Submitted to the Brazilian Journal of Probability and Statistics
arXiv: arXiv:1308.6626

Searching for the core variables in principal components analysis

Yanina Gimenez[a], **and Guido Giussani**[a]

[a] *Universidad de San Andrés and CONICET*

Abstract. In this article, we introduce a procedure for selecting variables in principal components analysis. It is developed to identify a small subset of the original variables that best explain the principal components through nonparametric relationships. There are usually some noisy uninformative variables in a dataset, and some variables that are strongly related to one another because of their general dependence. The procedure is designed to be used following the satisfactory initial principal components analysis with all variables, and its aim is to help to interpret the underlying structures. We analyze the asymptotic behavior of the method and provide some examples.

1 Introduction

Principal components analysis (PCA) is the best known dimensional reduction procedure for multivariate data. An important drawback of PCA is that it sometimes provides poor quality interpretation of data for practical problems, because the final output is a linear combination of the original variables. The aim of the present study is to identify a small subset of the original variables in a dataset, whilst retaining most of the information related to the first k principal components.

There is a large body of literature that focuses on trying to interpret principal components. Jolliffe (1995) has introduced rotation techniques. Vines (2000) proposed to restrict the value of the loadings for PCA to a small set of allowable integers such as {-1, 0, 1}. McCabe (1984) presented a different strategy that aims to select a subset of the original variables with a similar criterion to PCA.

A few years ago a whole literature of variable selection appeared, inspired by the LASSO (least absolute shrinkage and selection operator). This technique was introduced by Tibshirani (1996). The way the LASSO works was described thus: " It shrinks some coefficients and sets others to 0, and hence

MSC 2010 subject classifications: Primary 62H25; secondary 62G08
Keywords and phrases. Informative Variables, Multivariate Analysis, Principal Components, Selection of Variables

tries to retain the good features of both subset selection and ridge regression... The 'LASSO' minimizes the residual sum of squares subject to the sum of the absolute values of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models."

Jolliffe et al. (2003) proposed SCoTLASS, that imposes a bound on the sum of the absolute values of the loadings in the component, using a similar idea to the one that LASSO used in regression. Zou et al. (2006) presented SPCA (sparse PCA), that extends the "elastic net" of Zou & Hastie (2005) that was a generalization of the LASSO. As they mentioned "SPCA is built on the fact that PCA can be written as a regression-type optimization problem, with quadratic penalty; the LASSO penalty (via the elastic net) can then be directly integrated into the regression criterion, leading to a modified PCA with sparse loadings."

Luss & d'Aspremont (2010) studied the application of sparse PCA to clustering and problems of feature selection. Sparse PCA seeks sparse factors, or linear combinations of variables in the dataset, to explain as much variance in the data as possible while limiting the number of nonzero coefficients as far as possible. The authors applied their results to some classic biological clustering and feature selection problems.

Recently, Witten & Tibshirani (2008) introduced the notion of Lassoed principal components for identifying differentially-expressed genes, and considered the problem of testing the significance of features in high dimensional data.

Our approach is rather different and is designed to be used after satisfactory PCA has been achieved rather than, as in other methods, to produce principal components with particular characteristics (e.g., some coefficients that are zero) so that only interpretable principal components are produced. We first perform classical PCA and then look for a small subset of the original variables that contain almost all the relevant information to explain the principal components. However, our method can also be used after performing any sparse PCA method as those described previously.

To develop our method, we borrowed some ideas for selecting variables from Fraiman et al. (2008), who introduced two procedures for selecting variables in cluster analysis, and classification rules. Both of these procedures are based on the idea of "blinding" unnecessary variables. To cancel the effect of a variable, they substituted all its values with the marginal mean (in the first procedure) or with the conditional mean (in the second). The marginal mean approach was mainly intended to identify "noisy" uninformative variables, but the conditional mean approach could also deal with

dependence. We adapted the idea behind the second procedure to PCA.

In Section 2, we introduce the main definitions, a population version of our proposed method, and the empirical version; we also present our main results. In Section 3, a simulation study is conducted, where the results are compared with other well-known PCA variable selection procedures. Finally, in Section 4, we study a real data example. Proofs are given in the Appendix.

2 Our method: Definitions and properties

We begin by defining our notation and stating the problem in terms of the underlying distribution of the random vector $\mathbf{X}$. Then we give our estimates based on the sample data, via the empirical distribution.

2.1 Population version

We define $\mathbf{X} \in \mathbb{R}^p$ as a random vector with distribution $\mathbb{P}$. The coordinates of the vector $\mathbf{X}$ are defined as $X[i], \quad i = 1, \ldots, p$. The covariance matrix of $\mathbf{X}$ is denoted Σ.

For a given random vector $\tilde{\mathbf{X}}$, we say that it satisfy the **Assumption H1** if:

i) $\mathbb{E}(\|\tilde{\mathbf{X}}\|^2) < \infty$;
ii) The covariance matrix of $\tilde{\mathbf{X}}$ is positive definite.
iii) All the eigenvalues of the covariance matrix are different.

Throughout the manuscript, $\mathbf{X}$ fulfills **Assumption H1**.

As is well known, the first principal component associated with the vector $\mathbf{X}$ is defined as

$$\begin{aligned} \alpha^1(\mathbb{P}) := \alpha^1 &= \underset{\|\alpha\|=1}{arg\ max}\ Var\left(\alpha'\mathbf{X}\right) \\ &= \underset{\|\alpha\|=1}{arg\ max}\ \alpha'\Sigma\alpha, \end{aligned}$$

and the next principal components are defined as

$$\begin{aligned} \alpha^k(\mathbb{P}) := \alpha^k &= \underset{\|\alpha\|=1,\ \alpha\perp[\alpha^1,\ldots,\alpha^{k-1}]}{arg\ max}\ Var\left(\alpha'\mathbf{X}\right) \\ &= \underset{\|\alpha\|=1,\ \alpha\perp[\alpha^1,\ldots,\alpha^{k-1}]}{arg\ max}\ \alpha'\Sigma\alpha \quad \forall\ 2 \le k \le p, \end{aligned}$$

where $[\alpha^1, \ldots, \alpha^{k-1}]$ is the subspace generated by the vectors $\{\alpha^1, \ldots, \alpha^{k-1}\}$.

From the spectral theorem, it follows that, if $\lambda^1 > \lambda^2 > \ldots > \lambda^p$ are the Σ eigenvalues, the solutions to the PCA are the corresponding eigenvectors, $\alpha^k, \quad k = 1, \ldots, p$.

Given a subset of indices $I \subset \{1, \ldots, p\}$ with cardinality $d \leq p$, we define $\mathbf{X}[I]$ as the subset of random variables $\{X[i], i \in I\}$. With a slight abuse of notation, if $I = \{i_1 < \ldots < i_d\}$, we can also denote $\mathbf{X}[I]$ to the vector $(X[i_1], \ldots, X[i_d])$, and define the vector $\mathbf{Y}^I := (Y^I[1], \ldots, Y^I[p])$, where

$$Y^I[i] = \begin{cases} X[i] & \text{if } i \in I \\ \mathbb{E}(X[i]|\mathbf{X}[I]) & \text{if } i \notin I. \end{cases}$$

$\mathbf{Y}^I \in \mathbb{R}^p$ depends only on the $\{X[i], i \in I\}$ variables and then the principal components associated with the vector $\mathbf{Y}^I$ depend only on the variables in $\mathbf{X}[I]$. The distribution of $\mathbf{Y}^I$ is denoted $\mathbb{P}_{\mathbf{Y}^I}$, the covariance matrix is $\Sigma_{\mathbf{Y}^I}$ and $g^i(z) = \mathbb{E}(X[i]|\mathbf{X}[I] = z)$ for $i \notin I$ is the regression function.

In what follows we assume that $\mathbf{Y}^I$ fulfills the **Assumption H1**.

Remark: This assumptions will not hold if there is a variable $X[i]$ such that $g^i(z)$ is a linear combination of the variables in I. For instance, if $\mathbf{X}$ is Gaussian or $X[i]$ is independent of the variables in I. In practice we will see how this problem can be tackled in the example 2 of section 3.

We are looking for a subset I of small cardinality that minimizes the distance between the original principal components and the principal components that are function of the variables in the subset I. This can be done following two different approaches.

Local approach

We define the objective function $h^1(I)$ as

$$h^1(I, \mathbb{P}, \mathbb{P}_{\mathbf{Y}^I}) := h^1(I, \mathbb{P}) := h^1(I) = \|\alpha^1(\mathbb{P}) - \alpha^1(\mathbb{P}_{\mathbf{Y}^I})\|^2, \tag{2.1}$$

which measures the squared distance between the first original principal component and the first principal component that is a function of the variables in the subset I.

Given a fixed integer d, $1 \leq d << p$, $\mathcal{I}_d$ is the family of all subsets of $\{1, \ldots, p\}$ with cardinality d and $\mathcal{I}_{1,0} \subset \mathcal{I}_d$ is the family of subsets in which the minimum $h^1(I)$ is attained for $I \in \mathcal{I}_d$, i.e.,

$$\mathcal{I}_{1,0} = \{I_1 \in \mathcal{I}_d : I_1 = argmin\ h^1(I)\},$$

or, equivalently,

$$h^1(I_1) = min_{I \in \mathcal{I}_d} h^1(I) \text{ for all } I_1 \in \mathcal{I}_{1,0}. \tag{2.2}$$

Analogously, we define

$$h^k(I, \mathbb{P}, \mathbb{P}_{\mathbf{Y}^I}) := h^k(I, \mathbb{P}) := h^k(I) = \|\alpha^k(\mathbb{P}) - \alpha^k(\mathbb{P}_{\mathbf{Y}^I})\|^2, \tag{2.3}$$

and

$$\mathcal{I}_{k,0} = \{I_k \in \mathcal{I}_d : I_k = argmin\ h^k(I)\},$$

for $k = 2, \ldots, p$.

For $k = 1, \ldots, p$ (2.3) measures the squared distance between the k-th original principal component and the k-th principal component that is a function of the variables in the subset I.

Principal components may, in some cases, be rather difficult to interpret, which is an important issue in practice. If we can find a small cardinal subset I, for which the objective function (2.3) is small enough, then the k-th principal component will be well explained by a subset of the original variables.

Global approach

In this case, the objective function is

$$h(I) := \sum_{k=1}^{q} p_k h^k(I), \tag{2.4}$$

with $p_k \geq 0,\ \sum_{k=1}^{q} p_k = 1,\ 2 \leq q < p$,
and we define

$$\tilde{\mathcal{I}}_{q,0} = argmin_{I \in \mathcal{I}_d} h(I).$$

This time the objective function (2.4) deals with finding a unique subset I to explain the first q principal components at once. If we think that the q components are equally important then we choose $p_k = \frac{1}{q}$. Otherwise, if some components are more important than others then we can put different weights. As a default we suggest choosing weights proportional to the variance that each component is explaining, i.e. $p_k = \lambda^k / \sum_{k=1}^{q} \lambda^k$.

On the local approach, we consider a different subset for each principal component, using the objective function (2.3) for each k. On the global approach, we seek a unique subset for all the first q principal components using (2.4). In practice, we can choose q, so that the first q principal components explain a high percent of the total variance and consider a subset for this principal components.

These subsets tell us which are the original variables that "best explain" the first q principal components. In what follows we refer to this method as the *Blinding Procedure* **(BP)**.

An important issue is how to choose d, the cardinality of the set I. On the one hand we have that the objective function (2.3) decreases when d

increases for every k. On the other hand we look for a subset I with small cardinality, so we have a constant tug of war. Since the components have unitary norms there is a direct relationship between (2.3) and the angle between the two vectors. It is clear that the smaller the angle is the closer the components are.

Hence, for the k-th component we propose to fix an angle, γ, and choose d as the smallest value for which $\mathcal{I}_{k,0}(d)$ makes the angle between the blinding and the original component smaller than γ.

If we want to "explain" the first q principal components at once, we propose to fix an angle, γ, and choose d as the smallest value for which $\tilde{\mathcal{I}}_{q,0}(d)$ makes the largest of the q angles smaller than γ. In both cases, the angle γ must be chosen by the user. As default, we propose to fix an angle not larger than 25 degrees.

2.2 Empirical version. Consistent estimates of the optimal subset

We aimed to consistently estimate the sets $I_1, \ldots, I_q$ from a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of i.i.d. random vectors with distribution $\mathbb{P}$. Remember that the subset I_k is a small cardinal subset that minimize the objective function (2.3) to "explain" the k-th principal component with a subset of the original variables.

Given a subset $I \in \mathcal{I}_d$, the first step is to build a sample $\mathbf{Y}_1^I, \ldots, \mathbf{Y}_n^I$ of random vectors in $\mathbb{R}^p$, which only depends on $\mathbf{X}[I]$, using nonparametric estimates of the conditional expectation (the regression function). Below, we assume that

Assumption H2 For all $i \notin I$, $g_n^i(z)$ is a strongly consistent estimate of $g^i(z)$ uniformly in z. Conditions under which **H2** holds can be found in Hansen (2008).

First, we define the empirical version of the *blinded* observations. As an example, we consider the nearest neighbours estimator. We therefore set an integer value r (the number of nearest neighbours that we are going to use) with respect to an appropriate metric (typically Euclidean or Mahalanobis distance), only considering the coordinates on I.

More precisely, for each $j \in \{1, \ldots, n\}$, we find the set of indices $C_j =: C_j(X_j[I])$ of the r-nearest neighbours of $\mathbf{X}_j[I]$ among $\{\mathbf{X}_1[I], \ldots, \mathbf{X}_n[I]\}$, where $\mathbf{X}_j[I] = \{X_j[i], i \in I\}$.

Next, we define the random vectors $\mathbf{Y}_j^I, 1 \leq j \leq n$, verifying:

$$Y_j^I[i] = \begin{cases} X_j[i] & \text{if } i \in I \\ \frac{1}{r}\sum_{m \in C_j} X_m[i] & \text{otherwise,} \end{cases} \tag{2.5}$$

where $X_j[i]$ stands for the ith-coordinate of the vector $\mathbf{X}_j$.

This corresponds to use for $i \notin I$ the nonparametric estimate

$$g_n^i(z) = g_n^i(z[I]) = \frac{1}{r} \sum_{m \in C_j(z[I])} X_m[i],$$

$g_n^i(z)$ being a strongly consistent estimate of $g^i(z)$, where $g^i(z) = \mathbb{E}(X[i]|\mathbf{X}[I] = z)$. Moreover, observe that the notation $z[I]$ is to emphasize that the function depends on the coordinates indicated by the subset I.

$\mathbb{P}_n$ stands for the empirical distribution associated with $\{\mathbf{X}_j, 1 \leq j \leq n\}$ and $\mathbb{P}_{n,\mathbf{Y}^I}$ stands for the empirical distribution of $\{\mathbf{Y}_j^I, 1 \leq j \leq n\}$.

In our examples to consistently estimate the optimal number of nearest neighbours we use the generalized cross validation procedure proposed by Li & Gong (1987), i.e.

$$\widehat{r}_{opt}(i, I) = \arg\min_r \frac{\frac{1}{n}\sum_{j=1}^n (X_j[i] - Y_j[i])^2}{\left(1 - \frac{1}{r}\right)^2}, \quad i \notin I.$$

When d is large, nonparametric estimators perform poorly due to the curse of dimensionality. In this case, a semi-parametric approach should be used like that proposed by He & Shi (1996). Also a recent proposal by Biau et al. (2013) called COBRA can be used to avoid the curse of dimensionality. The consistency result given in the following theorem will still be valid as long as the semi-parametric estimates verify the consistency assumptions required for the purely nonparametric estimates.

We will be mainly interested in cases where d is small. In our experience a good idea is to start the search with some genetic algorithm which provides an initial solution with a few variables (d small) and then improve the result working with $card(I) \leq d$. Also a forward-backward algorithm as the one proposed in Fraiman et al. (2008) can be used.

Finally, we define, for each $I \in \mathcal{I}_d$, the corresponding empirical versions

$$\hat{\alpha}_n^k(I) := \alpha^k(\mathbb{P}_{n,\mathbf{Y}^I}), \hat{h}_n^k(I) = h^k(I, \mathbb{P}_n, \mathbb{P}_{n,\mathbf{Y}^I}), \hat{I}_{kn} = argmin_{I \in \mathcal{I}_d} \hat{h}_n^k(I) \quad (2.6)$$

and

$$\hat{h}_n(I) = \sum_{k=1}^q p_k \hat{h}_n^k(I), \quad \hat{I}_n = argmin_{I \in \mathcal{I}_d} \hat{h}_n(I) \quad (2.7)$$

respectively.

Let us observe that $\hat{\alpha}_n^k(I)$ corresponds to the principal component associated with the sample $\{\mathbf{Y}_j^I, 1 \leq j \leq n\}$. $\hat{h}_n^k(I)$ corresponds to the objective function of the k-th principal component, it measures the square distance between the k-th original principal component and the k-th principal component that is a function of the variables in the subset I. The function $\hat{h}_n^k(I)$ determines which are the variables that retain most of the information related to the k-th principal component. The variables we are choosing are the ones that minimize the function $\hat{h}_n^k(I)$. The subset $\hat{I}_{kn}$ indicates which this variables are. In case we are looking for a unique subset I to explain the first q principal components, we consider the function $\hat{h}_n(I)$.

A robust version can be obtained using robust principal components (see for instance Maronna et al. (2006)) and replacing in (2.5) the local mean by local median.

Theorem 2.1. *Under assumptions H1 for* $\mathbf{X}$ *and* $\mathbf{Y}^I$ *and H2 we have that* $\hat{I}_{kn} \in \mathcal{I}_{k,0}$ *ultimately for* $k = 1, \ldots, q$, *i.e.,* $\hat{I}_{kn} = I_k$ *with* $I_k \in \mathcal{I}_{k,0}$ $\forall n > n_0(\omega)$, *with probability one. We also have that* $\hat{I}_n \in \tilde{\mathcal{I}}_{q,0}$ *ultimately.*

Proof. The proof is given in the Appendix. □

3 Some simulated examples

In this section we consider two simulated experiments to analyze the behavior of our method and compare it with other methods proposed in the literature.

3.1 Example 1

To better understand the heart of our procedure we start with a simple simulation example in dimension four.

There are two "hidden" factors:

$$V_1 \sim N(0, 1.25^2),$$
$$V_2 \sim N(0, 0.55^2),$$

where V_1, V_2 are independent.

We construct the 4 observable variables as follows:

$$\begin{aligned} X_1 &= V_1 + \epsilon_1, \\ X_2 &= |V_1| + \epsilon_2, \\ X_3 &= V_2 + \epsilon_3, \\ X_4 &= V_1 V_2 + \epsilon_4, \end{aligned} \tag{3.1}$$

where ε_j, $1 \leq j \leq 4$ are i.i.d. $N(0, \sigma^2)$, with $\sigma = 0.01$ (Case 1), $\sigma = 0.1$ (Case 2) and $\sigma = 0.25$ (Case 3).

When we perform the variable selection it is clear that two variables containing the information given by V_1 and V_2 should be kept. This means that $\{X_1, X_3\}$, $\{X_1, X_4\}$ or $\{X_3, X_4\}$ would be "good" choices while $\{X_1, X_2\}$ is clearly a "bad" choice since only retains information of V_1. Additionally, $\{X_2, X_3\}$ (respec. $\{X_2, X_4\}$) is not a "good" choice since can not recover all the information of V_1 (respec. V_1 and V_2).

In this example, in all the cases, we compare our procedure with other proposals: algorithms $B2$ and $B4$ (Jolliffe (2002)), a variable selection approach proposed by McCabe (1984) and sparce PCA introduced by Zou et al. (2006). To retain q variables, algorithm $B2$ associates one of the original variables to each of the $p-q$ last PCA vectors and deletes those variables, while $B4$ associates one of the original variables to each of the first q PCA vectors and retains those variables.

For each case, we perform 500 replicates and on each of them we generate samples of size 100 from model (3.1) and compute the covariance matrix.

- **Case 1:** The first two principal components explain between 70 and 85 percent of the total variance. For the **BP** 85% of the times the maximum angle conformed by the first two principal components is smaller than 25 degrees, hence we keep two variables.
  The results exhibitted in Table 1 on page 10 show that 78% of the times the **BP** makes a "good" choice of variables, $B2$ only does it 27% of the times, while the other methods fail in more than 75% of the replications.
- **Case 2:** The first two principal components explain between 69 and 85 percent of the total variance. For the **BP** more than 78% of the times the maximum angle conformed by the first two principal components is smaller than 25 degrees, hence we keep two variables.
  The results exhibitted in Table 2 on page 10 show that that 75% of the times the **BP** makes a "good" choice of variables, $B2$ only does it 29% of the times, $B4$ and SPCA 27%, while McCabe fails in more than 75% of the replications.

Table 1 *Proportion of times where each method selects each pair of variables.(Case 1)*

	BP	B2	B4	McCabe	SPCA
X_1, X_2	0.22	0.732	0.76	0.782	0.76
X_1, X_3	0.37	0.004	0.002	0.002	0.002
X_1, X_4	0.366	0.264	0.238	0.216	0.238
X_2, X_3	0	0	0	0	0
X_2, X_4	0	0	0	0	0
X_3, X_4	0.044	0	0	0	0

Table 2 *Proportion of times where each method selects each pair of variables.(Case 2)*

	BP	B2	B4	McCabe	SPCA
X_1, X_2	0.248	0.714	0.728	0.752	0.728
X_1, X_3	0.354	0.002	0.002	0.002	0.002
X_1, X_4	0.368	0.284	0.27	0.246	0.27
X_2, X_3	0	0	0	0	0
X_2, X_4	0	0	0	0	0
X_3, X_4	0.03	0	0	0	0

- **Case 3:** The first two principal components explain between 67 and 82 percent of the total variance. For the **BP**, 69% of the times the maximum angle conformed by the first two principal components is smaller than 25 degrees, hence we keep two variables.
  The results exhibitted in Table 3 on page 10 show that 74% of the times the **BP** makes a "good" choice of variables, $B2$ does it 51% of the times, $B4$ and SPCA 48%, while McCabe 46%.

Table 3 *Proportion of times where each method selects each pair of variables.(Case 3)*

	BP	B2	B4	McCabe	SPCA
X_1, X_2	0.264	0.488	0.516	0.536	0.516
X_1, X_3	0.192	0.002	0	0	0
X_1, X_4	0.54	0.51	0.484	0.464	0.484
X_2, X_3	0	0	0	0	0
X_2, X_4	0	0	0	0	0
X_3, X_4	0.004	0	0	0	0

From the definition of X_1 and X_2 we have that X_2 is a function of X_1 plus an error. We can see from Table 1 on page 10, Table 2 on page 10 and Table 3 on page 10 that our procedure selects X_1 and X_2 only between 0.22 and 0.264 proportion of the times, when the other procedures select it in around a 0.5 proportion of the times or more. In most of the cases our procedure detects that X_2 is a "function" of X_1. That is, if X_1 is selected to be one of the two explanatory variables, for the other one, the **BP**, select

between X_3 and X_4. This way, the **BP** gains information about V_2 instead of getting redundant information by choosing X_2. Note that $\{X_3, X_4\}$ is also a "good" choice.

To make the problem a bit more challenging we enlarge the dimension of the dataset repeating the same variables plus noisy noninformative errors and also add some i.i.d. noisy variables. More precisely we consider the following model:

$$X_j = \begin{cases} V_1 + \varepsilon_j, & \text{if } 1 \le j \le 5, \\ |V_1| + \varepsilon_j, & \text{if } 6 \le j \le 10, \\ V_2 + \varepsilon_j, & \text{if } 11 \le j \le 15, \\ V_1 V_2 + \varepsilon_j, & \text{if } 16 \le j \le 20, \\ \varepsilon_j, & \text{if } 21 \le j \le 23, \end{cases}$$

where ε_j, $1 \le j \le 23$ are i.i.d. $N(0, \sigma^2)$, with $\sigma = 0.01$ (Case 1), $\sigma = 0.1$ (Case 2) and $\sigma = 0.25$ (Case 3).

We keep the first and second principal component because on 98% of the 500 replicates they explain between 70% and 85% of the total variance in Case 1, between 70% and 83% in Case 2 and between 65% and 78% in Case 3. Again we require all the methods to select two variables. In the three cases, at least 93% of the times the largest angle is smaller than 25 degrees, hence we consider it is a good decision to look for two variables.

There are five groups of variables $A_1, \ldots, A_5$. Within each group all the variables only differ on a noisy noninformative error. More precisely, the first four groups are $A_j = \{X_k, k = 5j - i, i = 0, \ldots, 4\}$ $(j = 1, \ldots, 4)$ and the last one is $A_5 = \{X_k, k = 21, 22, 23\}$. Clearly $\{\{A_j, A_j\},\ j = 1, \ldots 5\}$, $\{A_1, A_2\}$ and $\{\{A_j, A_5\},\ j = 1, \ldots 4\}$ are "bad" choices and also $\{A_2, A_3\}$ and $\{A_2, A_4\}$ are not good choices either.

Table 4 on page 12, Table 5 on page 12, and Table 6 on page 13 exhibit the proportion of replicates where each method selects one variable per group of variables. In the three cases the **BP** achieves the desired results more than 90% of the times, while the other methods fail more than 73% of the times.

3.2 Example 2

Zou et al. (2006) introduced the following simulation example. They have two "hidden" factors:

$$V_1 \sim N(0, 290),$$
$$V_2 \sim N(0, 300),$$

Table 4 *Proportion of replicates where each method selects one variable per group of variables.(Case1)*

	BP	B2	B4	McCabe	SPCA
A_1, A_1	0.012	0.184	0	0	0
A_1, A_2	0.026	0.202	0.744	0.784	0.744
A_1, A_3	0.214	0.08	0.004	0	0.004
A_1, A_4	0.57	0.114	0.252	0.216	0.252
A_1, A_5	0.028	0.086	0	0	0
A_2, A_2	0	0.076	0	0	0
A_2, A_3	0	0.048	0	0	0
A_2, A_4	0	0.058	0	0	0
A_2, A_5	0	0.032	0	0	0
A_3, A_3	0	0.016	0	0	0
A_3, A_4	0.124	0.02	0	0	0
A_3, A_5	0	0.016	0	0	0
A_4, A_4	0	0.036	0	0	0
A_4, A_5	0	0.026	0	0	0
A_5, A_5	0.026	0.006	0	0	0

Table 5 *Proportion of replicates where each method selects one variable per group of variables.(Case2)*

	BP	B2	B4	McCabe	SPCA
A_1, A_1	0.006	0.188	0	0	0
A_1, A_2	0.024	0.218	0.744	0.79	0.74
A_1, A_3	0.254	0.07	0.006	0	0.006
A_1, A_4	0.598	0.114	0.25	0.21	0.25
A_1, A_5	0.026	0.076	0	0	0
A_2, A_2	0	0.074	0	0	0
A_2, A_3	0	0.044	0	0	0
A_2, A_4	0	0.072	0	0	0
A_2, A_5	0	0.03	0	0	0
A_3, A_3	0	0.014	0	0	0
A_3, A_4	0.092	0.026	0	0	0
A_3, A_5	0	0.012	0	0	0
A_4, A_4	0	0.034	0	0	0
A_4, A_5	0	0.022	0	0	0
A_5, A_5	0	0.006	0	0	0

and a linear combination of them,

$$V_3 = -0.3V_1 + 0.925V_2 + \varepsilon$$

where V_1, V_2 and ε are independent, $\varepsilon \sim N(0, 1)$.

Then 10 observable variables are constructed as follows:

Table 6 *Proportion of replicates where each method selects one variable per group of variables.(Case3)*

	BP	B2	B4	McCabe	SPCA
A_1, A_1	0.028	0.17	0	0	0
A_1, A_2	0.004	0.234	0.736	0.762	0.728
A_1, A_3	0.292	0.094	0.008	0.004	0.008
A_1, A_4	0.604	0.104	0.256	0.234	0.256
A_1, A_5	0.026	0.084	0	0	0
A_2, A_2	0	0.072	0	0	0
A_2, A_3	0	0.032	0	0	0
A_2, A_4	0	0.07	0	0	0
A_2, A_5	0	0.03	0	0	0
A_3, A_3	0	0.022	0	0	0
A_3, A_4	0.046	0.024	0	0	0
A_3, A_5	0	0.012	0	0	0
A_4, A_4	0	0.04	0	0	0
A_4, A_5	0	0.006	0	0	0
A_5, A_5	0	0.006	0	0	0

$$X_j = \begin{cases} V_1 + \varepsilon_j, & \text{if } 1 \leq j \leq 4, \\ V_2 + \varepsilon_j, & \text{if } 5 \leq j \leq 8, \\ V_3 + \varepsilon_j, & \text{if } j = 9, 10, \end{cases}$$

where ε_j, $1 \leq j \leq 10$ are i.i.d. $N(0,1)$.

Zou et al. (2006) used the true covariance matrix of $(X_1, \ldots, X_{10})$ to perform PCA, SPCA and Simple Thresholding.

There are three groups of variables that share the same information. The first group, which we denote by A_1 corresponds to the variables X_1 to X_4, the second one, A_2, to the variables X_5 to X_8, and the third group, A_3, to the variables X_9 and X_{10}. By definition V_3 is also a function of V_1 and V_2, up to a noisy noninformative error.

Zou et al. (2006) show that SPCA correctly identifies the sets of important variables, the first SPCA identifies the group of the variables A_2 and the second SPCA the group of the variables A_1. Simple Thresholding incorrectly mixes two variables of A_2 with two of A_3.

We first consider the population version, using the true covariance matrix, where the first two principal components explain 99.6% of the total variance.

This example is outside of the theoretical framework stated in Section 2 because, for example, if $I = \{1, 2\}$, $g^l(z) = 0$ for $l = 5, 6, 7, 8$ due to the independence between $\{X_1, X_2\}$ and $\{X_5, X_6, X_7, X_8\}$, therefore Y^I does not fulfill the **Assumption H1**.

But this will not be a problem for the implementation of our procedure.

Clearly it is not adequate to keep one variable because when the cardinality of I is 1, $\Sigma_{\mathbf{Y}^I}$ has only one positive eigenvalue and then we can not recover two principal components.

If we choose a subset I of cardinal 2, then $\Sigma_{\mathbf{Y}^I}$ has two positive eigenvalues of multiplicity one and then, we can recover two principal components. If we select one variable of each of the three groups will be a "good" solution. A "bad" solution should be to choose the 2 variables within the same group.

Table 7 *Theoretical objective's value $h(I)$ when two variables are chosen (one of each group).*

	$h(I)$
A_1, A_1	1.656
A_1, A_2	2.257×10^{-6}
A_1, A_3	2.536×10^{-5}
A_2, A_2	1.028
A_2, A_3	0.001
A_3, A_3	1

In Table 7 on page 14 we can see that $h(I)$ takes large values (larger than one) for the "bad" choices and small values (smaller than 0.001) for the "good" choices. Moreover, for the "good" choices, the largest angle between the original and the blinded components is less than two dergees in all cases. Then, the **BP** is going to select a "good" choice. So our method performs perfectly well when the cardinal of I is 2.

Next, we consider a more realistic situation and perform a small simulation, where we estimate the covariance matrix, generating a sample of size 50 which we iterate 500 times. We perform the analysis considering the two first principal components, which explain more than the 99% of the variance.

We apply our procedure to select two variables for the first two principal components and in only 4.8% of the cases the larger angle was above 15 degrees and in 2.2% of the cases larger than 20 degrees.

In Table 8 on page 15 we can see that 99.2% of the times the **BP** makes a "good" choice. In addition McCabe and SPCA always make a "good" choice, $B4$ on 99% of the times and $B2$ only makes a "good" choice in 76.6% of the times.

4 A real data example

We consider a dataset obtained from the University of California, Irvine repository (Frank & Asuncion (2010)) as an example. This dataset contains

Table 8 *Proportion of times where each method selects one variable per group of variables.*

	BP	B2	B4	McCabe	SPCA
A_1, A_1	0	0.134	0.01	0	0
A_1, A_2	0.426	0.57	0.562	1	0.736
A_1, A_3	0.222	0.108	0.428	0	0.264
A_2, A_2	0	0.096	0	0	0
A_2, A_3	0.344	0.088	0	0	0
A_3, A_3	0.008	0.004	0	0	0

the values of six bio-mechanical characteristic that were used to classify orthopaedic patients into three classes (normal, disc hernia and spondylolisthesis groups), and includes data for one hundred normal patients, sixty patients with a disc hernia, and one hundred and fifty patients with spondylolisthesis. Each patient is represented in the dataset in terms of the six bio-mechanical attributes associated with the shape and orientation of the pelvis and the lumbar spine, namely (a) pelvic incidence, (b) pelvic tilt, (c) lumbar lordosis angle, (d) sacral slope, (e) pelvic radius and (f) grade of spondylolisthesis.

We use our procedure to select one variable for the first two principal components, which explain 85% of the variance. We give the same weight to all components, that is $p_k = 1/2$ in equation (2.4), and the "grade of spondylolisthesis", variable (f) is selected. In this case the largest angle is 7.5 degrees, hence we decide to keep one variable. Cross-validation finds 55 nearest neighbours for variables (a) and (b), 70 for variable (c), 102 for variable (d) and 39 for variable (e), while the empirical objective function (2.7) attains the value 0.017. We use the Euclidean distance in our procedure, but with the Mahalanobis distance we get the same results.

The plot on the first two principal components for the example data looks very similar to the plot on the principal components calculated using our procedure (Figure 1). It also shows that the patients with spondylolisthesis are separated from the rest of the patients, but the normal patients and the patients with disc hernias are mixed up together.

Next, we calculate the two principal components using the traditional method and our procedure for the subset of normal and disc hernia patients, using now two variables. The first two principal components explain 75% of the variance then we decide to keep them. We discard to select one variable because in that case the largest angle is close to 78 degrees. For two variables the largest angle is close to 21 degrees and for three variables the angle does not decrease very much (19 degrees), so we decide to keep two variables.

We consider both, the Euclidean and the Mahalanobis distance, and in

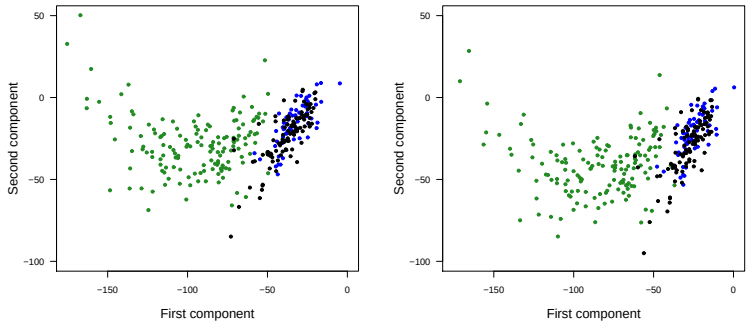

Figure 1 *Left: Map of the projection of the data on the original principal components. Right: Map of the projection of the data on the blinded based principal components. Blue: Disk Hernia patients. Black: Normal patients. Green: Spondylolisthesis patients.*

both cases the variables selected are "lumbar lordosis angle" and "pelvic radius". The value of the objective function (2.7) with the Euclidean distance is 0.125, while with the Mahalanobis distance is 0.141. Figure 2 shows that the plots for the principal components produced by these procedures look very similar.

5 Conclusions

In this paper we introduce a new variable selection procedure for PCA that aims to gain interpretation in principal components.

On the one hand we consider a local approach, where we deal with finding a subset of variables that best explains each principal component, and on the other hand we analyze a global approach, where the objective is to find a unique subset to explain the first q principal components at once.

We explain how we can choose the variables through the conditional expectation.

Under wide regular conditions in the conditional expectation estimates and in the covariance matrix, strong consistency results are studied and numerical aspects are analyzed.

The performance of the procedure is compared with several well-known variable selection techniques by using real and simulated data sets, showing the strengths of the method.

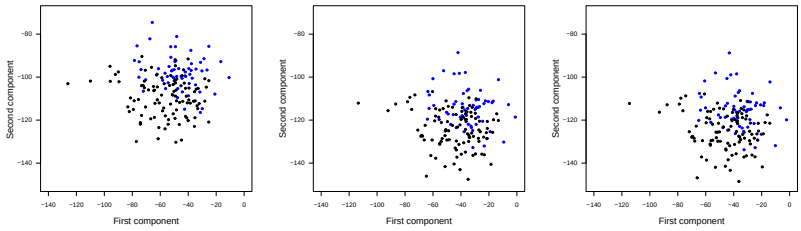

Figure 2 *Left: Map of the projection of the data on the original principal components. Middle: Map of the projection of the data on the blinded based principal components using Euclidean distance. Right: Map of the projection of the data on the blinded based principal components using Mahalanobis distance. Blue: Disk Hernia patients. Black: Normal patients.*

Appendix: Proof of Theorem 2.1.

First we are going to prove the following Proposition:

Proposition. *For $k \in \{1, ..., p\}$, if*

$$\hat{h}_n^k(I) \to h^k(I) \text{ a.s. for all } I \in \mathcal{I}_d, \tag{5.1}$$

then it converges uniformly, and

$$arg\min_{I \in \mathcal{I}_d} \hat{h}_n^k(I) \to arg\min_{I \in \mathcal{I}_d} h^k(I) \text{ a.s.}$$

Proof. Since $\mathcal{I}_d$ is a discrete and finite space , the convergence of (5.1) is uniform.

For the sake of simplicity, let $k = 1$ (the proof is analogue for $k \in \{2, ..., p\}$).

For all $\delta > 0$ there exists $n_0(\omega)$ such that for every $n \geq n_0(\omega)$, with probability one

$$\left|\hat{h}_n^1(I) - h^1(I)\right| < \delta \text{ for all } I \in \mathcal{I}_d,$$

then,

$$h^1(I) - \delta < \hat{h}_n^1(I) < h^1(I) + \delta \text{ for all } I \in \mathcal{I}_d \tag{5.2}$$

and

$$\hat{h}_n^1(I) - \delta < h^1(I) < \hat{h}_n^1(I) + \delta \text{ for all } I \in \mathcal{I}_d. \tag{5.3}$$

From (2.2), we know that exists $\delta_0 > 0$ such that

$$h^1(I_1) < h^1(I) - \delta_0 \text{ for all } I \notin \mathcal{I}_{1,0}, \text{ for all } I_1 \in \mathcal{I}_{1,0}.$$

By choosing in (5.2) $\delta = \frac{\delta_0}{2}$ we obtain

$$\hat{h}^1_n(I_1) < h^1(I_1) + \frac{\delta_0}{2} < h^1(I) - \delta_0 + \frac{\delta_0}{2} = h^1(I) - \frac{\delta_0}{2} < \hat{h}^1_n(I)$$

for all $I \notin \mathcal{I}_{1,0}$, for all $I_1 \in \mathcal{I}_{1,0}$, if $n \geq n_0(\omega)$ with probability one.

That means, that there exists $n_0(\omega)$ such that, if $n \geq n_0(\omega)$, with probability one, if we choose $I_1 \in \mathcal{I}_{1,0}$, then I_1 minimizes $\hat{h}^1_n(I)$.

Moreover, from (2.6), we have that for each fixed n, exists $\delta_1 > 0$, such that

$$\hat{h}^1_n(I^1) < \hat{h}^1_n(I) - \delta_1 \text{ for all } I \notin \hat{I}_{1n}, \text{ for all } I^1 \in \hat{I}_{1n}.$$

Let $\delta = \frac{\delta_1}{2}$ in (5.3) and $n \geq n_1(\omega)$,

$$h^1(I^1) < \hat{h}^1_n(I^1) + \frac{\delta_1}{2} < \hat{h}^1_n(I) - \delta_1 + \frac{\delta_1}{2} = \hat{h}^1_n(I) - \frac{\delta_1}{2} < h^1(I)$$

for all $I \notin \hat{I}_{1n}$, and for all $I^1 \in \hat{I}_{1n}$ with probability one.

That means, that exists $n_1(\omega)$ such that, if $n \geq n_1(\omega)$ for all $I^1 \in \hat{I}_{1n}$, I^1 minimizes $h^1(I)$ with probability one.

In conclusion, if $n > max\{n_0, n_1\}$, for all $I_1 \in \mathcal{I}_{1,0}$, I_1 minimizes $\hat{h}^1_n(I)$ and for all $I^1 \in \hat{I}_{1n}$, I^1 minimizes $h^1(I)$ with probability one.

As the proof is analogue for $k \in \{2, ..., p\}$,

$$\mathcal{I}_{k,0} = argmin_{I \in \mathcal{I}_d} h^k(I),$$

and

$$\hat{I}_{kn} = argmin_{I \in \mathcal{I}_d} \hat{h}^k_n(I)$$

this implies,

$$arg \min_{I \in \mathcal{I}_d} \hat{h}^k_n(I) \to arg \min_{I \in \mathcal{I}_d} h^k(I) \text{ a.s. for } k \in \{1, ..., p\}.$$

□

Proof. (of Theorem)

To prove the Theorem 2.1 it is enough to see that for each I the empirical objective function (2.6) converges a.s. to the theoretical objective function (2.3). To prove it, let us see that

$$\left\|\hat{\alpha}_n^k(\mathbb{P}_n) - \hat{\alpha}_n^k(\mathbb{P}_{n,\mathbf{Y}^I})\right\| \to \left\|\alpha^k(\mathbb{P}) - \alpha^k(\mathbb{P}_{\mathbf{Y}^I})\right\| \quad \text{a.s. for all } k.$$

Dauxois et al. (1982) proved that under **H1**, if

$$\left\|\hat{\Sigma}_n - \Sigma\right\|_\infty = \underset{\|u\|=1}{max} \left\|\left(\hat{\Sigma}_n - \Sigma\right)(u)\right\| \to 0 \ a.s.,$$

then

$$\left\|\hat{\alpha}_n^k(\mathbb{P}_n) - \alpha^k(\mathbb{P})\right\| \to 0 \quad \text{a.s. for all } 1 \leq k \leq p,$$

where $\hat{\alpha}_n^k(\mathbb{P}_n)$ (respec. $\alpha^k(\mathbb{P})$) are the eigenvectors of $\hat{\Sigma}_n$, that denotes the empirical covariance matrix associated with $\mathbb{P}_n$ (respec. Σ denotes the covariance matrix associated with $\mathbb{P}$).

So, if we prove that

$$\underset{\|u\|=1}{max} \left\|\left(\hat{\Sigma}_n(I) - \Sigma(I)\right)(u)\right\| \to 0 \ a.s.$$

then

$$\|\hat{\alpha}_n^k(\mathbb{P}_{n,\mathbf{Y}^I}) - \alpha^k(\mathbb{P}_{\mathbf{Y}^I})\| \to 0 \quad a.s. \quad \forall 1 \leq k \leq p,$$

where $\hat{\alpha}_n^k(\mathbb{P}_{n,\mathbf{Y}^I})$ (respec. $\alpha^k(\mathbb{P}_{\mathbf{Y}^I})$) are the eigenvectors of $\hat{\Sigma}_n(I)$, that denotes the empirical covariance matrix associated with $\mathbb{P}_{n,\mathbf{Y}^I}$ (respec. $\Sigma(I)$ denotes the covariance matrix associated with $\mathbb{P}_{\mathbf{Y}^I}$).

Note, that $I = \{1, \ldots, p\}$ is the classic case. We will show that this also holds for any $I \subseteq \{1, \ldots, p\}$, that is

$$\underset{\|u\|=1}{max} \left\|\left(\hat{\Sigma}_n(I) - \Sigma(I)\right)(u)\right\| \to 0 \ a.s.$$

To simplify the notation, we can assume (without losing generality) that $I = \{1, \ldots, d\}$, then

$$\mathbf{Y}_j = \left(X_j[1], \ldots, X_j[d], g_n^1(\mathbf{X}_j[I]), \ldots, g_n^{p-d}(\mathbf{X}_j[I])\right)^t,$$

where $g_n^i(z)$ is a uniformly consistent nonparametric estimate of $g^i(z)$. Specifically, g_n^i will fulfil the following assumption,

$$g_n^l(\mathbf{X}_j[I]) \to g^l(\mathbf{X}_j[I]) \ a.s. \text{ for any } l, \text{ for all } j, \text{ uniformly.}$$

We define a non observable auxiliary vector,

$$\mathbf{Z}_j = \Big(X_j[1], \ldots, X_j[d], g^1(\mathbf{X}_j[I]), \ldots, g^{p-d}(\mathbf{X}_j[I])\Big)^t,$$

where $g^i(z) = E(X[d+i]|\mathbf{X}[I] = z)$.

The proof will be complete if we show that

$$\underset{\|u\|=1}{max} \|(\Sigma_{\mathbf{Z}} - \Sigma(I))(u)\| \to 0 \ a.s., \tag{5.4}$$

and that

$$\underset{\|u\|=1}{max} \|(\Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{Z}})(u)\| \to 0 \ a.s., \tag{5.5}$$

where $\Sigma_{\mathbf{Y}} = \hat{\Sigma}_n(I)$.

Considering,

$$\overline{X[i]} = \frac{1}{n}\sum_{j=1}^{n} X_j[i], \quad \overline{g_n^l(\mathbf{X}[I])} = \frac{1}{n}\sum_{j=1}^{n} g_n^l(\mathbf{X}_j[I]), \quad \overline{g^l(\mathbf{X}[I])} = \frac{1}{n}\sum_{j=1}^{n} g^l(\mathbf{X}_j[I])$$

the three matrices are

$$(\Sigma(I))_{ii'} = \begin{cases} cov(X[i], X[i']) \text{ if } 1 \le i, i' \le d, \\ cov(X[i], \mathbb{E}(X[i']|\mathbf{X}[I])) \text{ if } 1 \le i \le d < i' \le p, \\ cov(\mathbb{E}(X[i]|\mathbf{X}[I]), X[i']) \text{ if } 1 \le i' \le d < i \le p, \\ cov(\mathbb{E}(X[i]|\mathbf{X}[I]), \mathbb{E}(X[i']|\mathbf{X}[I])) \text{ if } d+1 \le i, i' \le p, \end{cases}$$

$$(\Sigma_{\mathbf{Y}})_{ii'} = \begin{cases} \frac{1}{n}\sum_{j=1}^{n}(X_j[i] - \overline{X[i]})(X_j[i'] - \overline{X[i']}) \text{ if } 1 \le i, i' \le d, \\ \frac{1}{n}\sum_{j=1}^{n}(X_j[i] - \overline{X[i]})(g_n^{i'-d}(\mathbf{X}_j[I]) - \overline{g_n^{i'-d}(\mathbf{X}[I])}) \text{ if } 1 \le i \le d < i' \le p, \\ \frac{1}{n}\sum_{j=1}^{n}(g_n^{i-d}(\mathbf{X}_j[I]) - \overline{g_n^{i-d}(\mathbf{X}[I])})(X_j[i'] - \overline{X[i']}) \text{ if } 1 \le i' \le d < i \le p, \\ \frac{1}{n}\sum_{j=1}^{n}(g_n^{i-d}(\mathbf{X}_j[I]) - \overline{g_n^{i-d}(\mathbf{X}[I])})(g_n^{i'-d}(\mathbf{X}_j[I]) - \overline{g_n^{i'-d}(\mathbf{X}[I])}) \text{ if } d+1 \le i, i' \le p, \end{cases}$$

$$(\Sigma_{\mathbf{Z}})_{ii'} = \begin{cases} \frac{1}{n}\sum_{j=1}^{n}(X_j[i] - \overline{X[i]})(X_j[i'] - \overline{X[i']}) \text{ if } 1 \le i, i' \le d, \\ \frac{1}{n}\sum_{j=1}^{n}(X_j[i] - \overline{X[i]})(g^{i'-d}(\mathbf{X}_j[I]) - \overline{g^{i'-d}(\mathbf{X}[I])}) \text{ if } 1 \le i \le d < i' \le p, \\ \frac{1}{n}\sum_{j=1}^{n}(g^{i-d}(\mathbf{X}_j[I]) - \overline{g^{i-d}(\mathbf{X}[I])})(X_j[i'] - \overline{X[i']}) \text{ if } 1 \le i' \le d < i \le p, \\ \frac{1}{n}\sum_{j=1}^{n}(g^{i-d}(\mathbf{X}_j[I]) - \overline{g^{i-d}(\mathbf{X}[I])})(g^{i'-d}(\mathbf{X}_j[I]) - \overline{g^{i'-d}(\mathbf{X}[I])}) \text{ if } d+1 \le i, i' \le p. \end{cases}$$

We will now prove the convergence in (5.4) and (5.5).

First, we show that

$$\underset{\|u\|=1}{max} \|(\Sigma_{\mathbf{Z}} - \Sigma(I))(u)\| \to 0 \ a.s.$$

It is sufficient to prove that each of the coordinates of the matrix $\Sigma_{\mathbf{Z}}-\Sigma(I)$ converge to zero.

Set $1 \leq i, i' \leq p$, then $(\Sigma_{\mathbf{Z}})_{ii'} \to (\Sigma(I))_{ii'}$ $a.s.$ and then $(\Sigma_{\mathbf{Z}} - \Sigma(I))_{ii'} \to 0$ $a.s.$ $\forall i, i' = 1, ..., p$.

Because we are using a finite dimensional space, and each of the coordinates converge to zero, it holds that

$$\underset{\|u\|=1}{max} \|(\Sigma_{\mathbf{Z}} - \Sigma(I))(u)\| \to 0 \ a.s.$$

Now, we show that

$$\underset{\|u\|=1}{max} \|(\Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{Z}})(u)\| \to 0 \ a.s.$$

As before, we are going to prove that each of the coordinates of the matrix $\Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{Z}}$ converge to zero.

For better understanding, we define

$$\mathcal{G}(l, j, n) = g_n^l(\mathbf{X}_j[I]) - \overline{g_n^l(\mathbf{X}[I])} - g^l(\mathbf{X}_j[I]) + \overline{g^l(\mathbf{X}[I])}.$$

- If $1 \leq i, i' \leq d$, $(\Sigma_{\mathbf{Y}})_{ii'} = (\Sigma_{\mathbf{Z}})_{ii'}$ then $(\Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{Z}})_{ii'} = 0$.

- If $1 \leq i \leq d, d+1 \leq i' \leq p$,

$$\begin{aligned}(\Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{Z}})_{ii'} &= \frac{1}{n}\sum_{j=1}^{n}(X_j[i] - \overline{X[i]})(g_n^{i'-d}(\mathbf{X}_j[I]) - \overline{g_n^{i'-d}(\mathbf{X}[I])} - g^{i'-d}(\mathbf{X}_j[I]) + \overline{g^{i'-d}(\mathbf{X}[I])}) \\ &= \frac{1}{n}\sum_{j=1}^{n}(X_j[i] - \overline{X[i]})\mathcal{G}(i'-d, j, n).\end{aligned}$$

From the Cauchy-Schwarz inequality, we have that

$$\left|\frac{1}{n}\sum_{j=1}^{n}(X_j[i] - \overline{X[i]})\mathcal{G}(i'-d, j, n)\right| \leq \left(\frac{1}{n}\sum_{j=1}^{n}(X_j[i] - \overline{X[i]})^2\right)^{\frac{1}{2}} \left(\frac{1}{n}\sum_{j=1}^{n}(\mathcal{G}(i'-d, j, n))^2\right)^{\frac{1}{2}}$$

On the other hand,

$$\frac{1}{n}\sum_{j=1}^{n}(X_j[i] - \overline{X[i]})^2 \to Var(X[i]) \ a.s. \tag{5.6}$$

In addition, since $g_n^l(z)$ is a strongly consistent estimate of $g^l(z)$ uniformly in z for each fixed l, given $\epsilon > 0$ exist $N = N(\epsilon)$ such that for $n \geq N$

$$\left|g_n^l(z) - g^l(z)\right| < \epsilon \text{ for all } z.$$

Therefore

$$\left|\overline{g_n^l(\mathbf{X}[I])} - \overline{g^l(\mathbf{X}[I])}\right| = \left|\frac{1}{n}\sum_{j=1}^{n} g_n^l(\mathbf{X}_j[I]) - g^l(\mathbf{X}_j[I])\right| \leq \frac{1}{n}\sum_{j=1}^{n}\left|g_n^l(\mathbf{X}_j[I]) - g^l(\mathbf{X}_j[I])\right| < \epsilon.$$

Moreover, for each fixed l if $n \geq N$,

$$|\mathcal{G}(l,j,n)| \leq \left|g_n^l(\mathbf{X}_j[I]) - g^l(\mathbf{X}_j[I])\right| + \left|\overline{g_n^l(\mathbf{X}[I])} - \overline{g^l(\mathbf{X}[I])}\right| < 2\epsilon,$$

then,

$$\mathcal{G}(l,j,n) \to 0 \ a.s. \text{ for each fixed } l, \text{ uniformly.}$$

This implies that, for any l,

$$\left|\frac{1}{n}\sum_{j=1}^{n}(\mathcal{G}(l,j,n))^2\right| \leq \frac{1}{n}\sum_{j=1}^{n}\left|(\mathcal{G}(l,j,n))^2\right| < (2\epsilon)^2 = 4\epsilon^2$$

That is

$$\frac{1}{n}\sum_{j=1}^{n}(\mathcal{G}(l,j,n))^2 \to 0 \ a.s. \text{ for any } l. \tag{5.7}$$

Thus, by (5.6) and (5.7) we have

$$\left|\frac{1}{n}\sum_{j=1}^{n}(X_j[i] - \overline{X[i]})\mathcal{G}(i'-d,j,n)\right| \to 0 \ a.s. \tag{5.8}$$

which is what we required for our proof.

• If $d+1 \leq i \leq p, 1 \leq i' \leq d$, from the symmetry of the matrices and (5.8), we have that $(\Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{Z}})_{ii'} \to 0 \ a.s.$

• If $d+1 \leq i \leq p, d+1 \leq i' \leq p$,

$$(\Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{Z}})_{ii'} = \frac{1}{n}\sum_{j=1}^{n}\left[(g_n^{i-d}(\mathbf{X}_j[I]) - \overline{g_n^{i-d}(\mathbf{X}[I])})(g_n^{i'-d}(\mathbf{X}_j[I]) - \overline{g_n^{i'-d}(\mathbf{X}[I])})\right] - \left[(g^{i-d}(\mathbf{X}_j[I]) - \overline{g^{i-d}(\mathbf{X}[I])})(g^{i'-d}(\mathbf{X}_j[I]) - \overline{g^{i'-d}(\mathbf{X}[I])})\right].$$

If we add and subtract $(g^{i-d}(\mathbf{X}_j[I]) - \overline{g^{i-d}(\mathbf{X}[I])})(g_n^{i'-d}(\mathbf{X}_j[I]) - \overline{g_n^{i'-d}(\mathbf{X}[I])})$, and rearrange, we get that $(\Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{Z}})_{ii'}$ is the sum of

$$\frac{1}{n}\sum_{j=1}^{n}(g_n^{i'-d}(\mathbf{X}_j[I]) - \overline{g_n^{i'-d}(\mathbf{X}[I])})\mathcal{G}(i-d,j,n) \tag{5.9}$$

and

$$\frac{1}{n}\sum_{j=1}^{n}(g^{i-d}(\mathbf{X}_j[I]) - \overline{g^{i-d}(\mathbf{X}[I])})\mathcal{G}(i'-d,j,n)$$

where (5.9) can be rewritten as

$$\frac{1}{n}\sum_{j=1}^{n}[\mathcal{G}(i'-d,j,n) + (g^{i'-d}(\mathbf{X}_j[I]) - \overline{g^{i'-d}(\mathbf{X}[I])})]\mathcal{G}(i-d,j,n)$$

or

$$\frac{1}{n}\sum_{j=1}^{n}\mathcal{G}(i'-d,j,n)\mathcal{G}(i-d,j,n) + \frac{1}{n}\sum_{j=1}^{n}(g^{i'-d}(\mathbf{X}_j[I]) - \overline{g^{i'-d}(\mathbf{X}[I])})\mathcal{G}(i-d,j,n).$$

Then we have that

$$\begin{aligned}(\Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{Z}})_{ii'} &= \frac{1}{n}\sum_{j=1}^{n}\mathcal{G}(i'-d,j,n)\mathcal{G}(i-d,j,n)\\ &+ \frac{1}{n}\sum_{j=1}^{n}(g^{i'-d}(\mathbf{X}_j[I]) - \overline{g^{i'-d}(\mathbf{X}[I])})\mathcal{G}(i-d,j,n)\\ &+ \frac{1}{n}\sum_{j=1}^{n}(g^{i-d}(\mathbf{X}_j[I]) - \overline{g^{i-d}(\mathbf{X}[I])})\mathcal{G}(i'-d,j,n).\end{aligned}$$

From triangular and Cauchy-Schwarz inequality,

$$
\begin{aligned}
|(\Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{Z}})_{ii'}| \quad \leq \quad & \left(\frac{1}{n}\sum_{j=1}^{n}(\mathcal{G}(i'-d,j,n))^2\right)^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n}(\mathcal{G}(i-d,j,n))^2\right)^{1/2} \\
+ \quad & \left(\frac{1}{n}\sum_{j=1}^{n}(g^{i'-d}(\mathbf{X}_j[I]) - \overline{g^{i'-d}(\mathbf{X}[I])})^2\right)^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n}(\mathcal{G}(i-d,j,n))^2\right)^{1/2} \\
+ \quad & \left(\frac{1}{n}\sum_{j=1}^{n}(g^{i-d}(\mathbf{X}_j[I]) - \overline{g^{i-d}(\mathbf{X}[I])})^2\right)^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n}(\mathcal{G}(i'-d,j,n))^2\right)^{1/2}
\end{aligned}
$$

On the other hand, we have that

$$\frac{1}{n}\sum_{j=1}^{n}(g^l(\mathbf{X}_j[I]) - \overline{g^l(\mathbf{X}[I])})^2 \to Var(g^l(\mathbf{X}[I])) \ a.s. \text{ for any } l.$$

If $Var(g^l(\mathbf{X}[I])) < \infty$ for any l and I, (5.7) entails $|(\Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{Z}})_{ii'}| \to 0$.

As $Var(X[l]) < \infty$, then for any l and any I,

$$Var(X[l]) = E(Var(X[l]|\mathbf{X}[I])) + Var(E(X[l]|\mathbf{X}[I])) = E(Var(X[l]|\mathbf{X}[I])) + Var(g^l(\mathbf{X}[I]))$$

We now that $E(Var(X[l]|\mathbf{X}[I])) > 0$, so we conclude that $Var(g^l(\mathbf{X}[I])) < \infty$

We have proved that all the coordinates of the matrix $\Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{Z}}$ converge to zero, so

$$\underset{\|u\|=1}{sup} \ \|(\Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{Z}})(u)\| \to 0 \ a.s.$$

Now we are ready to complete the proof of Theorem 2.1. From (5.4) and (5.5) we are able to use the result from Dauxois et al. (1982) to derive that,

$$\hat{h}_n^k(I) = \|\hat{\alpha}_n^k(\mathbb{P}_n) - \hat{\alpha}_n^k(\mathbb{P}_{n,\mathbf{Y}^I})\|^2 \to \|\alpha^k(\mathbb{P}) - \alpha^k(\mathbb{P}_{\mathbf{Y}^I})\|^2 = h^k(I) \ \ a.s. \quad (5.10)$$

Indeed, we have that

$$\|\hat{\alpha}_n^k(\mathbb{P}_n) - \hat{\alpha}_n^k(\mathbb{P}_{n,\mathbf{Y}^I})\| - \|\alpha^k(\mathbb{P}) - \alpha^k(\mathbb{P}_{\mathbf{Y}^I})\| \leq$$
$$\|\hat{\alpha}_n^k(\mathbb{P}_n) - \alpha^k(\mathbb{P})\| + \|\alpha^k(\mathbb{P}) - \alpha^k(\mathbb{P}_{\mathbf{Y}^I})\| +$$
$$\|\alpha^k(\mathbb{P}_{\mathbf{Y}^I}) - \hat{\alpha}_n^k(\mathbb{P}_{n,\mathbf{Y}^I})\| - \|\alpha^k(\mathbb{P}) - \alpha^k(\mathbb{P}_{\mathbf{Y}^I})\| =$$
$$\|\hat{\alpha}_n^k(\mathbb{P}_n) - \alpha^k(\mathbb{P})\| + \|\alpha^k(\mathbb{P}_{\mathbf{Y}^I}) - \hat{\alpha}_n^k(\mathbb{P}_{n,\mathbf{Y}^I})\| \to 0 \ a.s.$$

and

$$\|\alpha^k(\mathbb{P}) - \alpha^k(\mathbb{P}_{\mathbf{Y}^I})\| - \|\hat{\alpha}_n^k(\mathbb{P}_n) - \hat{\alpha}_n^k(\mathbb{P}_{n,\mathbf{Y}^I})\| \leq$$
$$\|\alpha^k(\mathbb{P}) - \hat{\alpha}_n^k(\mathbb{P}_n)\| + \|\hat{\alpha}_n^k(\mathbb{P}_n) - \hat{\alpha}_n^k(\mathbb{P}_{n,\mathbf{Y}^I})\| +$$
$$\|\hat{\alpha}_n^k(\mathbb{P}_{n,\mathbf{Y}^I}) - \alpha^k(\mathbb{P}_{\mathbf{Y}^I})\| - \|\hat{\alpha}_n^k(\mathbb{P}_n) - \hat{\alpha}_n^k(\mathbb{P}_{n,\mathbf{Y}^I})\| =$$
$$\|\alpha^k(\mathbb{P}) - \hat{\alpha}_n^k(\mathbb{P}_n)\| + \|\hat{\alpha}_n^k(\mathbb{P}_{n,\mathbf{Y}^I}) - \alpha^k(\mathbb{P}_{\mathbf{Y}^I})\| \to 0 \ a.s.$$

which entails (5.10).

Finally, (5.10) and the Proposition implies

$$arg\min_{I\in\mathcal{I}_d} \hat{h}_n^k(I) \to arg\min_{I\in\mathcal{I}_d} h^k(I) \text{ a.s.}$$

which concludes the proof.

□

Acknowledgements

This work has been partially supported by Grant pict 2008-0921 from ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica), Argentina. We are very grateful to Dr. Ricardo Fraiman for his invaluable assistance and Dra. Marcela Svarc for her helpful suggestions. The authors would like to thank the referees for their constructive comments which improve significantly this work.

References

Biau, G., Fischer, A., Guedj, B. and Malley, J.D. (2013), "COBRA: A collective regression strategy", arxiv.org/pdf/1303.2236.

Dauxois, J., Pousse, A. and Romain, Y., (1982),"Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference", *Journal of Multivariate Analysis* **12,** 136–154.

Fraiman, R., Justel, A. and Svarc, M., (2008), "Selection of variables for cluster analysis and classification rules", *Journal of the American Statistical Association.* ***103(483),*** 1294–1303.

Frank, A. and Asuncion, A., (2010), UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science.

Hansen, B.E. (2008), "Uniform convergence rates for kernel estimation with dependent data", *Econometric Theory.* ***24,*** 726–748.

He, X. and Shi, P., (1996), "Bivariate Tensor-Product B-Splines in Partly Linear Models", *Journal of Multivariate Analysis.* ***58,*** 162–181.

Jolliffe, I., (1995), " Rotation of principal components: choise of normalization constraints", *Journal of Applied Statistics.****22,*** 29–35.

Jolliffe, I. (2002), "Principal Components Analysis. Second Edition", *Ed. Springer.*

Jolliffe, I., Trendafilov, N. and Uddin, M. (2003), " A modified principal component technique based on the LASSO", *Journal of Computational and Graphical Statistics.* ***12,*** 531–547.

Li, R. and Gong, G. (1987), "K-nn Nonparametric Estimation Of Regression Functions In the Presence of Irrelevant Variables", *Econometrics Journal,* ***00,*** 1-12.

Luss, R. and d'Aspremont, A., (2010), "Clustering and feature selection using sparse principal component analysis", *Optimization and Engineering.* ***11(1),*** 145–157.

Maronna, R. A., Martin, R. D. and Yohai, V. J., (2006), "Robust Statistics: Theory and Methods", Wiley, London.

McCabe, G. P. (1984), "Principal Variables", *Technometrics.* ***26,*** 137–144.

Tibshirani, R., (1996), "Regression shrinkage and selection via the LASSO", *Journal of the Royal statistical society. Series B* ***58(1),*** 267–288.

Vines,S., (2000), "Simple principal components", *Applied Statistics.* ***49,*** 441–451.

Witten, D.M. and Tibshirani, R., (2008), "Testing significance of features by lassoed principal components", *Annals of Applied Statistics.* ***2(3),*** 986–1012.

Zou, H. and Hastie, T., (2005), "'Regularizations and variable selection via the elastic net", *Journal of the Royal Statistical Society. Series B,* ***67,*** 301–320.

Zou, H., Hastie, T. and Tibshirani, R. (2006), "Sparse principal component analysis", *Journal of Computational and Graphical Statistics,* ***15,*** 265–286.

Universidad de San Andrés, Vito Dumas 284, 1644 Victoria, Buenos Aires, Argentina and CONICET
E-mail: yanugimenez@gmail.com; ggiussani@udesa.edu.ar