Selection of variables in principal components analysis

Ricardo Fraiman\textsuperscript{a,b}, Yanina Gimenez\textsuperscript{a,c} and Guido Giussani\textsuperscript{a}
\textsuperscript{a} Departamento de Matemática y Ciencia, Universidad de San Andrés, Vito Dumas 284, Victoria, Buenos Aires, Argentina
\textsuperscript{b} Departamento de Matemáticas, Facultad de Ciencias, Universidad de la República, Iguá 4225, Montevideo, Uruguay
\textsuperscript{c} CONICET

MSC: primary 62H25, 62G08

Keywords: Principal Components; Selection of variables, High dimensional data

Abstract

In this article, we introduce a procedure for selecting variables in principal components analysis. The procedure was 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 each other because of their general interdependence. The procedure is designed to be used following the satisfactory initial use of a principal components analysis with all variables, and its aim is to help to interpret underlying structures, particularly in high dimensional data. We analyse the asymptotic behaviour of the method and provide an example by applying the procedure to some real data.

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 was 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.

To develop our method, we borrowed some ideas for selecting variables from Fraiman et al. [2], 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.

There is a large body of literature on selecting variables, the most relevant publication being those related to the Lasso (least absolute shrinkage and selection operator), which was introduced by Tibshirani [3], and its application to different cases. The way the Lasso works was described thus: “It shrinks some coefficients and sets others to 0, and hence 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 value 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.”

Luss and d’Aspremont [4] 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 and Tibshirani [7] 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) such 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. This will become clear in the next section.

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 we apply the proposed method to a real dataset, as an 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 $X[i], \ i = 1, \ldots, p$.

We make the following assumptions throughout the manuscript,

**Assumption H1**

i) $\mathbb{E}(\|\mathbf{X}\|^2) < \infty$;

ii) The covariance matrix $\Sigma_{\mathbb{P}}$ is positive definite;

iii) The $\Sigma_{\mathbb{P}}$ eigenvalues are all different.

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

$$\alpha_1(\mathbb{P}) := \alpha^1 = \arg \max_{\|\alpha\|=1} \text{Var}_{\mathbb{P}}(\alpha' \mathbf{X}) = \arg \max_{\|\alpha\|=1} \alpha' \Sigma_{\mathbb{P}} \alpha,$$

and the next principal components are defined as

$$\alpha_k(\mathbb{P}) := \alpha^k = \arg \max_{\|\alpha\|=1, \ \alpha \perp [\alpha^1, \ldots, \alpha^{k-1}]} \text{Var}_{\mathbb{P}}(\alpha' \mathbf{X}) = \arg \max_{\|\alpha\|=1, \ \alpha \perp [\alpha^1, \ldots, \alpha^{k-1}]} \alpha' \Sigma_{\mathbb{P}} \alpha \ \forall \ 2 \leq k \leq p,$$

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 $\Sigma_{\mathbb{P}}$ eigenvalues, the solutions to the PCA are the corresponding eigenvectors, $\alpha^k, \ k = 1, \ldots, p$.

Given a subset of indices $I \subset \{1, \ldots, p\}$ with cardinal $d \leq p$, 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 \( X[I] \) to the vector \( (X[i_1], \ldots, X[i_d]) \), and define vector \( Y^I := Y = (Y[1], \ldots, Y[p]) \), where

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

\( Y^I \in \mathbb{R}^p \) depends only on the \( \{X[i], i \in I\} \) variables. The distribution of \( Y^I \) is denoted \( P_{Y^I} \). Finally, \( g^I(z) = E(X[i]|X[I] = z) \) for \( i \notin I \) is the regression function.

Next, we define the objective function \( h^1(I) \) as

\[
h^1(I, P, P_{Y^I}) := h^1(I, P) := h^1(I) = \| \alpha^1(P) - \alpha^1(P_{Y^I}) \|^2.
\]

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.,

\[
\mathcal{I}_{1,0} = \text{argmin}_{I \in \mathcal{I}_d} h^1(I),
\]

or, equivalently,

\[
h^1(I_1) = \text{min}_{I \in \mathcal{I}_d} h^1(I) \text{ for all } I_1 \in \mathcal{I}_{1,0}.
\]

Analogously, we define

\[
h^k(I, P, P_{Y^I}) := h^k(I, P) := h^k(I) = \| \alpha^k(P) - \alpha^k(P_{Y^I}) \|^2,
\]

and

\[
\mathcal{I}_{k,0} = \text{argmin}_{I \in \mathcal{I}_d} h^k(I),
\]

for \( k = 2, \ldots, q \), with \( q \leq p \).

We also define the subset \( \tilde{\mathcal{I}}_{q,0} \subset \mathcal{I}_d \), where

\[
\tilde{\mathcal{I}}_{q,0} = \text{argmin}_{I \in \mathcal{I}_d} h(I), \text{ for } h(I) := \sum_{k=1}^{q} p_k h^k(I), \quad (1)
\]

with \( p_k \geq 0, \sum_{k=1}^{q} p_k = 1 \).

Principal components may, in some cases be rather difficult to interpret. If we can find a small cardinal subset \( I \), for which the objective functions \( h^k(I) \) or \( h(I) \) are small enough, an important issue in practice is that the principal components will be well explained by a small subset of the original variables.
We can consider a different subset for each principal component, using objective functions $h^k$, $1 \leq k \leq q$, or a unique subset for all the first $q$ principal components using $h$.

These subsets tell us which are the original variables that “best explain” the first $q$ principal components.

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 $X_1, \ldots, X_n$ of iid random vectors with the distribution $\mathbb{P}$.

Given a subset $I \in \mathcal{I}_d$, the first step was to build a sample $Y_1^I, \ldots, Y_n^I$ of random vectors in $\mathbb{R}^p$, which only depends on $X(I)$, using nonparametric estimates of the conditional expectation (the regression function). Below, we assume that

**Assumption H2** $g_n^i(z)$ is a strongly consistent estimate of $g^i(z)$ for all $i \notin I$.

The simplest way to achieve this is to use nearest neighbour estimators. We therefore set an integer value $r$ (the number of nearest neighbours that we were going to use) and took the Euclidean distance, only considering the coordinates on $I$.

More precisely, for each $j \in \{1, \ldots, n\}$, we found the set of indices $C_j$ of the $r$-nearest neighbours of $X_j[I]$ among $\{X_1[I], \ldots, X_n[I]\}$, where $X_j[I] = \{X_j[i], i \in I\}$.

Next, we defined the random vectors $Y_j^I$, $1 \leq j \leq n$, verifying:

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

where $X_j[i]$ stands for the $i$th-coordinate of the vector $X_j$.

$\mathbb{P}_{n,Y_I}$ stands for the empirical distribution of $\{Y_j^I, 1 \leq j \leq n\}$.

Finally, we defined, for each $I \in \mathcal{I}_d$,

$$ \hat{\alpha}_n^k(I) := \alpha^k(\mathbb{P}_{n,Y_I}), \quad \hat{h}_n^k(I) = h^k(I, \mathbb{P}_n, \mathbb{P}_{n,Y_I}), \quad \hat{I}_kn = \arg \min_{I \in \mathcal{I}_d} \hat{h}_n^k(I) $$

respectively.
A robust version can be obtained using robust principal components (see for instance Maronna et al. [5]) and replacing in (2) the local mean by local medians.

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

The proof is given in the Appendix.

### 3 An example using real data

We used a data obtained from the University of California, Irvine repository (Frank and Asuncion (2010) [3]) as an example. This dataset contains 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 (1) pelvic incidence, (2) pelvic tilt, (3) lumbar lordosis angle, (4) sacral slope, (5) pelvic radius and (6) grade of spondylolisthesis.

We used our procedure to select one variable for the first two principal components, giving the same weight to all components, that is $p_k = 1/2$ in (1), and the “grade of spondylolisthesis” variable was selected. Cross-validation gave 23 nearest neighbours. We used the Euclidean distance in our procedure, but the Mahalanobis distance gave very similar results.

The plot of the first two principal components for the example data looks very similar to the plot of the principal components calculated using our procedure (Figure 1). The 2-D plots of the first two principal components show that the patients with spondylolisthesis were separated from the other, but that the normal patients and the patients with disc hernias had similar plotting positions.

Next, we calculated the two principal components using the traditional method, using our procedure for only the normal and disc hernia patients. Again, we gave the same weight to all of the components when applying our procedure, that is $p_k = 1/2$ in (1). From this, our procedure identified the fourth variable, “sacral slope” (using the four nearest neighbours, again obtained by cross-validation).
Figure 1: Left: The original first two principal components for the example dataset. Right: The principal components based only on the “grade of spondylolisthesis”.

Figure 2: Left: The original first two principal components for normal and disc hernia patients. Right: Principal components based only on “sacral slope”.

The two 2-D plots in Figure 2 again look very similar, but this time the components for the normal patients differ slightly from those of the disc hernia patients.

Finally, we used our procedure to select two variables for the first two principal components. We used both the Euclidean distance and the Mahalanobis distance, and the two selected variables were “pelvic incidence” and “grade of spondylolisthesis”. Figure 3 shows that the 2-D plots for the principal components produced by these procedures look very similar.
Figure 3: (a) The original first two principal components for the example data. (b) The principal components based only on the “pelvic incidence” and the “grade of spondylolisthesis” using the Euclidean distance. (c) The principal components based only on the “pelvic incidence” and the “grade of spondylolisthesis” using the Mahalanobis distance.

Taking only the normal patients and the patients with disc hernias (because these types of patient were mixed together in the plots), and applying our procedure to select two variables for the first two principal components, we found that the selected variables were “lumbar lordosis angle” and “pelvic radius”. We used both the Euclidean and the Mahalanobis distances. Figure 4 shows that the plots for the principal components produced by these procedures look very similar.

4 Appendix

Proof of Theorem II

We use $\hat{\Sigma}_n(I)$ to denote the empirical covariance matrix associated with $P_{n,Y^I}$ (for the sample $Y_1^I, \ldots, Y_n^I$). The spectral theorem implies that $\hat{\alpha}_k^I(I)$, $k = 1, \ldots, q$ are the eigenvectors of $\hat{\Sigma}_n(I)$ associated with the respective $k$ largest eigenvalues.

To achieve strong consistency between the eigenvalues and their associated eigenvectors, we assumed, as in Propositions 2 and 4 in the work published by Dauxois et al. [II], that it is sufficient to show the convergence of the covariance matrix in the operator space norm. More specifically, we used the following results,
Figure 4: (a) The original first two principal components for normal and disc hernia patients. (b) The principal components based only on the “pelvic incidence” and the “grade of spondylolisthesis” using the Euclidean distance. (c) The principal components based only on the “pelvic incidence” and the “grade of spondylolisthesis” using the Mahalanobis distance.

**Proposition 1.** If

\[
\sup_{\|u\|=1} \left\| \left( \Sigma_n - \Sigma \right) (u) \right\| \to 0 \text{ a.s.,}
\]

then

\[
\lambda_n^k \to \lambda^k \text{ a.s. } \forall 1 \leq k \leq p,
\]

and

\[
\alpha_n^k \to \alpha^k \text{ a.s. } \forall 1 \leq k \leq p,
\]

where \( \lambda^k_n \) are the eigenvalues and \( \alpha^k \) the eigenvectors for \( \Sigma_n \).

So, if

\[
\sup_{\|u\|=1} \left\| \left( \hat{\Sigma}_n(I) - \Sigma(I) \right) (u) \right\| \to 0 \text{ a.s.,}
\]

then

\[
\| \hat{\alpha}_n^k(\mathbb{P}_{n,Y}) - \alpha^k(\mathbb{P}_Y) \|^2 \to 0 \text{ a.s. } \forall 1 \leq k \leq p,
\]

when \( I = \{1, \ldots, p\} \) (because in this case: \( Y_j = X_j \) for all \( j = 1, \ldots, n \), \( \mathbb{P}_{n,Y} = \mathbb{P}_n \) and \( \mathbb{P}_Y = \mathbb{P} \), being the classic case.)

We will show that this also holds for any \( I \subseteq \{1, \ldots, p\} \), that is

\[
\sup_{\|u\|=1} \left\| \left( \hat{\Sigma}_n(I) - \Sigma(I) \right) (u) \right\| \to 0 \text{ a.s.}
\]
To simplify the notation, we can assume (without losing generality) that \( I = \{1, \ldots, d\} \), which is denoted in

\[
\mathbf{Y}_j = (X_j[1], \ldots, X_j[d], g_n^1(X_j[I]), \ldots, g_n^{d-1}(X_j[I]))^t,
\]
where \( g_n^i(z) = g_n(X_1[I], X_1[d+i], X_2[I], X_2[d+i], \ldots, X_n[I], X_n[d+i], z) \). \( g_n^i(z) \) is a consistent nonparametric estimate of \( g^i(z) \). Specifically, \( g_n \) will fulfil the following assumption,

\[
g_n^i(X_j[I]) \to g^i(X_j[I]) \text{ a.s. for any } l, \text{ for all } j.
\]

The non observable auxiliary vector is defined as

\[
\mathbf{Z}_j = (X_j[1], \ldots, X_j[d], g^1(X_j[I]), \ldots, g^{d-1}(X_j[I]))^t,
\]
where \( g^i(z) = E(X[d+i]|X[I] = z) \).

The proof will be complete if we show that

\[
\sup_{||u||=1} ||(\mathbf{Σ}_Z - \mathbf{Σ}(I)) (u)|| \to 0 \text{ a.s.,}
\]

and that

\[
\sup_{||u||=1} ||(\mathbf{Σ}_Y - \mathbf{Σ}_Z) (u)|| \to 0 \text{ a.s.,}
\]

where \( \mathbf{Σ}_Y = \hat{Σ}_n(I) \).

\[
(\mathbf{Σ}(I))_{ii'} = \begin{cases}
\text{cov}(X[i], X[i']) & \text{if } 1 \leq i, i' \leq d, \\
\text{cov}(X[i], E(X[i']|X[I])) & \text{if } 1 \leq i \leq d < i' \leq p, \\
\text{cov}(E(X[i]|X[I]), X[i']) & \text{if } 1 \leq i' \leq d < i \leq p, \\
\text{cov}(E(X[i]|X[I]), E(X[i']|X[I])) & \text{if } d + 1 \leq i, i' \leq p.
\end{cases}
\]

\[
(\mathbf{Σ}_Y)_{ii'} = \begin{cases}
\frac{1}{n} \sum_{j=1}^n (X_j[i] - \overline{X}[i])(X_j[i'] - \overline{X}[i']) & \text{if } 1 \leq i, i' \leq d, \\
\frac{1}{n} \sum_{j=1}^n (X_j[i] - \overline{X}[i])(g_n^{i-d}(X_j[I]) - g_n^{i-d}(\overline{X}[I])) & \text{if } 1 \leq i \leq d < i' \leq p, \\
\frac{1}{n} \sum_{j=1}^n (g_n^{i-d}(X_j[I]) - g_n^{i-d}(\overline{X}[I]))(X_j[i'] - \overline{X}[i']) & \text{if } 1 \leq i' \leq d < i \leq p, \\
\frac{1}{n} \sum_{j=1}^n (g_n^{i-d}(X_j[I]) - g_n^{i-d}(\overline{X}[I]))(g_n^{i-d}(X_j[I]) - g_n^{i-d}(\overline{X}[I])) & \text{if } d + 1 \leq i, i' \leq p.
\end{cases}
\]

\[
(\mathbf{Σ}_Z)_{ii'} = \begin{cases}
\frac{1}{n} \sum_{j=1}^n (X_j[i] - \overline{X}[i])(X_j[i'] - \overline{X}[i']) & \text{if } 1 \leq i, i' \leq d, \\
\frac{1}{n} \sum_{j=1}^n (X_j[i] - \overline{X}[i])(g^i(X_j[I]) - g^i(\overline{X}[I])) & \text{if } 1 \leq i \leq d < i' \leq p, \\
\frac{1}{n} \sum_{j=1}^n (g^i(X_j[I]) - g^i(\overline{X}[I]))(X_j[i'] - \overline{X}[i']) & \text{if } 1 \leq i' \leq d < i \leq p, \\
\frac{1}{n} \sum_{j=1}^n (g^i(X_j[I]) - g^i(\overline{X}[I]))(g^i(X_j[I]) - g^i(\overline{X}[I])) & \text{if } d + 1 \leq i, i' \leq p.
\end{cases}
\]
We will now prove the convergence in (3) and (4).

First, we show that
\[ \sup_{\|u\|=1} \|(\Sigma_Z - \Sigma(I)) \cdot u\| \to 0 \text{ a.s.} \]

It is sufficient to prove that each of the coordinates of the matrix \(\Sigma_Z - \Sigma(I)\) converges to zero. We considered four different cases for \((\Sigma_Z - \Sigma(I))_{ii'}:\)

- If \(1 \leq i, i' \leq d\)
  \[ \frac{1}{n} \sum_{j=1}^{n} (X_j[i] - \overline{X[i]})(X_j[i'] - \overline{X[i']}) - \text{cov}(X[i], X[i']) \to 0 \text{ a.s.} \]

- If \(1 \leq i \leq d, d + 1 \leq i' \leq p\)
  \[ \frac{1}{n} \sum_{j=1}^{n} (X_j[i] - \overline{X[i]})(g_i^{i'-d}(X_j[I]) - g_{i'}^{d}(X[I]) - \text{cov}(X[i], E(X[i']|X[I])) \to 0 \text{ a.s.} \]

- If \(d + 1 \leq i \leq p, 1 \leq i' \leq d\)
  \[ \frac{1}{n} \sum_{j=1}^{n} (g_i^{i-d}(X_j[I]) - g_{i'}^{d}(X[I]))(X_j[i'] - \overline{X[i']}) - \text{cov}(E(X[i]|X[I]), X[i']) \to 0 \text{ a.s.} \]

- If \(d + 1 \leq i \leq p, d + 1 \leq i' \leq p\)
  \[ \frac{1}{n} \sum_{j=1}^{n} (g_i^{i-d}(X_j[I]) - g_{i'}^{d}(X[I]))(g_i^{i'-d}(X_j[I]) - g_{i'}^{d}(X[I]) - \text{cov}(E(X[i]|X[I]), E(X[i']|X[I])) \to 0 \text{ a.s.} \]

Because we are using a finite dimensional space, and because each of the coordinates converges to zero, it holds that
\[ \sup_{\|u\|=1} \|(\Sigma_Z - \Sigma(I)) \cdot u\| \to 0 \text{ a.s.} \]

Now, we show that
\[ \sup_{\|u\|=1} \|(\Sigma_Y - \Sigma_Z) \cdot u\| \to 0 \text{ a.s.} \]

As before, we have proved that each of the coordinates of the matrix \(\Sigma_Y - \Sigma_Z\) converges to zero. We can write the terms \((\Sigma_Y - \Sigma_Z)_{ii'}\) as
• If $1 \leq i, i' \leq d$
\[
\frac{1}{n} \sum_{j=1}^{n} (X_j[i] - \bar{X}[i]) (X_j[i'] - \bar{X}[i']) - \frac{1}{n} \sum_{j=1}^{n} (X_j[i] - \bar{X}[i]) (X_j[i'] - \bar{X}[i']) = 0.
\]

• If $1 \leq i \leq d, d + 1 \leq i' \leq p$
\[
\frac{1}{n} \sum_{j=1}^{n} (X_j[i] - \bar{X}[i]) (g_n^{i'-d}(X_j[I]) - g_n^{i'-d}(\bar{X}[I])) - \\
\frac{1}{n} \sum_{j=1}^{n} (X_j[i] - \bar{X}[i]) (g_n^{i'-d}(X_j[I]) - g_n^{i'-d}(\bar{X}[I])) = \\
\frac{1}{n} \sum_{j=1}^{n} (X_j[i] - \bar{X}[i]) (g_n^{i'-d}(X_j[I]) - g_n^{i'-d}(X_j[I]) - g_n^{i'-d}(\bar{X}[I]) + g_n^{i'-d}(\bar{X}[I])).
\]

From the Cauchy-Schwarz inequality, we know that
\[
\left| \frac{1}{n} \sum_{j=1}^{n} (X_j[i] - \bar{X}[i]) (g_n^{i'-d}(X_j[I]) - g_n^{i'-d}(X_j[I]) - g_n^{i'-d}(\bar{X}[I]) + g_n^{i'-d}(\bar{X}[I])) \right| \leq \\
\left( \frac{1}{n} \sum_{j=1}^{n} (X_j[i] - \bar{X}[i])^2 \right)^{\frac{1}{2}} \times \\
\left( \frac{1}{n} \sum_{j=1}^{n} (g_n^{i'-d}(X_j[I]) - g_n^{i'-d}(X_j[I]) - g_n^{i'-d}(\bar{X}[I]) + g_n^{i'-d}(\bar{X}[I]))^2 \right)^{\frac{1}{2}}. \quad (5)
\]

We have
\[
\frac{1}{n} \sum_{j=1}^{n} (X_j[i] - \bar{X}[i])^2 \to Var(X[i]) \text{ a.s..} \quad (6)
\]

In addition,
\[
g_n^i(X_j(I)) \to g^i(X_j(I)) \text{ a.s. for any } l, \text{ for all } i,
\]
then
\[
\bar{g}_n^i(X(I)) \to \bar{g}^i(X(I)) \text{ a.s. for any } l,
\]
and therefore,
\[
(g_n^i(X_j(I)) - \bar{g}_n^i(X(I))) \to (g^i(X_j(I)) - \bar{g}^i(X(I))) \text{ a.s. for any } l, \text{ for all } j. \quad (7)
\]
This implies that
\[
\frac{1}{n} \sum_{j=1}^{n} (g_{n}^{i'-d}(X_j[I]) - g^{i'-d}(X_j[I])) - \frac{g_{n}^{i'-d}(X[I])}{n} + \frac{g^{i'-d}(X[I])}{n}^2 \to 0 \text{ a.s.. (8)}
\]

Thus, by (6), (8) and (5) we have
\[
\left| \frac{1}{n} \sum_{j=1}^{n} (X_j[i] - X[i])(g_{n}^{i'-d}(X_j[I]) - g^{i'-d}(X_j[I]) - \frac{g_{n}^{i'-d}(X[I])}{n} + \frac{g^{i'-d}(X[I])}{n}) \right| \to 0 \text{ a.s.,}
\]
which is what we required for our proof.

- If \( d + 1 \leq i \leq p, 1 \leq i' \leq d \)

\[
\frac{1}{n} \sum_{j=1}^{n} (g_{n}^{i'-d}(X_j[I]) - g^{i'-d}(X_j[I]))(X_j[i'] - X[i']) - \\
\frac{1}{n} \sum_{j=1}^{n} (g^{i-d}(X_j[I]) - g^{i-d}(X_j[I]))(X_j[i'] - X[i']) = \\
\frac{1}{n} \sum_{j=1}^{n} (g^{i-d}(X_j[I]) - g^{i-d}(X_j[I]) - \frac{g_{n}^{i-d}(X[I])}{n} + \frac{g^{i-d}(X[I])}{n})(X_j[i'] - X[i']) \to 0 \text{ a.s.}
\]

- If \( d + 1 \leq i \leq p, d + 1 \leq i' \leq p \)

\[
\frac{1}{n} \sum_{j=1}^{n} (g_{n}^{i'-d}(X_j[I]) - g^{i'-d}(X_j[I]))(g_{n}^{i'-d}(X_j[I]) - g^{i'-d}(X_j[I])) - \\
\frac{1}{n} \sum_{j=1}^{n} (g^{i-d}(X_j[I]) - g^{i-d}(X_j[I]))(g^{i'-d}(X_j[I]) - g^{i'-d}(X_j[I])) = \\
\frac{1}{n} \sum_{j=1}^{n} ((g_{n}^{i-d}(X_j[I]) - g^{i-d}(X_j[I]))(g_{n}^{i'-d}(X_j[I]) - g^{i'-d}(X_j[I])) - \\
(g^{i-d}(X_j[I]) - g^{i-d}(X_j[I]))(g^{i'-d}(X_j[I]) - g^{i'-d}(X_j[I])))
\]

By (7):
\[
(g_{n}^{l}(X_j[I]) - g^{l}(X_j[I])) \to (g^{l}(X_j[I]) - g^{l}(X_j[I])) \text{ a.s. for any l, for all i,}
\]
therefore
\[
\frac{1}{n} \sum_{j=1}^{n} ((g_{n}^{i-d}(X_j[I]) - g^{i-d}(X_j[I]))(g_{n}^{i'-d}(X_j[I]) - g^{i'-d}(X_j[I])) - \\
(g^{i-d}(X_j[I]) - g^{i-d}(X_j[I]))(g^{i'-d}(X_j[I]) - g^{i'-d}(X_j[I]))) \to 0 \text{ a.s.}
\]
We have proved that all the coordinates of the matrix $\Sigma_Y - \Sigma_Z$ converge to zero, so

$$\sup_{\|u\|=1} \|(\Sigma_Y - \Sigma_Z)(u)\| \to 0 \ a.s..$$

Now we can prove

$$\hat{h}_n^k(I) = \|\hat{\alpha}_n^k(P_n) - \hat{\alpha}_n^k(P_{n,Y}|I)|^2 \to \|\alpha^k(P) - \alpha^k(P_Y)|^2 = h^k(I) \ a.s..$$

To do this, we show that

$$\|\hat{\alpha}_n^k(P_n) - \hat{\alpha}_n^k(P_{n,Y}|I)| \to \|\alpha^k(P) - \alpha^k(P_{n,Y}|I)| \ a.s..$$

Indeed,

$$\|\hat{\alpha}_n^k(P_n) - \hat{\alpha}_n^k(P_{n,Y}|I)| \leq \|\alpha^k(P) - \hat{\alpha}_n^k(P_{n,Y}|I)| + \|\hat{\alpha}_n^k(P_{n,Y}|I) - \alpha^k(P_{n,Y}|I)| =$$

$$\|\hat{\alpha}_n^k(P_n) - \hat{\alpha}_n^k(P_{n,Y}|I)| \to 0 \ a.s..$$

On the other hand, we have

$$\|\alpha^k(P) - \alpha^k(P_{n,Y}|I)| \leq \|\alpha^k(P) - \hat{\alpha}_n^k(P)| + \|\hat{\alpha}_n^k(P) - \hat{\alpha}_n^k(P_{n,Y}|I)| + \|\hat{\alpha}_n^k(P_{n,Y}|I) - \alpha^k(P_{n,Y}|I)| =$$

$$\|\alpha^k(P) - \hat{\alpha}_n^k(P)| \to 0 \ a.s..$$

Finally, because we are using a discrete and finite space $\mathcal{I}_d$, if $\hat{h}_n^k(I) \to h^k(I) \ a.s.$ for all $I$ with cardinal $d$, it converges uniformly, so

$$\arg\min\hat{h}_n^k(I) \to \arg\min h^k(I) \ a.s..$$

which implies that

$$\hat{I}_{kn} = I_k \ \text{ultimately for} \ k = 1, \ldots, q.$$
5 Acknowledgements

This work has been partially supported by Grant pict 2008-0921 from ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica), Argentina.

6 References

References

[1] 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”, J. Multivariate Anal. 12, 136–154.

[2] 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.

[3] 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.

[4] Luss, R. and d’Aspremont, A., (2010), “Clustering and feature selection using sparse principal component analysis”, Optimization and Engineering. 11(1), 145–157.

[5] Maronna, R. A., Martin, R. D. and Yohai, V. J., (2006), “Robust Statistics: Theory and Methods”, Wiley, London.

[6] Tibshirani, R., (1996), “Regression shrinkage and selection via the LASSO”, Journal of the Royal statistical society. Series B 58(1), 267–288.

[7] Witten, D.M. and Tibshirani, R., (2008), “Testing significance of features by lassoed principal components”, Annals of Applied Statistics. 2(3), 986–1012.