbol{K} - \boldsymbol{P}_{\mathbf{1}}\boldsymbol{K} - \boldsymbol{K}\boldsymbol{P}_{\mathbf{1}} + \tfrac{\mathbf{1}^\top\boldsymbol{K}\mathbf{1}}{n}\boldsymbol{P}_{\mathbf{1}} && (25)\\
&= \boldsymbol{K} - \boldsymbol{P}_{\mathbf{1}}\boldsymbol{K} - \boldsymbol{K}\boldsymbol{P}_{\mathbf{1}} + n\|\boldsymbol{\mu}\|^2\boldsymbol{P}_{\mathbf{1}}, && (26)
\end{aligned}$$

where relation (21) is applied on $\boldsymbol{K}$, and the last equality is due to (6). These expressions reveal the *double centering*, which corresponds to subtracting the row and column means of the matrix $\boldsymbol{K}$ from its entries, and adding the grand mean. A byproduct of the double centering is obtained from (24) and (20): 0 is the eigenvalue of $\boldsymbol{K}_{\mathrm{c}}$ associated to the eigenvector $\frac{1}{n}\mathbf{1}$.

A first study to understand the distribution of the data associated to each matrix, $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$, is given by their traces, since they correspond to the sum of all eigenvalues. The following lemma provides a measure of the variance reduction due to centering.

Lemma 1. *Let $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$ be respectively the Gram matrix and its centered counterpart, then their corresponding traces verify the following relationships:*

$$\mathrm{tr}(\boldsymbol{K}_{\mathrm{c}}) = \mathrm{tr}(\boldsymbol{K}) - n\|\boldsymbol{\mu}\|^2,$$

and

$$\frac{\mathrm{tr}(\boldsymbol{K}_{\mathrm{c}})}{\mathrm{tr}(\boldsymbol{K})} = 1 - \frac{\mathbf{1}^\top\boldsymbol{K}\mathbf{1}}{n\,\mathrm{tr}(\boldsymbol{K})}.$$

Hence, their eigenvalues verify $\sum_{i=1}^n\lambda_{\mathrm{c}i} = \sum_{i=1}^n\lambda_i - n\|\boldsymbol{\mu}\|^2$.

Proof: The proof is straightforward, by applying (23) to $\boldsymbol{K}$ given in definition (24), with $\mathbf{1}^\top\boldsymbol{K}\mathbf{1} = n^2\|\boldsymbol{\mu}\|^2$. □

Next, we explore beyond the sum of the eigenvalues. In Section 3.1.1, we show that eigenvalues of $\boldsymbol{K}$ are interlaced with those of $\boldsymbol{K}_{\mathrm{c}}$. In Section 3.1.2, we provide bounds on a sum of the largest t eigenvalues, for any t, and in particular a lower bound on the largest eigenvalue of $\boldsymbol{K}_{\mathrm{c}}$.

3.1.1 Eigenvalue interlacing theorems for $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$

Before proceeding, the following theorem from [51, Theorem 5.9] is central to our study. It is worth noting that this theorem has been known in the literature for some time, see for instance [52, Appendix A], and is obtained from the Poincaré Separation Theorem [53], [54] (see also [29, Corollary 4.3.37 in page 248]).

Theorem 2 (Separation Theorem [51], [52])**.** *Let $\boldsymbol{M}$ be a d-by-n matrix. Let two orthogonal projection matrices be $\boldsymbol{P}_{\mathrm{left}}$, of size d-by-d, and $\boldsymbol{P}_{\mathrm{right}}$, of size n-by-n. Then,*

$$\sigma_{j+t}(\boldsymbol{M}) \leq \sigma_j(\boldsymbol{P}_{\mathrm{left}}\ \boldsymbol{M}\ \boldsymbol{P}_{\mathrm{right}}) \leq \sigma_j(\boldsymbol{M}),$$

where $\sigma_j(\cdot)$ denotes the j-th largest singular value of the matrix, $t = d - r(\boldsymbol{P}_{\mathrm{right}}) + n - r(\boldsymbol{P}_{\mathrm{left}})$ and $r(\cdot)$ is the rank of the matrix.

Theorem 3. *Let $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$ be respectively the Gram matrix and its centered counterpart, then their eigenvalues are interlaced, such that*

$$\lambda_{j+1} \leq \lambda_{\mathrm{c}j} \leq \lambda_j,$$

with $\lambda_{\mathrm{c}n} = 0$, where λ_j and $\lambda_{\mathrm{c}j}$ denote respectively the j-th largest eigenvalue of the matrices $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$.

Proof: To prove this, we apply the Separation Theorem 2 with $\boldsymbol{M} = \boldsymbol{X}$, $\boldsymbol{P}_{\mathrm{left}}$ being the d-by-d identity matrix, and $\boldsymbol{P}_{\mathrm{right}} = (\mathbf{I} - \boldsymbol{P}_{\mathbf{1}})$, where $r(\boldsymbol{P}_{\mathrm{left}}) = d$ and $r(\boldsymbol{P}_{\mathrm{right}}) = n - 1$, and thus $t = 1$. In this case, we get

$$\sigma_{j+1}(\boldsymbol{X}) \leq \sigma_j(\boldsymbol{X}(\mathbf{I} - \boldsymbol{P}_{\mathbf{1}})) \leq \sigma_j(\boldsymbol{X}).$$

Relations (5) and (24) are used to conclude the proof. □

Corollary 4. *Let $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$ be respectively the Gram matrix and its centered counterpart, then their proportion of total variation accounted for by the eigenvectors are interlaced, such that*

$$\pi_{j+1} \leq \gamma\ \pi_{\mathrm{c}j} \leq \pi_j,$$

where $\pi_j = \frac{\lambda_j}{\sum_i \lambda_i}$, $\pi_{\mathrm{c}j} = \frac{\lambda_{\mathrm{c}j}}{\sum_i \lambda_{\mathrm{c}i}}$ and $\gamma = \frac{\mathrm{tr}(\boldsymbol{K}_{\mathrm{c}})}{\mathrm{tr}(\boldsymbol{K})}$.

Proof: The proof is direct, on the one hand by dividing the inequalities of Theorem 3 by the trace of $\boldsymbol{K}$, and on the other hand by setting $\gamma = \frac{\mathrm{tr}(\boldsymbol{K}_{\mathrm{c}})}{\mathrm{tr}(\boldsymbol{K})}$ with the direct application of (1). □

All these results show the impact of centering the data on the distribution of the eigenvalues, where the eigenvalues of $\boldsymbol{K}_{\mathrm{c}}$ are *sandwitched* between the eigenvalues of $\boldsymbol{K}$. This illustrates that $\boldsymbol{K}_{\mathrm{c}}$ behaves like a "coarse" matrix compared to $\boldsymbol{K}$.

3.1.2 Bounds on the eigenvalues of $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$

In this section, we provide lower bounds on the largest eigenvalues of the matrices $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$. To this end, we state the Schur–Horn Theorem [55]. See [56, Chapter 9] for a recent review on the theory of majorization.

Theorem 5 (Schur–Horn Theorem)**.** *For any n-by-n symmetric matrix with diagonal entries $d_1, d_2, \ldots, d_n$ and eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ given in non-increasing order, we have:*

$$\sum_{i=1}^{t} d_i \leq \sum_{i=1}^{t} \lambda_i,$$

for any $t = 1, 2, \ldots, n$, with equality for $t = n$.

The Schur–Horn Theorem has been proven in the situation when the diagonal entries $d_1, d_2, \ldots, d_n$ are also given in non-increasing order. Still, one can also use any subset of the diagonal entries, although the resulting lower bound in the above theorem may not be as tight as when using the statement $d_n \leq \ldots \leq d_2 \leq d_1$.

By applying this theorem to both matrices $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$, we get the following result.

Lemma 6. *Let $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$ be the Gram matrix and its centered counterpart with, respectively, diagonal entries $d_1, d_2, \ldots, d_n$ and $d_{\mathrm{c}1}, d_{\mathrm{c}2}, \ldots, d_{\mathrm{c}n}$; and eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and $\lambda_{\mathrm{c}1}, \lambda_{\mathrm{c}2}, \ldots, \lambda_{\mathrm{c}n}$, given in non-increasing order. Then*

$$\sum_{i=1}^{t} d_i \leq \sum_{i=1}^{t} \lambda_i, \qquad \sum_{i=1}^{t} d_{\mathrm{c}i} \leq \sum_{i=1}^{t} \lambda_{\mathrm{c}i},$$

with equalities for $t = n$.

The direct application of the Schur–Horn Theorem separately on each matrix, $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$, does not give any particular result. The following lemma is a first step towards a connection between the eigenvalues of both matrices, and allows to establish the bounds given in Theorem 8.

Lemma 7. *Let $\boldsymbol{K} = \boldsymbol{A}\,\boldsymbol{\Lambda}\,\boldsymbol{A}^\top$ and $\boldsymbol{K}_{\mathrm{c}} = \boldsymbol{A}_{\mathrm{c}}\,\boldsymbol{\Lambda}_{\mathrm{c}}\,\boldsymbol{A}_{\mathrm{c}}^\top$ be the spectral decompositions of the Gram matrices. Then the spectral decomposition of the matrix $\boldsymbol{A}^\top\boldsymbol{K}_{\mathrm{c}}\boldsymbol{A}$ is $\boldsymbol{A}^\top\boldsymbol{A}_{\mathrm{c}}\,\boldsymbol{\Lambda}_{\mathrm{c}}\,(\boldsymbol{A}^\top\boldsymbol{A}_{\mathrm{c}})^\top$ and of the matrix $\boldsymbol{A}_{\mathrm{c}}^\top\boldsymbol{K}\boldsymbol{A}_{\mathrm{c}}$ is $\boldsymbol{A}_{\mathrm{c}}^\top\boldsymbol{A}\,\boldsymbol{\Lambda}\,(\boldsymbol{A}_{\mathrm{c}}^\top\boldsymbol{A})^\top$.*

It is easy to prove these results, by replacing either $\boldsymbol{K}_{\mathrm{c}}$ or $\boldsymbol{K}$ by its spectral decomposition and verifying that the product of two orthonormal matrices is orthonormal. This lemma shows that the matrix $\boldsymbol{A}^\top\boldsymbol{K}_{\mathrm{c}}\boldsymbol{A}$ has the same eigenvalues as the matrix $\boldsymbol{K}_{\mathrm{c}}$, and its eigenvectors are the columns of $\boldsymbol{A}^\top\boldsymbol{A}_{\mathrm{c}}$, which is the matrix whose entries are the inner products between the eigenvectors of $\boldsymbol{K}$ and the eigenvectors of $\boldsymbol{K}_{\mathrm{c}}$ (*i.e.*, $\boldsymbol{\alpha}_i^\top\boldsymbol{\alpha}_{\mathrm{c}j}$). Since both matrices $\boldsymbol{K}_{\mathrm{c}}$ and $\boldsymbol{A}^\top\boldsymbol{K}_{\mathrm{c}}\boldsymbol{A}$ share the same eigenvalues, we propose next to apply the Schur–Horn Theorem on the latter matrix.

Theorem 8. *The sum of the largest t (for any $t = 1, 2, \ldots, n$) eigenvalues of the matrix $\boldsymbol{K}_{\mathrm{c}}$ is lower bounded as follows:*

$$\sum_{i=1}^{t} d'_i \leq \sum_{i=1}^{t} \lambda_{\mathrm{c}i},$$

where d'_i is the i-th largest value of $\lambda_i + (\|\boldsymbol{\mu}\|^2 - \frac{2}{n}\lambda_i)\,(\boldsymbol{\alpha}_i^\top\mathbf{1})^2$, for $i = 1, 2, \ldots, n$. Here, $(\lambda_i, \boldsymbol{\alpha}_i)$ is an eigenpair of the matrix $\boldsymbol{K}$. This expression provides a lower bound on the largest

eigenvalue of $\boldsymbol{K}_{\mathrm{c}}$, by setting $t = 1$:

$$\max_{i=1,\ldots,n} \lambda_i + (\|\boldsymbol{\mu}\|^2 - \tfrac{2}{n}\lambda_i)\ (\boldsymbol{\alpha}_i^\top \mathbf{1})^2 \quad \leq \quad \lambda_{\mathrm{c}1}. \qquad (27)$$

Proof: To prove this theorem, we describe the diagonal entries of the matrix $\boldsymbol{A}^\top \boldsymbol{K}_{\mathrm{c}} \boldsymbol{A}$, namely for any i:

$$\begin{aligned}
&\boldsymbol{\alpha}_i^\top \boldsymbol{K}_{\mathrm{c}} \boldsymbol{\alpha}_i \\
&\quad = \boldsymbol{\alpha}_i^\top \big(\boldsymbol{K} - \boldsymbol{P}_1 \boldsymbol{K} - \boldsymbol{K}\boldsymbol{P}_1 + n\|\boldsymbol{\mu}\|^2 \boldsymbol{P}_1\big)\boldsymbol{\alpha}_i \\
&\quad = \lambda_i \boldsymbol{\alpha}_i^\top \boldsymbol{\alpha}_i - \lambda_i \boldsymbol{\alpha}_i^\top \boldsymbol{P}_1 \boldsymbol{\alpha}_i - \lambda_i \boldsymbol{\alpha}_i^\top \boldsymbol{P}_1 \boldsymbol{\alpha}_i + n\|\boldsymbol{\mu}\|^2 \boldsymbol{\alpha}_i^\top \boldsymbol{P}_1 \boldsymbol{\alpha}_i \\
&\quad = \lambda_i \boldsymbol{\alpha}_i^\top \boldsymbol{\alpha}_i + (n\|\boldsymbol{\mu}\|^2 - 2\lambda_i)\ \boldsymbol{\alpha}_i^\top \boldsymbol{P}_1 \boldsymbol{\alpha}_i \\
&\quad = \lambda_i + (\|\boldsymbol{\mu}\|^2 - \tfrac{2}{n}\lambda_i)\ (\boldsymbol{\alpha}_i^\top \mathbf{1})^2,
\end{aligned}$$

where the first equality follows from expression (26), the second equality is due to the fact that $(\lambda_i, \boldsymbol{\alpha}_i)$ is an eigenpair of $\boldsymbol{K}$, and the last equality follows from the definition of $\boldsymbol{P}_1$ given in (18). To conclude the proof, Theorem 5 is applied on the matrix $\boldsymbol{A}^\top \boldsymbol{K}_{\mathrm{c}} \boldsymbol{A}$, after observing from Lemma 7 that both matrices $\boldsymbol{K}_{\mathrm{c}}$ and $\boldsymbol{A}^\top \boldsymbol{K}_{\mathrm{c}} \boldsymbol{A}$ share the same eigenvalues. □

This theorem provides a further characterization of the eigenvalues of both matrices $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$, beyond the relation of their traces given in Lemma 1. Moreover, the latter lemma is obtained as a particular case of our theorem, when $t = n$ where the equality in Theorem 8 holds, namely $\sum_{i=1}^n d'_i = \sum_{i=1}^n \lambda_{\mathrm{c}i}$. To see this, first observe that $\sum_{i=1}^n \lambda_{\mathrm{c}i} = \mathrm{tr}(\boldsymbol{K}_{\mathrm{c}})$. Then, we have

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{K}_{\mathrm{c}}) &= \sum_{i=1}^n d'_i \\
&= \sum_{i=1}^n \lambda_i + (\|\boldsymbol{\mu}\|^2 - \tfrac{2}{n}\lambda_i)\ (\boldsymbol{\alpha}_i^\top \mathbf{1})^2 \\
&= \sum_{i=1}^n \lambda_i + \|\boldsymbol{\mu}\|^2 \sum_{i=1}^n (\boldsymbol{\alpha}_i^\top \mathbf{1})^2 - \tfrac{2}{n} \sum_{i=1}^n \lambda_i (\boldsymbol{\alpha}_i^\top \mathbf{1})^2 \\
&= \mathrm{tr}(\boldsymbol{K}) - n\|\boldsymbol{\mu}\|^2
\end{aligned}$$

where the expression of $\|\boldsymbol{\mu}\|^2$ follows from (6), and we have used $\sum_{i=1}^n (\boldsymbol{\alpha}_i^\top \mathbf{1})^2 = \sum_{i=1}^n \mathbf{1}^\top \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^\top \mathbf{1} = \mathbf{1}^\top \boldsymbol{A}\boldsymbol{A}^\top \mathbf{1} = \mathbf{1}^\top \mathbf{1} = n$. This illustrates the tightness of the derived bounds.

Furthermore, it is worth noting that the largest value of $\lambda_i + (\|\boldsymbol{\mu}\|^2 - \frac{2}{n}\lambda_i)\ (\boldsymbol{\alpha}_i^\top \mathbf{1})^2$ needs not to be given by the largest eigenvalue λ_1 and the corresponding eigenvector $\boldsymbol{\alpha}_1$. To see this, we show that $\|\boldsymbol{\mu}\|^2 - \frac{2}{n}\lambda_i < 0$ in this case. To this end, we apply the celebrated Courant-Fischer Theorem on the matrix $\boldsymbol{K}$, which states that $\lambda_1 = \max_{\boldsymbol{v}} \frac{\boldsymbol{v}^\top \boldsymbol{K} \boldsymbol{v}}{\boldsymbol{v}^\top \boldsymbol{v}}$. As a consequence, λ_1 is larger or equal to the special case when $\boldsymbol{v} = \mathbf{1}$. Consequently $\lambda_1 \geq \frac{1}{n}\mathbf{1}^\top \boldsymbol{K} \mathbf{1}$, and therefore we have $\|\boldsymbol{\mu}\|^2 < \frac{2}{n}\lambda_1$ from (6).

Finally, Theorem 8 reveals the terms $\lambda_i (\boldsymbol{\alpha}_i^\top \mathbf{1})^2$ in the lower bound. It turns out that these terms are the building blocks of the entropy estimate, as given in (17) for the (kernel) entropy component analysis.

3.1.3 Eigenvectors of $\boldsymbol{K}_{\mathrm{c}}$

We propose to go further beyond the analysis of the eigenvalues as given so far. In this section, we study the eigenvectors of the centered Gram matrix. The following theorem provides insights on the eigenvectors of the matrix $\boldsymbol{K}_{\mathrm{c}}$.

Theorem 9. *For any eigenvector $\boldsymbol{\alpha}_{\mathrm{c}j}$ of the matrix $\boldsymbol{K}_{\mathrm{c}}$ associated to a non-zero eigenvalue, its entries sum to zero, namely $\boldsymbol{\alpha}_{\mathrm{c}j}^\top \mathbf{1} = 0$ for any $\lambda_{\mathrm{c}j} \neq 0$.*

Proof: The proof is straightforward, with

$$\boldsymbol{\alpha}_{\mathrm{c}i}^\top \mathbf{1} = \tfrac{1}{\lambda_{\mathrm{c}i}} \boldsymbol{\alpha}_{\mathrm{c}i}^\top \boldsymbol{K}_{\mathrm{c}} \mathbf{1} = \tfrac{1}{\lambda_{\mathrm{c}i}} \boldsymbol{\alpha}_{\mathrm{c}i}^\top (\mathbf{I} - \boldsymbol{P}_1) \boldsymbol{K} (\mathbf{I} - \boldsymbol{P}_1) \mathbf{1} = \mathbf{0},$$

where the first equality follows from the eigenproblem (8) and the last equality is due to (20). □

It is therefore easy to see that $\boldsymbol{P}_1 \boldsymbol{\alpha}_{\mathrm{c}j} = \mathbf{0}$ for any eigenvector of $\boldsymbol{K}_{\mathrm{c}}$ associated to a non-zero eigenvalue, and we have in its dual form $(\mathbf{I} - \boldsymbol{P}_1)\boldsymbol{\alpha}_{\mathrm{c}j} = \boldsymbol{\alpha}_{\mathrm{c}j}$.

Theorem 10. *All entries of any eigenvector $\boldsymbol{\alpha}_{\mathrm{c}j}$ of the matrix $\boldsymbol{K}_{\mathrm{c}}$ of the centered data are bounded with*

$$-1 \leq \boldsymbol{\alpha}_{\mathrm{c}j} \leq 1,$$

where inequalities are applied element-wise.

Proof: It is well known that $\|\boldsymbol{\alpha}_{\mathrm{c}j}\|_\infty \leq \|\boldsymbol{\alpha}_{\mathrm{c}j}\| \leq \|\boldsymbol{\alpha}_{\mathrm{c}j}\|_1 \leq \sqrt{n}\|\boldsymbol{\alpha}_{\mathrm{c}j}\|_\infty$, where $\|\cdot\|_\infty$ is the supremum norm which takes the largest absolute value of the vector's entries. Since eigenvectors have a unit norm, we get the pair of inequalities. □

By combining Theorems 9 and 10, we have that each eigenvector of the centered Gram matrix verifies the sum-to-one and the boxed constraints. These constraints are equivalent to the well-known constraints of *SVM*. One can also describe data-driven bounds when a normalization is operated, as given in (15) for instance. In this case, the normalization $\|\boldsymbol{\alpha}_{\mathrm{c}i}\|^2 = 1/\lambda_{\mathrm{c}i}$ yields a modification on the box constraints, since $\|\boldsymbol{\alpha}_{\mathrm{c}j}\|_\infty \leq \|\boldsymbol{\alpha}_{\mathrm{c}j}\| = 1/\sqrt{\lambda_{\mathrm{c}i}}$. The latter can also be upper bounded by using the eigenvalues of $\boldsymbol{K}$, thanks to the interlacing property derived in Theorem 3.

The eigenvectors of the Gram matrix are seldom used directly, but often considered to define relevant axes. This is shown in Section 2.3.1 for the principal component analysis, and in Section 2.3.2 for the (kernel) entropy component analysis. In either methods, the relevant axes are determined by a weighted linear combination of the data, the weights being the eigenvectors of the Gram matrix, up to a normalization factor. By considering the normalization given in (15), we get

$$\boldsymbol{w}_j = \tfrac{1}{\sqrt{\lambda_i}} \boldsymbol{X} \boldsymbol{\alpha}_j, \qquad (28)$$

and its centered counterpart $\boldsymbol{w}_{\mathrm{c}j} = \boldsymbol{X}_{\mathrm{c}} \boldsymbol{\alpha}_{\mathrm{c}j} / \sqrt{\lambda_{\mathrm{c}i}}$. Therefore, the projection of the data on either axes is:

$$\boldsymbol{X}^\top \boldsymbol{w}_j = \tfrac{1}{\sqrt{\lambda_i}} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\alpha}_j = \tfrac{1}{\sqrt{\lambda_i}} \boldsymbol{K} \boldsymbol{\alpha}_j = \sqrt{\lambda_i}\, \boldsymbol{\alpha}_j$$

and likewise $\boldsymbol{X}_{\rm c}^\top \boldsymbol{w}_{{\rm c}j} = \sqrt{\lambda_{{\rm c}i}}\boldsymbol{\alpha}_{{\rm c}j}$. By using the normalization (16), we get a unit variance along the respective axes, with

$$\boldsymbol{X}^\top \boldsymbol{w}_j = \boldsymbol{\alpha}_j, \qquad \text{and} \qquad \boldsymbol{X}_{\rm c}^\top \boldsymbol{w}_{{\rm c}j} = \boldsymbol{\alpha}_{{\rm c}j}. \tag{29}$$

The analysis of these axes is derived next, by examining the outer product matrices.

3.2 Outer product – the covariance matrix $C_{\rm c}$

The relation between the covariance matrix $\boldsymbol{C}_{\rm c} = \frac{1}{n}\boldsymbol{X}_{\rm c}\boldsymbol{X}_{\rm c}^\top$, and $\boldsymbol{C} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$, *i.e.*, the second-order non-central moment matrix, is given by

$$\boldsymbol{C}_{\rm c} = \tfrac{1}{n}\boldsymbol{X}(\mathbf{I} - \boldsymbol{P}_1)\boldsymbol{X}^\top = \boldsymbol{C} - \boldsymbol{\mu}\boldsymbol{\mu}^\top. \tag{30}$$

Since $\frac{1}{\|\boldsymbol{\mu}\|^2}\boldsymbol{\mu}\boldsymbol{\mu}^\top$ denotes the d-by-d projection matrix onto the vector mean $\boldsymbol{\mu}$ then, by analogy with the definition of $\boldsymbol{K}_{\rm c}$ in (26), we have here a simpler expression. We can therefore revisit all the results given in Section 3.1 to describe relations between the eigenvectors of $\boldsymbol{C}$ and $\boldsymbol{C}_{\rm c}$. The eigenvalues of these matrices still satisfy the interlacing theorems given in Section 3.1.1, since the eigenvalues of $\boldsymbol{C}$ and $\boldsymbol{C}_{\rm c}$ are respectively $\frac{1}{n}\lambda_1, \frac{1}{n}\lambda_2, \frac{1}{n}\lambda_3 \ldots$, and $\frac{1}{n}\lambda_{{\rm c}1}, \frac{1}{n}\lambda_{{\rm c}2}, \frac{1}{n}\lambda_{{\rm c}3} \ldots$. By analogy with Lemma 1, we have $\mathrm{tr}(\boldsymbol{C}_{\rm c}) = \mathrm{tr}(\boldsymbol{C}) - \|\boldsymbol{\mu}\|^2$.

The following lemma provides an expression essential to our study.

Lemma 11. *For any eigenpair $(\lambda_i, \boldsymbol{w}_i)$ of $\boldsymbol{C}$ and any eigenpair $(\lambda_{{\rm c}i}, \boldsymbol{w}_{{\rm c}i})$ of $\boldsymbol{C}_{\rm c}$, we have the following relation with the mean vector $\boldsymbol{\mu}$:*

$$\left(\lambda_i - \lambda_{{\rm c}j}\right) \boldsymbol{w}_i^\top \boldsymbol{w}_{{\rm c}j} = n\ \boldsymbol{w}_i^\top \boldsymbol{\mu}\ \ \boldsymbol{w}_{{\rm c}j}^\top \boldsymbol{\mu}.$$

Proof: Since $\boldsymbol{w}_{{\rm c}j}$ is an eigenvector of $\boldsymbol{C}_{\rm c}$, we have from expression (13): $\boldsymbol{w}_i^\top \boldsymbol{w}_{{\rm c}j} = \frac{n}{\lambda_{{\rm c}j}}\, \boldsymbol{w}_i^\top \boldsymbol{C}_{\rm c}\, \boldsymbol{w}_{{\rm c}j}$. By substituting the definition of $\boldsymbol{C}_{\rm c}$ from (30), we get:

$$\begin{aligned}\boldsymbol{w}_i^\top \boldsymbol{w}_{{\rm c}j} &= \tfrac{n}{\lambda_{{\rm c}j}}\, \boldsymbol{w}_i^\top (\boldsymbol{C} - \boldsymbol{\mu}\boldsymbol{\mu}^\top)\, \boldsymbol{w}_{{\rm c}j} \\ &= \tfrac{n}{\lambda_{{\rm c}j}}\, \boldsymbol{w}_i^\top \boldsymbol{C}\, \boldsymbol{w}_{{\rm c}j} - \tfrac{n}{\lambda_{{\rm c}j}}\, \boldsymbol{w}_i^\top \boldsymbol{\mu}\, \boldsymbol{\mu}^\top\, \boldsymbol{w}_{{\rm c}j}.\end{aligned}$$

The first term in the right-hand-side can be simplified, since $\boldsymbol{w}_i$ is an eigenvector of $\boldsymbol{C}$, thus $\boldsymbol{w}_i^\top \boldsymbol{C} = \frac{1}{n}\lambda_i \boldsymbol{w}_i^\top$. □

In the above expression, $\boldsymbol{w}_i^\top \boldsymbol{\mu}$ corresponds to the so-called mean score over the i-th principal component $\boldsymbol{w}_i$ in non-centered PCA, since $\boldsymbol{w}_i^\top \boldsymbol{\mu} = \boldsymbol{w}_i^\top \frac{1}{n}\boldsymbol{X}\mathbf{1}$ corresponds to the mean of the *score vector* $\boldsymbol{X}^\top \boldsymbol{w}_i$. Moreover, we have

$$\boldsymbol{w}_i^\top \boldsymbol{\mu} = \tfrac{1}{\sqrt{\lambda_i}}\boldsymbol{\alpha}_i^\top \boldsymbol{X}^\top \tfrac{1}{n}\boldsymbol{X}\mathbf{1} = \tfrac{1}{n\sqrt{\lambda_i}}\boldsymbol{\alpha}_i^\top \boldsymbol{K}\mathbf{1} = \tfrac{\sqrt{\lambda_i}}{n}\, \boldsymbol{\alpha}_i^\top \mathbf{1}, \tag{31}$$

where the first equality follows from (28) and the last one is due to the fact that $\boldsymbol{\alpha}_i$ is an eigenvector of $\boldsymbol{K}$. Once again, we get the main building blocks of the entropy estimate (17). Therefore, it is easy to see that the entropy estimate from the ECA is simply

$$\int \hat{p}(\boldsymbol{x})^2 d\boldsymbol{x} = \tfrac{1}{n^2}\sum_{i=1}^{n} \lambda_i (\boldsymbol{\alpha}_i^\top \mathbf{1})^2 = \|\boldsymbol{W}^\top \boldsymbol{\mu}\|^2.$$

Expression (31) illustrates that Theorem 8 can be rewritten in terms of $\boldsymbol{w}_i^\top \boldsymbol{\mu}$. In the following theorem, we derive more effective expressions thanks to the simplified definition of $\boldsymbol{C}_{\rm c}$.

Theorem 12. *Theorem 8 can be expressed as follows:*

$$\sum_{i=1}^{t} d'_i \leq \sum_{i=1}^{t} \lambda_{{\rm c}i},$$

where d'_i is the i-th largest value of $\lambda_i - n(\boldsymbol{w}_i^\top \boldsymbol{\mu})^2$, for $i = 1, 2, \ldots, n$. In this expression, $(\frac{1}{n}\lambda_i, \boldsymbol{w}_i)$ is an eigenpair of the matrix $\boldsymbol{C}$. Therefore, we have:

$$\max_{i=1,\ldots,n} \lambda_i - n(\boldsymbol{w}_i^\top \boldsymbol{\mu})^2 \quad \leq \quad \lambda_{{\rm c}1}.$$

Proof: To prove this result, we use essentially the same steps given in the proof of Theorem 8, by considering the diagonal entries of the matrix $\boldsymbol{W}^\top \boldsymbol{C}_{\rm c} \boldsymbol{W}$, with

$$\boldsymbol{w}_i^\top \boldsymbol{C}_{\rm c} \boldsymbol{w}_i = \boldsymbol{w}_i^\top (\boldsymbol{C} - \boldsymbol{\mu}\boldsymbol{\mu}^\top)\boldsymbol{w}_i = \tfrac{\lambda_i}{n}\boldsymbol{w}_i^\top \boldsymbol{w}_i - \boldsymbol{w}_i^\top \boldsymbol{\mu}\, \boldsymbol{\mu}^\top \boldsymbol{w}_i,$$

and therefore $\boldsymbol{w}_i^\top \boldsymbol{C}_{\rm c} \boldsymbol{w}_i = \frac{\lambda_i}{n} - (\boldsymbol{w}_i^\top \boldsymbol{\mu})^2$. Finally, we apply Theorem 5 on the matrix $\boldsymbol{W}^\top \boldsymbol{C}_{\rm c} \boldsymbol{W}$, and observe by analogy to Lemma 7 that both matrices $\boldsymbol{C}_{\rm c}$ and $\boldsymbol{W}^\top \boldsymbol{C}_{\rm c} \boldsymbol{W}$ share the same eigenvalues, *i.e.*, $\frac{1}{n}\lambda_{{\rm c}1}, \frac{1}{n}\lambda_{{\rm c}2}, \ldots, \frac{1}{n}\lambda_{{\rm c}n}$. □

The previous theorem has several important consequences. The following theorem states that the mean vector is close to the eigenvector associated to the largest eigenvalue of $\boldsymbol{C}$.

Theorem 13. *We have the following lower bound on the inner product between the mean vector $\boldsymbol{\mu}$ and the first eigenvector of $\boldsymbol{C}$:*

$$\lambda_1 - \lambda_{{\rm c}1} \leq n(\boldsymbol{w}_1^\top \boldsymbol{\mu})^2.$$

Proof: From Theorem 12, $\lambda_i - \lambda_{{\rm c}1} \leq n(\boldsymbol{w}_i^\top \boldsymbol{\mu})^2$ for any i. This inequality can be investigated only when the left-hand-side is non-negative. As shown from the interlacing property in Theorem 3, $\lambda_{{\rm c}1} \leq \lambda_i$ if and only if $i = 1$. By considering this case, we conclude the proof. □

The following theorem shows that the eigenvectors associated to the largest eigenvalue of each matrix $\boldsymbol{C}$ and $\boldsymbol{C}_{\rm c}$, cannot be arbitrary different.

Theorem 14. *We have the following lower bound on the inner product between the first eigenvector of each of $\boldsymbol{C}$ and $\boldsymbol{C}_{\rm c}$:*

$$\frac{(\boldsymbol{w}_{{\rm c}1}^\top \boldsymbol{\mu})^2}{\|\boldsymbol{\mu}\|^2} \leq (\boldsymbol{w}_1^\top \boldsymbol{w}_{{\rm c}1})^2.$$

Proof: From Theorem 13, and by replacing the left-hand-side by the expression given in Lemma 11, we get

$$\frac{\boldsymbol{w}_1^\top \boldsymbol{\mu}\ \ \boldsymbol{w}_{{\rm c}1}^\top \boldsymbol{\mu}}{\boldsymbol{w}_1^\top \boldsymbol{w}_{{\rm c}1}} \leq (\boldsymbol{w}_1^\top \boldsymbol{\mu})^2.$$

By squaring and simplifying by $(\boldsymbol{w}_1^\top \boldsymbol{\mu})^2$, we obtain:

$$\frac{(\boldsymbol{w}_{{\rm c}1}^\top \boldsymbol{\mu})^2}{(\boldsymbol{w}_1^\top \boldsymbol{\mu})^2} \leq (\boldsymbol{w}_1^\top \boldsymbol{w}_{{\rm c}1})^2.$$

The above denominator can be upper bounded thanks to the Cauchy-Schwarz inequality, with $(\boldsymbol{w}_1^\top \boldsymbol{\mu})^2 \leq$

$\|\boldsymbol{w}_1\|^2\|\boldsymbol{\mu}\|^2 = \|\boldsymbol{\mu}\|^2$, due to the normalization of the eigenvectors. By combining these results, this concludes the proof. □

This theorem has an immediate result which shows that $\boldsymbol{w}_{\mathrm{c}1}$ is *closer* to $\boldsymbol{w}_1$ than to $\boldsymbol{\mu}$. This property is illustrated in the following corollary.

Corollary 15. *The cosine of the angle between the first eigenvectors of $\boldsymbol{C}$ and $\boldsymbol{C}_{\mathrm{c}}$ is lower bounded as follows:*

$$\cos(\boldsymbol{w}_{\mathrm{c}1}, \boldsymbol{\mu})^2 \leq \cos(\boldsymbol{w}_{\mathrm{c}1}, \boldsymbol{w}_1)^2.$$

Proof: The proof is straightforward from Theorem 14 and the definition of the inner product with $\|\boldsymbol{w}_1\| = \|\boldsymbol{w}_{\mathrm{c}1}\| = 1$, namely $\boldsymbol{w}_1^\top \boldsymbol{w}_{\mathrm{c}1} = \cos(\boldsymbol{w}_1, \boldsymbol{w}_{\mathrm{c}1})$ and $\boldsymbol{w}_1^\top \boldsymbol{\mu} = \|\boldsymbol{\mu}\| \cos(\boldsymbol{w}_1, \boldsymbol{\mu})$. □

We conclude this section by giving a summary of the relations obtained between the eigenvectors associated to the largest eigenvalues of the covariance matrix and its non-centered counterpart. In the non-centered case, we see from Theorem 13 that the eigenvector $\boldsymbol{w}_1$ of $\boldsymbol{C}$ tends to be collinear with the the mean vector $\boldsymbol{\mu}$. Now, consider the case when data are centered, which leads to the first eigenvector $\boldsymbol{w}_{\mathrm{c}1}$ of $\boldsymbol{C}_{\mathrm{c}}$. Theorem 14 and Corollary 15 provide inequalities that measure the fact that $\boldsymbol{w}_{\mathrm{c}1}$ is "closer" to $\boldsymbol{w}_1$ than to $\boldsymbol{\mu}$.

3.2.1 Connections to the work of Cadima and Jolliffe

The above results, obtained by confronting the covariance matrix and its non-centered counterpart, corroborate the work of Cadima and Jolliffe in [30]. In the latter, the eighth property in Proposition 3.1 gives a result equivalent to the above Lemma 11, however our proof is much shorter and significantly simpler than in [30, proof that spans nearly all the page 499]. Likewise, the ninth property in Proposition 3.1 is equivalent to Theorem 12, while our proof is slightly simpler.

Finally, Theorem 14 and Corollary 15 provide bounds that do not depend on the relation between $\boldsymbol{w}_1$ and $\boldsymbol{\mu}$, as opposed to the fourteenth property in Proposition 3.1 in [30]. In our case, we have more comprehensive expressions with simpler bounds, thus offering a straightforward interpretation.

4 Beyond conventional centering

In this section, we show that several research activities can take advantage of our study, apart from the conventional centering issue and beyond the scope of bridging the gap between PCA and ECA.

4.1 Weighted mean shift

The issue of centering the data has also been investigated with the use of a vector other than the mean of the data. Weighted means provide a generalization of the conventional mean. They have been commonly studied in the literature, such as in statistics with population studies [57]. More recently, a weighted mean is considered in [58] to derive a robust PCA algorithm. In [59], the authors study the use of a weighted mean in a k-nearest neighbor algorithm, in order to reduce hubs[6].

We propose to extend the our study to the issue of a weighted mean. Let $\boldsymbol{\omega}$ be a weight vector such as $\boldsymbol{\omega}^\top \mathbf{1} = 1$, and let $\boldsymbol{\mu}_{\boldsymbol{\omega}} = \boldsymbol{X}\boldsymbol{\omega}$ be the corresponding weighted mean. The matrix

$$\boldsymbol{P}_{\boldsymbol{\omega}} = \boldsymbol{\omega}\mathbf{1}^\top \tag{32}$$

maps the data such that their weighted mean becomes zero, with $\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{\omega}} = \boldsymbol{\mu}_{\boldsymbol{\omega}}\mathbf{1}^\top$. This matrix defines a projection map, since it is idempotent (*i.e.*, $\boldsymbol{P}_{\boldsymbol{\omega}}^2 = \boldsymbol{P}_{\boldsymbol{\omega}}$), but it is not necessary orthogonal. To define an orthogonal projection, the matrix needs to be symmetric, which means that $\boldsymbol{\omega} = \frac{1}{n}\mathbf{1}$ and therefore we get the particular case of the conventional mean studied so far.

As shown in the following, it turns out that this generalization of the projection can be easily studied with the analysis of the inner product matrices, as given in Section 3.1. Unfortunately, the analysis of the covariance matrix is no longer as easy as in Section 3.2. The main difficulty raises from the non-symmetric property of the matrix $\boldsymbol{P}_{\boldsymbol{\omega}}$ in the general case, due to the relaxation of the orthogonality in the projection.

Before proceeding, we revisit relations (20)–(23) in the light of this general definition, as follows:

$$\begin{aligned}(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{\omega}})^\top \boldsymbol{\omega} = \boldsymbol{\omega}^\top (\mathbf{I} - \boldsymbol{P}_{\boldsymbol{\omega}}) = \mathbf{0};\\ \boldsymbol{P}_{\boldsymbol{\omega}}^\top \boldsymbol{M} \boldsymbol{P}_{\boldsymbol{\omega}} = \mathbf{1}\boldsymbol{\omega}^\top \boldsymbol{M} \boldsymbol{\omega}\mathbf{1}^\top = n\,(\boldsymbol{\omega}^\top \boldsymbol{M}\boldsymbol{\omega})\,\boldsymbol{P}_{\mathbf{1}};\\ \mathrm{tr}(\boldsymbol{P}_{\boldsymbol{\omega}}^\top \boldsymbol{M}) = \mathrm{tr}(\boldsymbol{M}^\top \boldsymbol{P}_{\boldsymbol{\omega}}) = \tfrac{1}{n}\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{\omega}}^\top \boldsymbol{M} \boldsymbol{P}_{\boldsymbol{\omega}}) = \boldsymbol{\omega}^\top \boldsymbol{M}\mathbf{1}.\end{aligned}$$

By substituting $\boldsymbol{M}$ with $\boldsymbol{K}$ in the above expressions, we get $\boldsymbol{\omega}^\top \boldsymbol{K}\boldsymbol{\omega} = \boldsymbol{\omega}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\omega} = \|\boldsymbol{\mu}_{\boldsymbol{\omega}}\|^2$. Therefore, the definition of $\boldsymbol{K}_{\mathrm{c}}$ becomes

$$\boldsymbol{K}_{\mathrm{c}} = \boldsymbol{K} - \boldsymbol{P}_{\boldsymbol{\omega}}^\top \boldsymbol{K} - \boldsymbol{K}\boldsymbol{P}_{\boldsymbol{\omega}} + n\|\boldsymbol{\mu}_{\boldsymbol{\omega}}\|^2 \boldsymbol{P}_{\mathbf{1}}. \tag{33}$$

Lemma 1 becomes $\mathrm{tr}(\boldsymbol{K}_{\mathrm{c}}) = \mathrm{tr}(\boldsymbol{K}) - 2n\boldsymbol{\mu}_{\boldsymbol{\omega}}^\top \boldsymbol{\mu} + n\|\boldsymbol{\mu}_{\boldsymbol{\omega}}\|^2$, where we have used $\boldsymbol{\omega}^\top \boldsymbol{K}\mathbf{1} = \boldsymbol{\omega}^\top \boldsymbol{X}^\top \boldsymbol{X}\mathbf{1} = n\,\boldsymbol{\mu}_{\boldsymbol{\omega}}^\top \boldsymbol{\mu}$.

The analysis of the eigenvalues remains unchanged in the weighted mean case, including the interlacing property as given in Theorem 3. The only difference lies in the bounds proposed in Theorem 8. The general results are derived from expression (33) as follows:

$$\max_{i=1,\ldots,n} \lambda_i + \left(\|\boldsymbol{\mu}\|^2\, \boldsymbol{\alpha}_i^\top \mathbf{1} - 2\lambda_i\, \boldsymbol{\alpha}_i^\top \boldsymbol{\omega}\right) \boldsymbol{\alpha}_i^\top \mathbf{1} \quad \leq \quad \lambda_{\mathrm{c}1}.$$

It is also easy to verify that the eigenvectors satisfy $\boldsymbol{\alpha}_{\mathrm{c}j}^\top \boldsymbol{\omega} = 0$ for any non-zero eigenvalue.

The analysis of the covariance matrix is more complicated than the study derived in Section 3.2. This is due to the resulting expression of the covariance matrix in the general case of a weighted mean, with

$$\begin{aligned}\boldsymbol{C}_{\mathrm{c}} &= \tfrac{1}{n}\boldsymbol{X}(\mathbf{I} - \boldsymbol{P}_{\mathbf{1}})(\mathbf{I} - \boldsymbol{P}_{\mathbf{1}}^\top)\boldsymbol{X}^\top\\ &= \boldsymbol{C} - \boldsymbol{\mu}_{\boldsymbol{\omega}}\boldsymbol{\mu}^\top - \boldsymbol{\mu}\boldsymbol{\mu}_{\boldsymbol{\omega}}^\top + \boldsymbol{\mu}_{\boldsymbol{\omega}}\boldsymbol{\mu}_{\boldsymbol{\omega}}^\top.\end{aligned}$$

6. A hub is a sample that is very similar to many other samples of the dataset. Hubs emerge from the curse of dimensionality, and tend to be close to the data mean, *i.e.,* centroid. See [60] for more details on the concept of hubs in machine learning.

Still, one can derive several results. For instance, the lower bound in Theorem 12 becomes:

$$\max_{i=1,\ldots,n} \lambda_i - 2n\, \boldsymbol{w}_i^\top \boldsymbol{\mu}_{\boldsymbol{\omega}}\, \boldsymbol{w}_i^\top \boldsymbol{\mu} + n(\boldsymbol{w}_i^\top \boldsymbol{\mu}_{\boldsymbol{\omega}})^2 \leq \lambda_{\mathrm{c}1}.$$

4.2 Rank-one update of the covariance matrix

As given in expression (30), the matrix $\boldsymbol{C}_{\mathrm{c}}$ is a special case of the rank-one update of the matrix $\boldsymbol{C}$. It turns out that the study given in Sections 3.1 and 3.2 can be extended to any rank-one update of the covariance matrix. Such update is of great interest in covariance matrix adaptation within machines that rely on Gaussian random variations, such as evolutionary strategies [61] and ensemble optimization [62]. In [63], the author provides connections of these genetic machines to Monte Carlo-based methods, including particle filtering and population Monte Carlo. Without lost of generality[7], this section presents the issue of covariance matrix adaptation in evolutionary strategies [64]. See also [65], [66] for a survey.

The re-sampling techniques solve hard optimization problems by generating a set of candidate solutions. The performance highly depends on the population's distribution under investigation. Evolutionary strategies provide an elegant approach to derive (quasi) parameter-free techniques for the user. This principle of self-adaptation allows to adjust the distribution in the direction of more relevant regions in the search space.

Let $\boldsymbol{v}_t$ be a candidate solution generated from a zero-mean Gaussian distribution with covariance matrix $\boldsymbol{C}_t$. The latter is adapted according to the relevance of $\boldsymbol{v}_t$ in the optimization problem. Let $\nu_t \in]0,1[$ be a parameter that measures this relevance, where high values correspond to promising pertinent fitness progress. The rank-one update rule of the covariance matrix at iteration t is given by

$$\boldsymbol{C}_{t+1} = (1-\nu_t)\, \boldsymbol{C}_t + \nu_t\, \boldsymbol{v}_t \boldsymbol{v}_t^\top. \tag{34}$$

This rule allows to adjust the distribution towards the zero-mean Gaussian distribution with covariance matrix $\boldsymbol{v}_t \boldsymbol{v}_t^\top$, namely the distribution with the highest probability to generate $\boldsymbol{v}_t$ among all zero-mean Gaussian distributions.

We show next that one can take advantage of the mathematical statements presented in Section 3.2 in order to provide new insights to the update rule (34). Let $(\frac{1}{n}\lambda_{i,t}, \boldsymbol{w}_{i,t})$ be the i-th eigenpair of $\boldsymbol{C}_t$, namely

$$\boldsymbol{C}_t\, \boldsymbol{w}_{i,t} = \tfrac{1}{n}\lambda_{i,t}\, \boldsymbol{w}_{i,t}. \tag{35}$$

Firstly, the variance of the data can be measured with the sum of eigenvalues of the covariance matrix, which is given by its trace thanks to the relation (1). By following the same derivations as in Lemma 1, we get the following relation $\sum_{i=1}^n \lambda_{i,t+1} = (1-\nu_t)\sum_{i=1}^n \lambda_{i,t} + n\,\nu_t\,\|\boldsymbol{v}_t\|^2$. From expression (34), Lemma 11 becomes:

$$(\lambda_{j,t+1} - (1-\nu_t)\lambda_{i,t})\, \boldsymbol{w}_{i,t}^\top \boldsymbol{w}_{j,t+1} = n\,\nu_t\, \boldsymbol{w}_{i,t}^\top \boldsymbol{v}_t\ \boldsymbol{w}_{j,t+1}^\top \boldsymbol{v}_t. \tag{36}$$

Bounds on the eigenvalues of the covariance matrix can be easily derived, by following the same steps given in the proof of Theorem 12. Therefore, we have

$$\max_{i=1,\ldots,n} (1-\nu_t)\lambda_{i,t} + n\nu_t(\boldsymbol{w}_{i,t}^\top \boldsymbol{v}_t)^2 \quad \leq \quad \lambda_{1,t+1}.$$

In order to study the impact of the update rule on the eigenvectors of the covariance matrice, we revisit Theorems 13-14 and Corollary 15. From the above expression, we have for any $i = 1, 2, \ldots, n$:

$$\lambda_{1,t+1} - (1-\nu_t)\lambda_{i,t} \geq n\nu_t(\boldsymbol{w}_{i,t}^\top \boldsymbol{v}_t)^2.$$

By injecting expression (36) for $j = 1$, we get

$$\frac{n\,\nu_t\, \boldsymbol{w}_{i,t}^\top \boldsymbol{v}_t\ \boldsymbol{w}_{1,t+1}^\top \boldsymbol{v}_t}{\boldsymbol{w}_{i,t}^\top \boldsymbol{w}_{1,t+1}} \geq n\nu_t(\boldsymbol{w}_{i,t}^\top \boldsymbol{v}_t)^2.$$

By squaring and simplifying by $(\boldsymbol{w}_{i,t}^\top \boldsymbol{v}_t)^2$, we obtain

$$(\boldsymbol{w}_{i,t}^\top \boldsymbol{w}_{1,t+1})^2 \leq \frac{(\boldsymbol{w}_{1,t+1}^\top \boldsymbol{v}_t)^2}{(\boldsymbol{w}_{i,t}^\top \boldsymbol{v}_t)^2},$$

and equivalently

$$\cos(\boldsymbol{w}_{i,t}, \boldsymbol{w}_{1,t+1})^2 \leq \frac{\cos(\boldsymbol{w}_{1,t+1}, \boldsymbol{v}_t)^2}{\cos(\boldsymbol{w}_{i,t}, \boldsymbol{v}_t)^2}.$$

This bound shows that the first eigenvector of $\boldsymbol{C}_{t+1}$ forms a greater angle with all the eigenvectors of $\boldsymbol{C}_t$ than with the vector $\boldsymbol{v}_t$. This result, independent of the value of the parameter ν_t, illustrates the diversity introduced by applying the update rule (34).

4.3 Multidimensional scaling

Multidimensional scaling (MDS) is a well-known dimensionality reduction technique that seeks to preserve pairwise distances or dissimilarity measures [3]. The problem is to estimate all $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ from their available distances, denoted $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. By expanding this expression, we get $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 = \boldsymbol{x}_i^\top \boldsymbol{x}_i + \boldsymbol{x}_j^\top \boldsymbol{x}_j - 2\boldsymbol{x}_i^\top \boldsymbol{x}_j$. Therefore, one can define an inner product from distances, with $\boldsymbol{x}_i^\top \boldsymbol{x}_j = -\frac{1}{2}(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 - \|\boldsymbol{x}_i\|^2 - \|\boldsymbol{x}_j\|^2)$, or equivalently in matrix form

$$\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{\Delta} + \tfrac{1}{2}\boldsymbol{\delta}\mathbf{1}^\top + \tfrac{1}{2}\mathbf{1}\boldsymbol{\delta}^\top,$$

where $\boldsymbol{\Delta}$ is the matrix of entries $-\frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$ and $\boldsymbol{\delta}$ is the column vector whose i-th entry is $\|\boldsymbol{x}_i\|^2$. In order to remove the indeterminacy with respect to translation, the inner product of the centered data is considered. Thus, the double centering from (24) gives us:

$$\begin{aligned} \boldsymbol{K}_{\mathrm{c}} &= (\mathbf{I} - \boldsymbol{P}_1)\boldsymbol{X}^\top \boldsymbol{X}(\mathbf{I} - \boldsymbol{P}_1) \\ &= (\mathbf{I} - \boldsymbol{P}_1)(\boldsymbol{\Delta} + \tfrac{1}{2}\boldsymbol{\delta}\mathbf{1}^\top + \tfrac{1}{2}\mathbf{1}\boldsymbol{\delta}^\top)(\mathbf{I} - \boldsymbol{P}_1) \\ &= (\mathbf{I} - \boldsymbol{P}_1)\boldsymbol{\Delta}(\mathbf{I} - \boldsymbol{P}_1), \end{aligned} \tag{37}$$

7. For the sake of clarity, we examine the rank-one update of the covariance matrix. One could also study the rank-one update of the Gram matrix. In this case, simply replace $\boldsymbol{C}_t$, $\boldsymbol{w}_{i,t}$ and $\frac{1}{n}\lambda_{i,t}$, with $\boldsymbol{K}_t$, $\boldsymbol{\alpha}_{i,t}$ and $\lambda_{i,t}$.

where the last equality follows from (20). An eigendecomposition of this matrix provides the relevant axes to describe the samples. This is the classical MDS. Next, we study the MDS in the light of our work.

Expression (37) is similar to the double centering of the Gram matrix in (24). While, by construction, $\boldsymbol{K}_{\mathrm{c}}$ in both expressions is a Gram matrix, there is however one major difference: $\boldsymbol{\Delta}$ is not a positive definite matrix. It turns out that the corresponding function $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = -\frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$ is a conditionally positive definite kernel [67]. Next, we define this principle and give some properties, before studying its impact on the mathematical statements in this paper.

A conditionally positive definite kernel is a symmetric function that satisfies the inequality (11) for any $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}^\top \mathbf{1} = 0$. This is the case of $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = -\frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$, since we have for any $\boldsymbol{\beta}^\top \mathbf{1} = 0$:

$$\boldsymbol{\beta}^\top \boldsymbol{\Delta} \boldsymbol{\beta} = \boldsymbol{\beta}^\top (2\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{\delta}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\delta}^\top)\boldsymbol{\beta} = 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = 2\|\boldsymbol{X}\boldsymbol{\beta}\|^2 \geq 0.$$

In [67], the author provides a thorough description of conditionally positive definite kernels, and argues that they are "as good as" positive definite kernels whenever a translation invariant problem is investigated, such as in PCA and SVM. It is worth noting that one can include a positive bias b large enough such that $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) + b$ is positive definite, thus eliminating the term associated to the negative eigenvalue.

We return now to the main issue of this section, which is the analysis of the relations between the matrices $\boldsymbol{K}_{\mathrm{c}}$ and $\boldsymbol{\Delta}$. It turns out that one can take advantage of most of the mathematical statements derived in Section 3.1 for this purpose, as illustrated next by substituting $\boldsymbol{K}$ with $\boldsymbol{\Delta}$. To this end, we write expression (37) as follows:

$$\boldsymbol{K}_{\mathrm{c}} = \boldsymbol{\Delta} - \boldsymbol{P}_1 \boldsymbol{\Delta} - \boldsymbol{\Delta} \boldsymbol{P}_1 + \tfrac{\mathbf{1}^\top \boldsymbol{\Delta} \mathbf{1}}{n} \boldsymbol{P}_1.$$

This illustrates the analogy with (25), where

$$\mathbf{1}^\top \boldsymbol{\Delta} \mathbf{1} = -\frac{1}{2} \sum_{i,j=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2.$$

Moreover, since all diagonal entries of $\boldsymbol{\Delta}$ are null, then its trace is null, as well as the sum of its eigenvalues. Lemma 6 and Lemma 1 provide new insights. The former, namely the Schur-Horn Theorem given in Theorem 5 applied for the matrix $\boldsymbol{\Delta}$, shows that $0 \leq \sum_{i=1}^{t} \lambda_i$ for all $t = 1, 2, \ldots, n$, with equality for $t = n$, that is $\lambda_n = -\sum_{i=1}^{n-1} \lambda_i$. Lemma 1 shows the variability of the data with $\sum_{i=1}^{n} \lambda_{\mathrm{c}i} = -\frac{1}{n}\mathbf{1}^\top \boldsymbol{\Delta} \mathbf{1} = \frac{1}{2n}\sum_{i,j=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$.

An analysis of the distribution of the eigenvalues of each of $\boldsymbol{K}_{\mathrm{c}}$ and $\boldsymbol{\Delta}$ is given by considering once again the Separation Theorem given in Theorem 2. To this end, let $\boldsymbol{M} = \boldsymbol{\Delta}$, $\boldsymbol{P}_{\mathrm{left}} = \boldsymbol{P}_{\mathrm{right}} = (\mathbf{I} - \boldsymbol{P}_1)$, where $r(\boldsymbol{P}_{\mathrm{left}}) = r(\boldsymbol{P}_{\mathrm{right}}) = n - 1$, and thus $t = 2$. This leads to the following pair of inequalities:

$$\sigma_{j+2}(\boldsymbol{\Delta}) \leq \sigma_j(\boldsymbol{K}_{\mathrm{c}}) \leq \sigma_j(\boldsymbol{\Delta}).$$

The resulting inequalities are not as tight as the ones given in Theorem 3, due to the use of the decomposition $\boldsymbol{K}_{\mathrm{c}} = \boldsymbol{X}_{\mathrm{c}}^\top \boldsymbol{X}_{\mathrm{c}}$ in the latter case. Similar tight inequalities may be derived when one assumes that such decomposition is valid for $\boldsymbol{K}_{\mathrm{c}}$ defined in (37).

Bounds on the eigenvalues of $\boldsymbol{\Delta}$ and $\boldsymbol{K}_{\mathrm{c}}$ can be derived with the help of the Schur–Horn Theorem given in Theorem 5. By virtue of Lemma 7, one can describe results by following the same steps in the proof of Theorem 8. Thus, a lower bound on the largest eigenvalue of $\boldsymbol{K}_{\mathrm{c}}$ is given by

$$\max_{i=1,\ldots,n} \lambda_i + \left(\tfrac{\mathbf{1}^\top \boldsymbol{\Delta} \mathbf{1}}{n^2} - \tfrac{2}{n}\lambda_i\right) (\boldsymbol{\alpha}_i^\top \mathbf{1})^2 \quad \leq \quad \lambda_{\mathrm{c}1},$$

where $(\lambda_i, \boldsymbol{\alpha}_i)$ is an eigenpair of $\boldsymbol{\Delta}$.

When applied on noisy data in practice, the MDS technique considers the factorization of the resulting matrix $\boldsymbol{K}_{\mathrm{c}}$, such that $\boldsymbol{K}_{\mathrm{c}} = \boldsymbol{A}_{\mathrm{c}} \boldsymbol{\Lambda}_{\mathrm{c}} \boldsymbol{A}_{\mathrm{c}}^\top$ where only the largest non-negative eigenvalues are retained. From this expression, one defines a Gram matrix by setting $\boldsymbol{X}_{\mathrm{c}} = (\boldsymbol{\Lambda}_{\mathrm{c}})^{\frac{1}{2}} \boldsymbol{A}_{\mathrm{c}}^\top$. This construction leads to uncorrelated data, since

$$\boldsymbol{C}_{\mathrm{c}} = \tfrac{1}{n} \boldsymbol{X}_{\mathrm{c}} \boldsymbol{X}_{\mathrm{c}}^\top = \tfrac{1}{n} (\boldsymbol{\Lambda}_{\mathrm{c}})^{\frac{1}{2}} \boldsymbol{A}_{\mathrm{c}}^\top \boldsymbol{A}_{\mathrm{c}} (\boldsymbol{\Lambda}_{\mathrm{c}})^{\frac{1}{2}} = \tfrac{1}{n} \boldsymbol{\Lambda}_{\mathrm{c}},$$

and therefore, the analysis of the covariance matrix given in Section 3.2 is no longer required.

Scaling the data

We conclude this section by some interesting properties borrowed from [68] and naturally completes our work. We describe some interesting properties of a matrix $\boldsymbol{\Delta}$ obtained from the kernel $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = -\frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$. Consider scaling the data with some positive factor ξ. Since $\kappa(\xi \boldsymbol{x}_i, \xi \boldsymbol{x}_j) = \xi^2 \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$, then the corresponding matrix is $\boldsymbol{\Delta}_\xi = \xi^2 \boldsymbol{\Delta}$. From this relation, it is easy to see that both matrices $\boldsymbol{\Delta}$ and $\boldsymbol{\Delta}_\xi$ share the same eigenvectors, while any eigenvalue λ_i of the former defines the eigenvalue $\xi^2 \lambda_i$ of the latter. By considering the normalization given in (16) with $\|\boldsymbol{\alpha}_i\| = \frac{1}{\lambda_i}$, we obtain from (29): $\boldsymbol{X}^\top \boldsymbol{w}_j = \boldsymbol{\alpha}_j = \boldsymbol{X}_\xi^\top \boldsymbol{w}_{\xi j}$. Therefore, projections onto axes defined by either PCA or ECA provide scale-invariant features, as show here within the MDS approach. These results extends the work in [68] where only PCA is studies.

5 Experimental Results

All established mathematical statements can be easily verified. To show this, we consider the well-known *iris* dataset (available at the UCI Machine Learning Repository), which has been extensively studied in the pattern recognition literature since Fisher's seminal paper [6]. The dataset consists of 150 samples, divided equally into three classes, each sample having 4 attributes. TABLE 1 shows the interlacing property of the largest three eigenvalues of each Gram matrix, $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$, as settled in Theorem 3. Fig. 1 illustrates Theorem 8, where a lower bound on the largest t eigenvalues of $\boldsymbol{K}_{\mathrm{c}}$ is derived, while the specific cases of $t = 1$ and $t = n$ are shown.

We illustrate next the impact of data centering in kernel-based methods. To this end, we consider a set

TABLE 1
Illustration of the interlacing property of the eigenvalues of $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$. In these experiments, the values are obtained from the *iris* dataset.

	λ_5		λ_3		λ_2		λ_1
0.51	≤ 0.67	≤ 2.39	≤ 3.27	≤ 5.48	≤ 59.72	≤ 98.45	≤ 101.68
	$\lambda_{\mathrm{c}5}$		$\lambda_{\mathrm{c}3}$		$\lambda_{\mathrm{c}2}$		$\lambda_{\mathrm{c}1}$

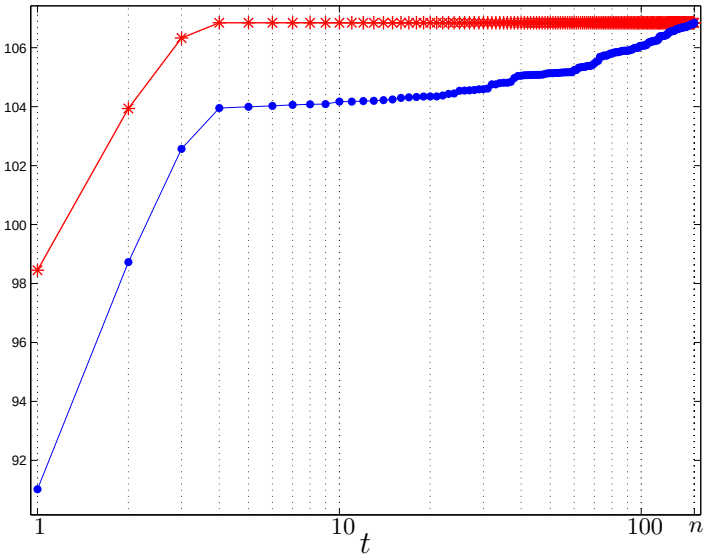

Fig. 1. Illustration of Theorem 8, in the case of the *iris* dataset with $n = 150$. For any $t = 1, 2, \ldots, n$, the cumulative sum $\sum_{i=1}^{t} \lambda_{\mathrm{c}i}$ (shown with $*$) is greater than the cumulative sum of $d'_i = \lambda_i + (\|\boldsymbol{\mu}\|^2 - \frac{2}{n}\lambda_i)\ (\boldsymbol{\alpha}_i^\top \mathbf{1})^2$ (shown with $\bullet$), given in non-increasing order. We see that, for $t = 1$, we have $\max_{i=1,\ldots,n} d'_i \leq \lambda_{\mathrm{c}1}$, while we get the equality $\sum_{i=1}^{t} d'_i = \sum_{i=1}^{t} \lambda_{\mathrm{c}i}$ when $t = n$.

of $n = 200$ two-dimensional data generated from a banana-shaped distribution, with $(x_j, y_j) = (\zeta_i, \zeta_i^2 + \xi)$ where ζ_i follows a uniform distribution on $[-1\ 1]$ and ξ follows a zero-mean Gaussian distribution with a 0.2 standard deviation. The Gram matrices were constructed by using the Gaussian kernel $\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$, where the bandwidth parameter was (naively) set to $\sigma = 0.5$. TABLE 2 illustrates the interlacing property of the largest eigenvalues of the two kernel matrices, the non-centered $\boldsymbol{K}$ with entries $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and the corresponding centered matrix $\boldsymbol{K}_{\mathrm{c}}$. Fig. 2 shows the contours of the first five principal functions, when data are centered (implicitly) in the feature space (first row), and when data is not centered (second row). This illustrates that the first principal function of the non-centered case is related to the data mean, while the other principal functions are similar to those obtained from the centered case, with one order higher, namely results from $(\lambda_{\mathrm{c}i}, \boldsymbol{\alpha}_{\mathrm{c}i})$ are comparable to results from $(\lambda_{i+1}, \boldsymbol{\alpha}_{i+1})$.

TABLE 2
Illustration of the interlacing property of the eigenvalues of the kernel matrices, $\boldsymbol{K}$ and $\boldsymbol{K}_{\mathrm{c}}$, where the Gaussian kernel is used. In these experiments, the values are obtained from the banana-shaped dataset.

	λ_5		λ_3		λ_2		λ_1
10.17	≤ 15.18	≤ 15.23	≤ 26.73	≤ 31.33	≤ 47.61	≤ 47.62	≤ 84.51
	$\lambda_{\mathrm{c}5}$		$\lambda_{\mathrm{c}3}$		$\lambda_{\mathrm{c}2}$		$\lambda_{\mathrm{c}1}$

6 Final Remarks

The main objective of this paper was to bridge the gap between centered and uncentered data. We studied the impact of centering data on the eigendecomposition of the Gram matrix, thereby of benefit to most kernel-based methods. To be more specific in this paper, we explored the eigendecomposition of the covariance matrix, with results that corroborate recent work on conventional PCA. Our key motivation was to reconcile the centered Gram matrix in PCA and the non-centered Gram matrix, such as with the ECA for nonparametric density estimation. Moreover, we provided several extensions of our main results, beyond the conventional centering issue.

Other techniques in manifold learning and dimensionality reduction can also take advantage of this work, include ISOMAP, locally-linear embedding, eigenmaps, and spectral clustering, to name a few. Further future work will address the issue of the impact of kernel functions in the centering issue, as well as the impact of centering the data in the input space, as opposed to the implicit centering in the feature space with kernel-based methods. We will also study connections with spectral analysis in random matrix theory [69].

References

[1] L. Sun, S. Ji, and J. Ye, "A least squares formulation for a class of generalized eigenvalue problems in machine learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, (New York, NY, USA), pp. 977–984, ACM, 2009.

[2] I. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer-Verlag, 1986.

[3] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling*. Monographs on Statistics and Applied Probability, London: Chapman and Hall / CRC, 2nd edition ed., September 2000.

[4] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica Chimica Acta*, vol. 185, no. 0, pp. 1 – 17, 1986.

[5] W. Härdle and L. Simar, "Canonical correlation analysis," in *Applied Multivariate Statistical Analysis*, pp. 321–330, Springer Berlin Heidelberg, 2007.

[6] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[7] M. Aizerman, E. Braverman, and L. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and Remote Control*, vol. 25, pp. 821–837, 1964.

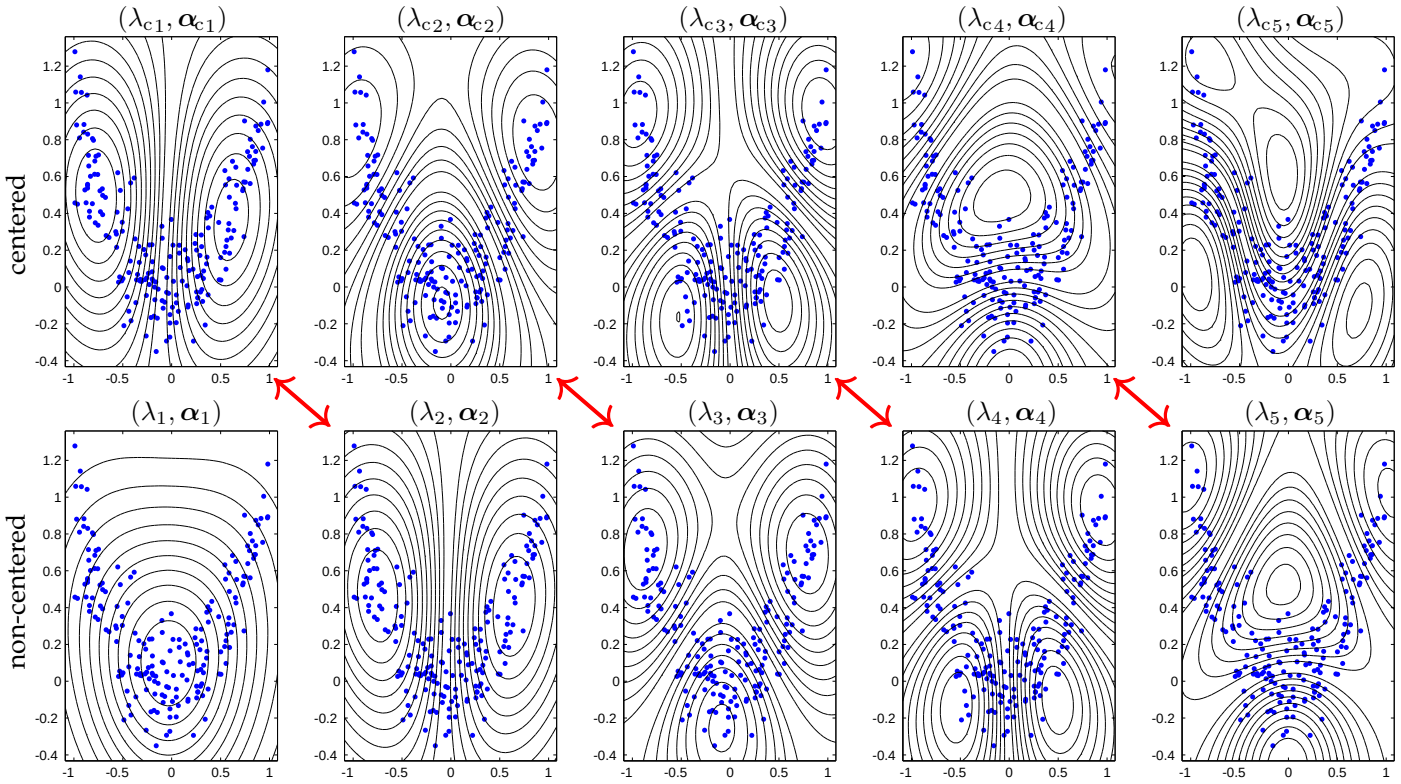

Fig. 2. Illustration of the contours of the first five principal functions from the kernel-based PCA with the Gaussian kernel, with $n = 200$ samples (given in blue dots $\cdot$) from a banana-shaped distribution. The first raw of figures is obtained with the eigendecomposition of the centered matrix $\boldsymbol{K}_{\mathrm{c}}$, while the second row corresponds to the non-centered case with $\boldsymbol{K}$.

[8] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," *Advances in kernel methods: support vector learning*, pp. 327–352, 1999.

[9] R. Rosipal and L. Trejo, "Kernel partial least squares regression in reproducing kernel Hilbert space," *Journal of Machine Learning Research*, vol. 2, pp. 97–123, 2002.

[10] K. Fukumizu, F. R. Bach, and A. Gretton, "Statistical consistency of kernel canonical correlation analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 361–383, Dec. 2007.

[11] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX* (Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, eds.), pp. 41–48, IEEE, 1999.

[12] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, UK: Cambridge University Press, 2004.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, Springer, second edition ed., 2009.

[15] P. Honeine, "Online kernel principal component analysis: a reduced-order model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1814–1826, September 2012.

[16] C.-D. Chang, C.-C. Wang, and B. Jiang, "Singular value decomposition based feature extraction technique for physiological signal analysis," *Journal of Medical Systems*, vol. 36, no. 3, pp. 1769–1777, 2012.

[17] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[18] M. Girolami, "Orthogonal series density estimation and the kernel eigenvalue problem," *Neural Computation*, vol. 14, pp. 669–688, 2002.

[19] R. Jenssen, "Kernel entropy component analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 847–860, 2010.

[20] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer Publishing Company, Incorporated, 1st ed., 2010.

[21] R. Jenssen, "Information theoretic learning and kernel methods," in *Information Theory and Statistical Learning* (F. Emmert-Streib and M. Dehmer, eds.), pp. 209–230, Springer US, 2009.

[22] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observations and Remote Sens.*, vol. 5, pp. 354–379, April 2012.

[23] K. Y. Yeung and W. L. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[24] J. Ruiz-del Solar and P. Navarrete, "Eigenspace-based face recognition: a comparative study of different approaches," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 35, no. 3, pp. 315–325, 2005.

[25] V. Deepu, S. Madhvanath, and A. Ramakrishnan, "Principal component analysis for online handwritten character recognition," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 2, pp. 327–330 Vol.2, 2004.

[26] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization.," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[27] P. Comon and C. Jutten, eds., *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, March 2010.

[28] A. Berman and R. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*. No. part 11 in Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 1994.

[29] R. A. Horn and C. R. Johnson, *Matrix analysis*. New York, NY,

USA: Cambridge University Press, 2nd edition ed., December 2012.
[30] J. Cadima and I. Jolliffe, "On relationships between uncentred and column-centred principal component analysis,," *Pakistan Journal of Statistics*, vol. 25, no. 4, pp. 473–503, 2009.
[31] T. V. Gestel, J. A. K. Suykens, G. Lanckriet, A. Lambrechts, B. D. Moor, and J. Vandewalle, "Bayesian framework for least-squares support vector machine classifiers, gaussian processes, and kernel fisher discriminant analysis," *Neural Comput.*, vol. 14, pp. 1115–1147, May 2002.
[32] C. H. Park and H. Park, "Nonlinear discriminant analysis using kernel functions and the generalized singular value decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 27, no. 1, pp. 87–102, 2005.
[33] C. Cortes, M. Mohri, and J. Weston, "A general regression technique for learning transductions," in *Proc. of the 22nd international conference on machine learning (ICML)*, (New York, NY, USA), pp. 153–160, ACM, 2005.
[34] I. Markovsky, "Matrix centering and low-rank approximation," tech. rep., Southampton, UK, 2009.
[35] S. W. Raudenbush, "Centering predictors in multilevel analysis: choices and consequences," *Multilevel Modelling Newsletter*, vol. 1, no. 2, pp. 10–12, 1989.
[36] N. T. Longford, "To center or not to center," *Multilevel Modelling Newsletter*, vol. 1, no. 3, pp. 7–11, 1989.
[37] D. A. Hofmann and M. B. Gavin, "Centering decisions in hierarchical linear models: Implications for research in organizations," *Journal of Management*, vol. 24, no. 5, pp. 623–641, 1998.
[38] O. Paccagnella, "Centering or not centering in multilevel models? the role of the group mean and the assessment of group effects," *Evaluation Review*, vol. 30, no. 1, pp. 66–85, 2006.
[39] I. Kreft, I. Kreft, and J. de Leeuw, *Introducing Multilevel Modeling*. ISM (London, England), SAGE Publications, 1998.
[40] O. Papaspiliopolous, G. O. Roberts, and M. Skôld, "Non-centered parameterizations for hierarchical models and data augmentation," in *Baysian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, eds.), pp. 307–326, Oxford University Press, 2003.
[41] Y. Yu and X.-L. Meng, "To center or not to center: That is not the question - an ancillarity-sufficiency interweaving strategy (asis) for boosting mcmc efficiency," *Journal of Computational and Graphical Statistics*, vol. 20, no. 3, pp. 531–570, 2011.
[42] O. E. Barndorff-Nielsen and N. Shephard, "Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics," *Econometrica*, vol. 72, pp. 885–925, May 2004.
[43] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," *Royal Society of London Philosophical Transactions Series A*, vol. 209, pp. 415–446, 1909.
[44] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, 1950.
[45] P. Honeine and C. Richard, "Preimage problem in kernel-based machine learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 77–88, 2011.
[46] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, July 1998.
[47] D. M. J. Tax and P. Juszczak, "Kernel whitening for one-class classification," in *Proc. of the First International Workshop on Pattern Recognition with Support Vector Machines*, (London, UK), pp. 40–52, Springer-Verlag, 2002.
[48] E. Parzen, "On estimation of a probability density function and mode," *Annals of Mathematical Statistics*, vol. 33, pp. 1065–1076, 1962.
[49] A. Izenman, "Recent developments in nonparametric density estimation," *Journal of the American Statistical Association*, vol. 86, no. 413, pp. 205–224, 1991.
[50] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.
[51] H. Yanai, K. Takeuchi, and Y. Takane, *Projection Matrices, Generalized Inverse Matrices, and Singular Value Decomposition*. New York, NY: Springer New York, 2011.
[52] Y. Takane and T. Shibayama, "Principal component analysis with external information on both subjects and variables," *Psychometrika*, vol. 56, no. 1, pp. 97–120, 1991.
[53] C. Rao, "Separation theorems for singular values of matrices and their applications in multivariate analysis," *Journal of Multivariate Analysis*, vol. 9, no. 3, pp. 362 – 377, 1979.
[54] C. R. Rao, "Matrix approximations and reduction of dimensionality in multivariate statistical analysis," *Multivariate Analysis V*, vol. 5, pp. 3–22, 1980.
[55] A. Horn, "Doubly stochastic matrices and the diagonal of a rotation matrix," *American Journal of Mathematics*, vol. 76, pp. 620–630, 1954.
[56] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities : Theory of Majorization and its Applications*. New York: Springer Science+Business Media, LLC, 2011.
[57] P. Meier, "Variance of a weighted mean," *Biometrics*, vol. 9, no. 1, pp. pp. 59–73, 1953.
[58] I. Higuchi and S. Eguchi, "Robust principal component analysis with adaptive selection for tuning parameters," *J. Mach. Learn. Res.*, vol. 5, pp. 453–471, Dec. 2004.
[59] I. Suzuki, K. Hara, M. Shimbo, M. Saerens, and K. Fukumizu, "Centering similarity measures to reduce hubs," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 613–623, ACL, 2013.
[60] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Hubs in space: Popular nearest neighbors in high-dimensional data," *J. Mach. Learn. Res.*, vol. 11, pp. 2487–2531, Dec. 2010.
[61] T. Suttorp, N. Hansen, and C. Igel, "Efficient covariance matrix update for variable metric evolution strategies," *Machine Learning*, vol. 75, no. 2, pp. 167–197, 2009.
[62] R. Fonseca, O. Leeuwenburgh, P. v. d. Hof, and J. Jansen, "Improving the ensemble optimization method through covariance matrix adaptation (CMA-EnOpt)," in *SPE Reservoir Simulation Symposium 2013*, (The Woodlands, TX), pp. 1124–1133, Society of Petroleum Engineers, February 2013.
[63] P. Djuric, "From nature to methods and back to nature," in *Signal Processing and Multimedia Applications (SIGMAP), Proceedings of the 2010 International Conference on*, pp. IS–7–IS–7, 2010.
[64] C. L. Müller and I. F. Sbalzarini, "Gaussian adaptation revisited: An entropic view on covariance matrix adaptation," in *Proceedings of the 2010 International Conference on Applications of Evolutionary Computation - Volume Part I*, EvoApplicatons'10, (Berlin, Heidelberg), pp. 432–441, Springer-Verlag, 2010.
[65] S. Meyer-Nieberg and H.-G. Beyer, "Self-adaptation in evolutionary algorithms," in *Parameter Setting in Evolutionary Algorithm*, pp. 47–76, Springer, 2006.
[66] O. Kramer, "Evolutionary self-adaptation: a survey of operators and strategy parameters," *Evolutionary Intelligence*, vol. 3, no. 2, pp. 51–65, 2010.
[67] B. Schölkopf, "The kernel trick for distances.," in *NIPS* (T. K. Leen, T. G. Dietterich, and V. Tresp, eds.), pp. 301–307, MIT Press, 2000.
[68] H. Sahbi, "Kernel pca for similarity invariant shape recognition," *Neurocomput.*, vol. 70, pp. 3034–3045, Oct. 2007.
[69] Z. Bai and J. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics, Springer, 2009.

Paul Honeine (M'07) was born in Beirut, Lebanon, on October 2, 1977. He received the Dipl.-Ing. degree in mechanical engineering in 2002 and the M.Sc. degree in industrial control in 2003, both from the Faculty of Engineering, the Lebanese University, Lebanon. In 2007, he received the Ph.D. degree in Systems Optimisation and Security from the University of Technology of Troyes, France, and was a Postdoctoral Research associate with the Systems Modeling and Dependability Laboratory, from 2007 to 2008. Since September 2008, he has been an assistant Professor at the University of Technology of Troyes, France. His research interests include nonstationary signal analysis and classification, nonlinear signal processing, sparse representations, machine learning, and wireless sensor networks. He is the co-author (with C. Richard) of the 2009 Best Paper Award at the IEEE Workshop on Machine Learning for Signal Processing.