Neural Process Lett (2008) 27:197–207
DOI 10.1007/s11063-007-9069-2

New Routes from Minimal Approximation Error to Principal Components

Abhilash Alexander Miranda · Yann-Aël Le Borgne · Gianluca Bontempi

Published online: 5 January 2008

Abstract We introduce two new methods of deriving the classical PCA in the framework of minimizing the mean square error upon performing a lower-dimensional approximation of the data. These methods are based on two forms of the mean square error function. One of the novelties of the presented methods is that the commonly employed process of subtraction of the mean of the data becomes part of the solution of the optimization problem and not a pre-analysis heuristic. We also derive the optimal basis and the minimum error of approximation in this framework and demonstrate the elegance of our solution in comparison with a recent solution in the framework.

Keywords Principal components analysis · Eigenvalue · Matrix trace

1 Introduction

The problem of approximating a given set of data using a weighted linear combination of a fewer number of vectors than the original dimensionality is classic. Many applications that require such a dimensionality reduction desire that the new representation retain the maximum variability in the data for further analysis. A popular method that attains simultaneous dimensionality reduction, minimum mean square error of approximation and retainment of maximum variance of the original data representation in the new representation is called the Principal Components Analysis (PCA) [7,11].

The most popular framework for deriving PCA starts with the analysis of variance. A very common derivation of PCA in this framework generates the basis by iteratively finding the orthogonal directions of maximum retained variances [7,10,11,14]. Since variance is implied in the statement of the problem here, the mean is subtracted from the data as a preliminary step. The second most predominant framework derives PCA by minimizing the

A. A. Miranda (✉) · Y.-A. Le Borgne · G. Bontempi
Machine Learning Group, Département d'Informatique, Université Libre de Bruxelles, Boulevard du Triomphe – CP212, 1050 Brussels, Belgium
e-mail: abalexan@ulb.ac.be

mean square error of approximation [1–3]. Aided by the derivation in the variance-based framework above, it has become acceptable to resort to mean subtraction of the data prior to any analysis in this framework too in order to keep the analysis simple. In this letter our focus is on the latter framework within which we demonstrate two distinct and elegant analytical methods of deriving the PCA. In each of these methods of derivation, subtraction of data mean becomes part of the solution instead of being an initial assumption.

The letter is organized as follows: in Sect. 2 we describe the motivation behind the need for yet another derivation of the classical PCA. In particular, we highlight the issue of mean centering in Sect. 2.1. The notations are introduced in Sect. 2.2 and the PCA problem and its interpretations are discussed in Sect. 3. After reviewing a recent solution in Sect. 4, we make it evident in Sect. 5 that our two methods are due to two forms of the optimization function. Then we introduce these two methods of solving the PCA problem in Sects. 6 and 7 and arrive at a simple common form of the optimization function in both these methods. This is analyzed further in Sect. 8 where we show the relation of the variance to the optimal basis as well as the minimum approximation error attained in PCA. In Sect. 8.3, we revisit a recent solution in our framework of PCA introduced in Sect. 4 and equate it with our approach.

2 Motivation

There are many standard textbooks of multivariate and statistical analysis [10,11,14] detailing PCA as a technique that seeks the best approximation of a given set of data points using a linear combination of a set of vectors which retain maximum variance along their directions. Since this framework of PCA starts by finding the covariances, the mean has to be subtracted from the data and becomes the *de facto* origin of the new coordinate system. The subsequent analysis is simple: find the eigenvector corresponding to the largest eigenvalue of the covariance matrix as the first basis vector. Then find the second basis vector on which the data components bear zero correlation with the data components on the first basis vector. This turns out to be the eigenvector corresponding to the second largest eigenvalue. In successively finding the basis vectors that have uncorrelated components as the eigenvectors of decreasing retained variances, the second order cross moments between the components are successively eliminated.[1] Computationally, a widely employed trick in this framework finds the eigenvectors using singular value decomposition of the mean centered data matrix which effectively diagonalizes the covariance matrix without actually computing it [11,14]. The set of orthogonal vectors corresponding to the largest few singular values proportional to the variances yields those directions which retain the maximum variance in the new representation of the data.

The second framework derives the PCA approximation by using its property of minimizing the mean square error. We think that this framework is more effective in introducing PCA to a novice because the two outcomes of optimal dimensionality reduction, viz. error minimization and retained variance maximization, are attained here simultaneously. Following the path of the retained variance maximization framework and to keep the analysis simple, many textbooks [2,9,10,20] advocate a mean subtraction for this framework too without sensible justification. Pearson stated in his now classical paper [18]:

> "The second moment of a system about a series of parallel lines is always least for the line going through the centroid. Hence: The best-fitting straight line for a system of points in a space of any order goes through the centroid of the system."

[1] Elimination of higher order cross moments is dealt in Independent Components Analysis (ICA) [9].

A procedure equivalent to rephrasing of this statement is followed in a much referenced textbook [3] which reasons that since the mean is the zero-dimensional hyperplane which satisfies the minimum average square error criterion, any higher dimensional hyperplane should be excused to pass through it too. In order to keep our analysis coherent with the concept of simultaneous dimensionality reduction, retained variance maximization and approximation error minimization, we do not invite the reader to such geometric intuitions. Note that the error minimization framework can also be viewed as a total least squares regression problem with all variables thought to be free so that the task is to fit a lower dimensional hyperplane that minimizes the perpendicular distances from the data points to the hyperplane [21].

We will also be reviewing [1] who derives PCA in the same framework as that of ours. Unlike in their approach, we neither undertake a complete orthogonal decomposition nor force any basis vectors to bear a common statistic enticed by the prospect of an eventual mean subtraction. Also for the benefit of practitioners who would like to deal data as realizations of a random variable, our treatment in the data samples domain can be readily extended to a population domain.

2.1 To Mean Center or Not

In the framework of finding the basis of a lower dimensional space which minimizes the mean square error of approximation, the process of mean subtraction has so far been part of the heuristics that the data needs to be centered before installing the new low-dimensional coordinate system motivated by the philosophy according to [18] that, had the mean of the data not been subtracted, the best fitting hyperplane would pass through the origin and not through the centroid. But there exist situations where a hyperplane is merely expected to partition the data space into orthogonal subspaces and as a result subtraction of mean is not desired. Note that in such situations, the term ‘principal component’ does not strictly hold as the basis vectors for the new space are not obtained from the data covariance matrix and the main concern there is the decomposition of the data rather than its approximation.

One such set of situations are addressed by the Fukunaga–Koontz Transform [5,16] and it works by not requiring a subtraction of mean but instead finds the principal components of the autocorrelation matrices of two classes of data. It is widely used in automatic target recognition where eigenvalue decomposition generates basis for a target space orthogonal to the clutter space. But such is the issue of mean subtraction in using this transform that researchers of [12] and [8] use autocorrelation and covariance matrices, respectively, for the same task without a justification of the impact of their choice to mean center or not. A similar approach called Eigenspace Separation Transformation [19] aimed at classification also does not involve mean subtraction. A family of techniques called Orthogonal Subspace Projection that is widely applied in noise rejection of signals use data that are not mean centered for the generalized PCA that follows [6].

Although the theory of PCA demands mean subtraction for optimal low dimensional approximation, for many applications it is not without consequence. For example, the researchers of ecology and climate studies have extensively debated the purpose and result of mean centering for their PCA-based data analysis. In [17], the characteristics and apparent advantages of the principal components generated without mean subtraction are compared for data sampled homogenously in the original space or otherwise. The claim made therein is that if data form distinct clusters, the influence of variance within a cluster on another can be minimized by not subtracting the mean. Another ongoing debate named ‘Hockey Stick’ controversy [15] involves the appropriateness of mean subtraction for PCA in a much cited global warming study [13].

It should be borne in mind that this letter is neither solely about the aforementioned issue of mean centering that researchers using PCA often take it for granted nor does it change the results of PCA that is previously known to them. But we demonstrate in a new comprehensive framework that (i) the mean subtraction becomes a solution to the optimization problem in PCA and we reach this solution through two simple distinct methods that borrow little from traditional textbook derivations of PCA, and (ii) the derivation of the basis for the low dimensional space converges to minimum approximation error and maximum retained variance in the framework. Consequently, we believe that many problems which raise questions about their choice regarding mean subtraction can be revisited with ease using our proposed PCA framework.

2.2 Notations

The notations that will be used throughout this letter are summarized in the table below

J_q	:	error function
q	:	new dimensionality
p	:	original dimensionality
n	:	number of samples
x_k	$\in$	$\mathbb{R}^p$; k^{th} data sample
$\hat{x}_k$	$\in$	$\mathbb{R}^p$; approximation of x_k
θ	$\in$	$\mathbb{R}^p$; new general origin
$\tilde{x}_k$	$=$	$x_k - \theta \in \mathbb{R}^p$
e_i	$\in$	$\mathbb{R}^p$; i^{th} orthonormal basis vector of $\mathbb{R}^p$
W	$=$	$[e_1 \cdots e_q] \in \mathbb{R}^{p\times q}$
B	$=$	$I - WW^T \in \mathbb{R}^{p\times p}$
$\widetilde{W}$	$=$	$[e_{q+1} \cdots e_p] \in \mathbb{R}^{p\times p-q}$
z_k	$\in$	$\mathbb{R}^q$; dependent on x_k
b	$\in$	$\mathbb{R}^{p-q}$; a constant
$\mathsf{Tr}(A)$	:	Trace of the matrix A
$\mathsf{rank}(A)$	:	Rank of the matrix A
μ	$\in$	$\mathbb{R}^p$; sample mean
S	$\in$	$\mathbb{R}^{p\times p}$; sample covariance matrix
λ_i	:	i^{th} largest eigenvalue of S
r	$=$	rank(S)

3 Problem Definition in the Sample Domain

Let $x_k \in \mathbb{R}^p, k = 1, \dots, n$ be a given set of data points. Suppose we are interested in orthonormal vectors $e_i \in \mathbb{R}^p, i = 1, \dots, q \le p$ whose resultant of weighted linear combination $\hat{x}_k \in \mathbb{R}^p$ can approximate x_k with a minimum average (sample mean) square error or in other words minimize

$$J_q(\hat{x}_k) = \frac{1}{n}\sum_{k=1}^{n} \|x_k - \hat{x}_k\|^2. \tag{1}$$

The problem stated above means that we need an approximation $x_k \simeq \hat{x}_k$ such that

$$\hat{x}_k = \sum_{i=1}^{q} \left(e_i^T x_k\right) e_i \tag{2}$$

so that we attain the minimum for J_q. This approximation assumes that the origin of all orthonormal e_i is the same as that of the coordinate system in which the data is defined. We assume orthonormality here because (i) orthogonality guarantees linearly independent e_i so that they form a basis for $\mathbb{R}^q$ (ii) normalizing e_i maintains notational simplicity in not having to divide the scalars $e_i^T x_k$ in (2) by the norm $\|e_i\|$ which is unity due to our assumption.

We reformulate the approximation

$$\hat{x}_k = \theta + \sum_{i=1}^{q} \left(e_i^T (x_k - \theta)\right) e_i \tag{3}$$

to assume that the new representation using basis vectors e_i has a general origin $\theta \in \mathbb{R}^p$ and not the origin as in the approximation (2). Hence, the PCA problem may be defined as

$$\underset{e_i,\theta}{\text{argmin}} \; \frac{1}{n} \sum_{k=1}^{n} \|x_k - \hat{x}_k\|^2 \; : \; \begin{array}{l} \hat{x}_k = \theta + \sum_{i=1}^{q} \left(e_i^T \left(x_k - \theta\right)\right) e_i; \\ e_i^T e_j = 0, \; i \neq j; \;\; e_i^T e_i = 1 \; \forall i, j. \end{array} \tag{4}$$

which seeks a set of orthonormal basis vectors e_i with a new origin θ which minimizes the error function in (1) in order to find a low-dimensional approximation $W^T(x_k - \theta) \in \mathbb{R}^q$ for any $x_k \in \mathbb{R}^p$, where

$$W = [e_1 \cdots e_q]. \tag{5}$$

It is now easy to see that (3) becomes

$$\hat{x}_k = \theta + WW^T (x_k - \theta). \tag{6}$$

Hence the displacement vector directed from the approximation $\hat{x}_k$ towards x_k is $x_k - \hat{x}_k = (x_k - \theta) - WW^T (x_k - \theta)$, which using $\tilde{x}_k = x_k - \theta$ can be written concisely as $x_k - \hat{x}_k = \tilde{x}_k - WW^T \tilde{x}_k$. By setting $B = I - WW^T$ for simplicity of notation, we write the displacement vector as

$$x_k - \hat{x}_k = B\tilde{x}_k. \tag{7}$$

4 Review of a Recent Solution

The most recent PCA solution in the framework of approximation error minimization, derived in [1], is reviewed here. They derive PCA by undertaking a complete decomposition

$$\hat{x}_k = Wz_k + \widetilde{W}b \tag{8}$$

into basis vectors contained in the columns of matrix W of (5) and $\widetilde{W} = [e_{q+1} \cdots e_p] \in \mathbb{R}^{p \times p-q}$ such that components of $z_k \in \mathbb{R}^q$ depend on x_k, whereas components of $b \in \mathbb{R}^{p-q}$ are constants common for all data points.

By taking the derivative of the error function with respect to b, they find that

$$b = \widetilde{W}^T \mu \tag{9}$$

so that the common components are those of the sample mean vector μ. This implies that by subtracting the sample mean they are no longer obliged to retain the $p - q$ dimensions corresponding to the columns of $\widetilde{W}$ which preserve little information regarding the variation in the data. The first drawback of this approach is that it couples the process of dimensionality reduction with mean subtraction although the two will be shown to be independent in our derivation. By taking the derivative of the error function with respect to z_k, they also show that $z_k = W^T x_k$. Hence the approximation they are seeking is

$$\hat{x}_k = WW^T x_k + \widetilde{W}\widetilde{W}^T \mu. \tag{10}$$

The second drawback of their approach is the requirement of yet another constrained minimization of the error function before they reach the solution for the optimal columns of W.

5 Methods of PCA

We have discussed the need for a new derivation of PCA by (i) explaining the lack of proper justification in the literature for subtracting the mean in a minimum mean square error framework, (ii) reminding its chronic necessity for the benefit of many applications in Sect. 2, and (iii) reviewing a recent attempt to solve this problem in Sect. 4. Our derivations of the solution for the problem in (4) are due to two simple forms of the error function J_q of (1) which we state as follows:

$$\text{Form 1}: J_q(\hat{x}_k) = \frac{1}{n}\sum_{k=1}^{n}\left(x_k - \hat{x}_k\right)^T\left(x_k - \hat{x}_k\right) \tag{11}$$

$$\text{Form 2}: J_q(\hat{x}_k) = \mathrm{Tr}\left(\frac{1}{n}\sum_{k=1}^{n}\left(x_k - \hat{x}_k\right)\left(x_k - \hat{x}_k\right)^T\right) \tag{12}$$

We analyze Form 1 in (11) in Sect. 6 to arrive at a simplified J_q which is exactly the same as we get by following a different method of analyzing Form 2 in (12) in Sect. 7. These two methods pursue different paths towards the common error function, viz., the first using straightforward expansion of the terms in J_q and the second using the property of matrix trace. The common form of J_q is subsequently treated in Sect. 8 to reveal the rest of the solution to our original problem.

6 Analysis of Form 1 of Error Function

Using (7), the error function J_q of Form 1 in (11) can be developed as

$$J_q(B,\theta) = \frac{1}{n}\sum_{k=1}^{n}\tilde{x}_k^T B^T B\tilde{x}_k. \tag{13}$$

The property that $B = I - WW^T$ is idempotent and symmetric, i.e.,

$$B = B^2 = B^T, \tag{14}$$

or B is simply an orthogonal projector, may be used to reduce J_q further as

$$J_q(B,\theta) = \frac{1}{n}\sum_{k=1}^{n}\tilde{x}_k^T B\tilde{x}_k. \tag{15}$$

Expanding J_q above using $\tilde{x}_k = x_k - \theta$ gives

$$J_q(B,\theta) = \frac{1}{n}\sum_{k=1}^{n}\left[x_k^T Bx_k - 2\theta^T Bx_k + \theta^T B\theta\right] \tag{16}$$

In order to get the θ which minimizes J_q, we find the partial derivative $\partial J_q/\partial\theta = -2B\left[\frac{1}{n}\sum_{k=1}^{n} x_k - \theta\right]$ and setting it to zero results in

$$\theta = \frac{1}{n}\sum_{k=1}^{n} x_k = \mu, \tag{17}$$

which is as simple as regarding the sample mean of the data points as the new origin. Henceforth, we can assume that $\tilde{x}_k$ is the data point x_k from which the sample mean has been subtracted.

6.1 Simplifying the Error Function

We may analyze the error function in (15) as follows:

$$\begin{aligned} J_q(W) &= \frac{1}{n}\sum_{k=1}^{n} \tilde{x}_k^T \left(I - WW^T\right)\tilde{x}_k \\ &= \frac{1}{n}\sum_{k=1}^{n} \tilde{x}_k^T \tilde{x}_k - \frac{1}{n}\sum_{k=1}^{n} \tilde{x}_k^T WW^T \tilde{x}_k \\ &= \frac{1}{n}\sum_{k=1}^{n} \tilde{x}_k^T \tilde{x}_k - \mathrm{Tr}\left(W^T \left[\frac{1}{n}\sum_{k=1}^{n} \tilde{x}_k \tilde{x}_k^T\right] W\right). \end{aligned}$$

We have the sample covariance matrix

$$S = \frac{1}{n}\sum_{k=1}^{n} \tilde{x}_k \tilde{x}_k^T \ |_{\theta=\mu} \tag{18}$$

so that the term $\frac{1}{n}\sum_{k=1}^{n} \tilde{x}_k^T \tilde{x}_k \ |_{\theta=\mu}$ equals $\mathrm{Tr}(S)$, and we can write

$$J_q(W) = \mathrm{Tr}(S) - \mathrm{Tr}\left(W^T S W\right). \tag{19}$$

7 Analysis of Form 2 of Error Function

We now analyze the Form 2 of the error function J_q by substituting (7) in (12) as

$$J_q(B,\theta) = \mathrm{Tr}\left(B\left[\frac{1}{n}\sum_{k=1}^{n} \tilde{x}_k \tilde{x}_k^T\right] B^T\right). \tag{20}$$

7.1 Finding θ

As in the previous section, we denote the sample mean and sample covariance matrix by μ and S, respectively, and we may develop the term in (20):

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\tilde{x}_k\tilde{x}_k^T &= \frac{1}{n}\sum_{k=1}^{n}(x_k-\theta)(x_k-\theta)^T\\ &= \frac{1}{n}\sum_{k=1}^{n}\left[x_kx_k^T - x_k\theta^T - \theta x_k^T + \theta\theta^T\right]\\ &= S+\mu\mu^T-\mu\theta^T-\theta\mu^T+\theta\theta^T, \end{aligned} \tag{21}$$

where we have used the sample autocorrelation matrix [4] given by $\frac{1}{n}\sum_{k=1}^{n} x_kx_k^T = S+\mu\mu^T$. We get $J_q(B) = \mathrm{Tr}\left(B\left(S+\mu\mu^T - \mu\theta^T - \theta\mu^T+\theta\theta^T\right)B^T\right)$ upon substituting (21) in (20). Using (14) and the cyclic permutation property of trace of matrix products[2] we get

$$J_q(B) = \mathrm{Tr}\left(B\left(S+\mu\mu^T-\mu\theta^T-\theta\mu^T+\theta\theta^T\right)\right) \tag{22}$$

and using the property of the derivative of trace[3] and the chain rule of the derivatives,[4] we find that $\partial J_q/\partial\theta = 2B\,(-\mu+\theta)$ which when equated to zero results in

$$\theta = \mu \tag{23}$$

leading to the same solution of Form 1 in (17).

7.2 Simplifying the Error Function

Having found θ, we can substitute it in (22) to get $J_q(B) = \mathrm{Tr}\,(BS)$. On substitution for B in terms of W, we may write $J_q(W) = \mathrm{Tr}\,(S) - \mathrm{Tr}\left(WW^TS\right)$. Utilizing the cyclic permutation property of matrix trace again, we get

$$J_q(W) = \mathrm{Tr}\,(S) - \mathrm{Tr}\left(W^TSW\right). \tag{24}$$

8 Optimal Basis and Minimum Error

Note that we have arrived at the same set of equations in both (19) and (24) of Form 1 and Form 2, respectively, whereby substituting W as defined in (5) in either of them gives

$$J_q(e_i) = \mathrm{Tr}(S) - \sum_{i=1}^{q} e_i^T S e_i. \tag{25}$$

8.1 Relation of Variance to Optimal Basis

Let us now find the variance λ_i of the data projected on the basis vector e_i. It is the average of the square of the difference between projections $e_i^T x_k$ of the data points and the projection

[2] $\mathrm{Tr}\,(\Upsilon\Phi\Psi) = \mathrm{Tr}\,(\Psi\Upsilon\Phi) = \mathrm{Tr}\,(\Phi\Psi\Upsilon)$.

[3] $\partial\left(\mathrm{Tr}\left(\Psi\Phi^T\right)\right)\Big/\partial\Phi = \Psi$.

[4] $\partial(\cdot)/\partial u = \left[\partial(\cdot)/\partial\left(uv^T\right)\right]v$.

$e_i^T \mu$ of the sample mean, i.e.,

$$\begin{aligned}\lambda_i &= \frac{1}{n}\sum_{k=1}^{n}\left(e_i^T x_k - e_i^T \mu\right)^2 \\ &= \frac{1}{n}\sum_{k=1}^{n}\left(e_i^T x_k - e_i^T \mu\right)\left(e_i^T x_k - e_i^T \mu\right)^T \\ &= e_i^T\left[\frac{1}{n}\sum_{k=1}^{n}(x_k-\mu)(x_k-\mu)^T\right]e_i \\ &= e_i^T S e_i. \end{aligned} \tag{26}$$

Thus, the term $\sum_{i=1}^{q} e_i^T S e_i$ in (25) gives the portion of the total variance $\mathsf{Tr}(S)$ retained along the directions of orthonormal e_i. Hence, we are looking for vectors e_i of the form $\lambda_i = e_i^T (S e_i)$, which is satisfied if $S e_i = \lambda_i e_i$. Such a relation implies (e_i, λ_i) form an eigen-pair of S. Note that since there is no unique basis for any nontrivial vector space, any basis that spans the q-dimensional space generated by the eigenvectors of S are solutions to e_i too. In (25), since

$$\underset{e_i}{\operatorname{argmin}}\, J_q = \underset{e_i}{\operatorname{argmax}} \sum_{i=1}^{q} e_i^T S e_i, \tag{27}$$

the vectors e_i have to be the eigenvectors corresponding to the q largest (‘principal’) eigenvalues of S. This is the classical result of the PCA.

8.2 Relation of Variance to Minimum Approximation Error

It follows from (26) that the term $\sum_{i=1}^{q} e_i^T S e_i = \sum_{i=1}^{q} \lambda_i$ of (25) is the sum of the q principal eigenvalues of S; this is the maximum variance that could be retained upon approximation using any q basis vectors. Also, $\mathsf{Tr}(S) = \sum_{i=1}^{r} \lambda_i$, $r = \mathsf{rank}(S)$ is the total variance in the data. Substituting these in J_q in (25) gives the difference of the total variance and the maximum retained variance; the result is the minimum of the eliminated variance. Hence, for $\lambda_i \geq \lambda_j$, $j > i$, the minimum mean square approximation error can be expressed as

$$J_q = \underbrace{\sum_{i=1}^{r} \lambda_i}_{\textit{total variance}} - \underbrace{\sum_{i=1}^{q} \lambda_i}_{\textit{retained variance}} = \underbrace{\sum_{i=q+1}^{r} \lambda_i}_{\textit{eliminated variance}}. \tag{28}$$

8.3 Comparison of the Reviewed Solution with the Present Work

In order to compare the solution of [1] reviewed in Sect. 4, let us first write the approximation in (6) as $\hat{x}_k = W W^T x_k + B\theta$. We know from (17) and (23) that $\theta = \mu$ and, hence,

$$\hat{x}_k = W W^T x_k + B\mu. \tag{29}$$

If $\widetilde{W}\widetilde{W}^T = B$, we have the approximation according to [1] in (10) of Sect. 4 equivalent to the approximation in (29).

While the drawbacks of (6) highlighted in Sect. 4 exist, let us outline the difference in these two approaches: we have demonstrated in the proposed framework that the new origin $\theta \in \mathbb{R}^p$ of the low dimensional coordinate system should be the mean $\mu \in \mathbb{R}^p$ so that the error of the approximation is reduced. But [1] necessitates an orthogonal projection of certain data-independent components $b \in \mathbb{R}^{p-q}$ to $\mu \in \mathbb{R}^p$ to achieve the same objective. The framework presented in this letter has shown that such a dimensionality reduction coupled with mean subtraction is unnecessary for deriving PCA.

8.4 Population PCA

For population PCA [10,14], where the samples that form the data are assumed to be realizations of a random variable, we have made it easy for the reader to follow our analysis by just replacing all occurrences of $\frac{1}{n}\sum_{k=1}^{n} \rightarrow \mathcal{E}$, the expectation operator; and bold faces for random variables as in $x_k \rightarrow \mathbf{x}$, $\hat{x}_k \rightarrow \hat{\mathbf{x}}$, and $\tilde{x}_k \rightarrow \tilde{\mathbf{x}}$.

9 Conclusion

Motivated by the need to justify the heuristics of pre-analysis mean centering in PCA and related questions, we have demonstrated through two distinct methods that the mean subtraction becomes part of the solution of the standard PCA problem in an approximation error minimization framework. We believe that the framework, in which we have compared a recent solution with ours, is more effective in justifying mean subtraction in PCA. Also, we have shown that the framework is comprehensive in the sense that the two outcomes of optimal dimensionality reduction, viz. approximation error minimization and retained variance maximization, are attained here simultaneously.

Acknowledgments This work was funded by the Project TANIA (WALEO II) of the Walloon Region, Belgium. The authors thank their colleague Olivier Caelen for his appreciable comments. Thanks are also due to Dr. P.P. Mohanlal of ISRO Inertial Systems Unit, India for his valuable insights. The authors are very grateful to the Editor-in-Chief and three anonymous reviewers for their excellent suggestions on an earlier version of this letter.

References

1. Bishop CM (2006) Pattern recognition and machine learning. Information science and statistics. Springer, New York
2. Diamantaras KI, Kung SY (1996) Principal component neural networks: theory and applications. John wiley, NewYork
3. Duda RO, Hart PE, Stork DG (2001) Pattern classification. 2nd edn. Wiley Interscience, New York
4. Fukunaga K (1990) Introduction to statistical pattern recognition. Computer science and scientific computing, 2nd edn. Academic Press, San Diego
5. Fukunaga K, Koontz WLG (1970) Application of the Karhunen–Loeve expansion to feature selection and ordering. IEEE Transac Comput C-19(4):311–318
6. Harsanyi JC, Chang C-I (1994) Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach. IEEE Transac Geosci Remote Sens 32(4):779–785
7. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:417–441
8. Huo X, Elad M, Flesia AG, Muise B, Stanfill R, Mahalanobis A et al (2003) Optimal reduced-rank quadratic classifiers using the Fukunaga–Koontz transform with applications to automated target recognition. Proc SPIE 5094:59–72
9. Hyvarinen A, Karhunen J, Oja E (2001) Independent component analysis, vol 27 of adaptive and learning systems for signal processing, communications and control. Wiley-Interscience, New York
10. Johnson RA, Wichern DW (1992) Applied multivariate statistical analysis, 3rd edn. Prentice-Hall, Inc., Upper Saddle River
11. Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York

12. Mahanalobis A, Muise RR, Stanfill SR, Van Nevel A (2004) Design and application of quadratic correlation filters for target detection. IEEE Transac Aerosp Electron Syst 40(3):837–850
13. Mann ME, Bradley RS, Hughes MK (1998) Global-scale temperature patterns and climate forcing over the past six centuries. Nature 392:779–788
14. Mardia K, Kent J, Bibby J (1979) Multivariate analysis. Academic Press, London
15. McIntyre S, McKitrick R (2005) Reply to comment by Huybers on "hockey sticks, principal components, and spurious significance". Geophys Res Lett 32:L20713
16. Miranda AA, Whelan PF (2005) Fukunaga–Koontz transform for small sample size problems. In: Proceedings of the IEE Irish signals and systems conference, pp 156–161, Dublin
17. Noy-Meir I (1973) Data transformations in ecological ordination: I. some advantages of non-centering. J Ecol 61(2):329–341
18. Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philos Mag 2:559–572
19. Plett GL, Doi T, Torrieri D (1997) Mine detection using scattering parameters and an artificial neural network. IEEE Transac Neural Netw 8(6):1456–1467
20. Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge
21. Van Huffel S (ed) (1997) Recent advances in total least squares techniques and errors-in-variables modeling. SIAM, Philadelphia