ESAIM: PS 20 (2016) 463–479
DOI: 10.1051/ps/2016022

ESAIM: Probability and Statistics
www.esaim-ps.org

SEEKING RELEVANT INFORMATION FROM A STATISTICAL MODEL

Ricardo Fraiman[1], Yanina Gimenez[2] and Marcela Svarc[2]

Abstract. We herein introduce a general variable selection procedure, which can be applied to several parametric multivariate problems, including principal components and regression, among others. The aim is to allow the identification of a small subset of the original variables that can 'better explain' the model through nonparametric relationships. The method typically yields some noisy uninformative variables and some variables that are strongly related because of their general dependence and our aim is to help understand the underlying structures in a given data–set. The asymptotic behaviour of the proposed method is considered and some real and simulated data–sets are analysed as examples.

Mathematics Subject Classification. 62H30, 68T10, 62G20.

Received December 19, 2014. Revised June 13, 2016. Accepted August 10, 2016.

1. Introduction

The problem of variable selection in many statistical models has become an enormous challenge over the last decade, and is now a main–stream research area in statistics. The large volume of data from internet traffic and from computers in general, and the impressive ability to store these data, has led to the need to develop new statistical procedures to handle these challenges. Retrieving information from these data–sets, which usually have redundancies, ambiguities, and noise effects among their attributes, is a difficult problem. Several strategies have been used to overcome these difficulties.

Probably the best known and most studied methods are the procedures for variable selection for the ordinary linear model. All types of statistical software have built–in routines for selecting variables using classical variable selection procedures such as Bayesian Information Criteria and Akaike criteria, among others. Nonetheless, the study of this issue is not restricted to these models. In recent years, ad hoc criteria for feature extraction have been introduced for use in many multivariate problems, such as principal components analysis, regression and both supervised and unsupervised classification. The most widely used strategies are built either on Bayesian model averaging (BMA) or on 'least absolute shrinkage and selection operator' (LASSO) type penalties. The first approach was typified in the Bayesian proposals of Frayley and Raftery ([7, 8]), in which they analysed the unsupervised classification problem. Hoeting *et al.* [14], among others, studied the problems involved in simultaneously selecting variables and transformations for the linear model. An alternative perspective was developed by Tibshirani [21], who introduced the LASSO: his proposal is to minimize the residual sum of squares

Keywords and phrases. Variable selection, regression, principal components analysis.

[1] Centro de Matemática, Universidad de la República, Iguá 4225, Malvín Norte 11400, Montevideo, Uruguay.
rfraiman@cmat.edu.uy

[2] Universidad de San Andrés and Conicet, Vito Dumas 284, Victoria 1644,Buenos Aires, Argentina.
yanugimenez@gmail.com; msvarc@udesa.edu.ar

Article published by EDP Sciences

subject to the sum of the absolute values of the coefficients being less than a constant typically some coefficients are shrunk and others set to zero. This procedure is especially suitable for large data–sets, specifically when there are more variables than observations. Some important references for that direction were provided by Witten and Tibshirani [23], who introduced a procedure for clustering the observations using an adaptive, chosen set of features that they considered to be lasso–type penalty functions. The same authors (Witten and Tibshirani [22]) introduced the notion of using lassoed principal components to identify differentially–expressed genes, and they considered the problem of testing the significance of features in high–dimensional data. A detailed description of this method can be found in [12].

All these variable selection criteria jointly select the variables and adjust the statistical model. We tackle the variable selection problem from a different perspective. In first place, we perform a statistical analysis of the data (such as regression or principal components). Once this analysis has succeeded, we proceed to find the subset of variables that better explain the output. We do not propose to perform the analysis with fewer variables, in fact, it can even be applied after having performed a classical model selection procedure.

We extend some ideas introduced in Fraiman *et al.* [6] for classification methods to regression and principal components analysis. Therein, two procedures for variable selection in cluster analysis and classification rules have been introduced. Both of them are based on the idea of *blinding* unnecessary variables. To cancel the effect of one variable, they substitute all their values by the marginal mean in the first proposal and by the conditional mean in the second one. The marginal mean approach is mainly intended for identifying *noisy* uninformative variables, whereas the conditional mean approach can also deal with dependence. In this article, we extend the potential uses of the second of these procedures.

To illustrate our aims, consider the following simple linear model:

$$Y = \beta'\mathbf{X} + \varepsilon = 3X_1 - 3X_2 + 4X_3 + \varepsilon, \tag{1.1}$$

where $X_3 = \exp(X_1X_2)$ and X_1 and X_2 are normal independent random variables centred on the origin with unit variance; the errors, ε, are independent, normally distributed random variables with zero mean and variance 0.25. One hundred observations were generated using this model. It is clear that the regression function is a function of a subset of $\{X_1, X_2, X_3\}$ with cardinality 2, but there is no linear model for two of the three variables, X_1, X_2, and X_3, that fits the data well. But if we estimate X_3 by $E(X_3|X_1, X_2)$ then the model $Y = \beta'(X_1, X_2, E(X_3|X_1, X_2)) + \varepsilon$ will fit properly.

Figure 1a shows the surface adjusted to match the complete model and the surface adjusted to match $Y = \beta_1X_1 + \beta_2X_2 + \varepsilon$. It is clear that all three variables are needed to predict the underlying model correctly. Figure 1b shows a scatter plot of $\widehat{\beta}'\mathbf{X}$ *versus* $\widehat{\beta}'(X_1, X_2, \hat{E}(X_3|X_1, X_2))$, where $\hat{E}(X_3|X_1, X_2)$ is the $k-$nearest neighbors estimate with $k = 10$ and $\widehat{\beta}$ represents the OLS β estimate for Equation (1.1). The linear relationship that can be seen in this figure predicts X_3 extremely well, highlighting that almost all the information contained in this variable can be described by the other two variables when it is estimated (*via* conditional expectation) using the data–set. This is the central idea expressed in this paper, and it can be applied to a great variety of models.

The remainder of this paper is organized as follows. We introduce the main definitions and the population version of our proposal in Section 2, and provide the empirical counterparts of the procedures and present our main results in Section 3. We provide some practical advice for implementing our proposals in Section 4, and apply our methods to simulated and real data–sets in Section 5. Some final comments are given in Section 6, and all the proofs are given in the Appendix.

2. Our setup: Main definitions and notation

We begin by defining some notation used throughout the paper. Then we state the problem in terms of the underlying data distribution. We then express the population statements in terms of the sample data using the empirical distribution in a plug–in fashion.

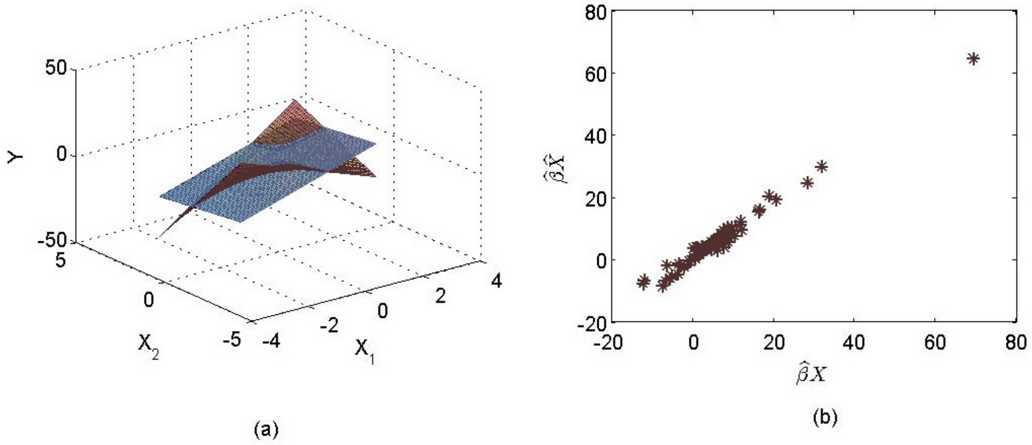

FIGURE 1. **(a)** The darker surface was adjusted using a linear model involving the three variables while the ligther surface was adjusted using a linear model involving only X_1 and X_2. **(b)** Scatter plot with the regression function for the complete data–set on the horizontal axis and, on the vertical axis, the regression function with X_3 predicted by X_1, X_2.

Even though we are presenting a variable selection procedure that can be applied to many parametric statistical problems, we are going to focus on regression models and principal components analysis. We first set the theoretical framework for the regression models, and then extend the ideas to principal components.

Let us consider the regression models,

$$Y = g(\mathbf{X}, \beta) + \varepsilon, \tag{2.1}$$

where, $Y, \varepsilon \in \mathbb{R}$ are random variables, $\mathbf{X} \in \mathbb{R}^p$ is a random vector, and $\beta \in \mathbb{R}^p$ is the vector of parameters to be estimated. The regressor variables, $\mathbf{X}$, have distribution $\mathcal{P}$, the errors, ε, have zero mean and are independent of $\mathbf{X}$. If $g(\mathbf{X}, \beta) = \mathbf{X}'\beta$ this is the classical linear model, otherwise it is a non-linear regression model.

The population target is given by β_0, fulfilling

$$\beta_0 = \text{argmin}_\beta E(\rho\,(Y - g(\mathbf{X}, \beta))). \tag{2.2}$$

If $\rho(\cdot) = \|\cdot\|^2$ we have the ordinary least squares estimates.

Suppose that we are satisfied with model (2.1). Our aim is to find a subset of variables from $\mathbf{X}$ that retains almost all the relevant information for the model, without redundancies or noise.

The coordinates of the vector $\mathbf{X}$ are denoted $X[i], \;\; i = 1, \ldots, p$. It is important to note that throughout this paper, p remains fixed.

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

$$Z(I)[i] = \begin{cases} X[i] & \text{if } \; i \in I \\ E(X[i]|\mathbf{X}(I)) & \text{if } \; i \notin I. \end{cases} \tag{2.3}$$

$\mathbf{Z}(I) \in \mathbb{R}^p$, but it depends only on $\{X[i], i \in I\}$ variables. The distribution of $\mathbf{Z}(I)$ is denoted $Q(I)$. Lastly, $\eta^i(z) = E(X[i]|\mathbf{X}(I) = z)$ for $i \notin I$, represents the regression function.

Typically we are interested in the case where $d \ll p$. Given a fixed integer d, $1 \leq d \ll p$, we let $\mathcal{I}_d$ be the family of all subsets of $\{1, \ldots, p\}$ with cardinality d. We seek a small subset, I, such that $g(\mathbf{X}, \beta_0)$ is as close as possible to $g(\mathbf{Z}(I), \beta_0)$.

More precisely, $\mathcal{I}_0 \subset \mathcal{I}_d$ is defined as the family of subsets in which the minimum $h(I)$ is attained for $I \in \mathcal{I}_d$, *i.e.*,

$$\mathcal{I}_0 = \text{argmin}_{I \in \mathcal{I}_d} h(I), \tag{2.4}$$

where

$$h(I) = E\left((g(\mathbf{X}, \beta_0) - g(\mathbf{Z}(I), \beta_0))^2 \right). \tag{2.5}$$

The objective function (2.5) measures the mean square distance between the regression function with the original variables and the regression function with the blinded variables.

We apply our procedure to the classical dimension reduction technique, principal components analysis (PCA). We seek a linear function of the components of $\mathbf{X}$ that accounts for most of the information contained in $\mathbf{X}$. More precisely, we seek the linear combination that has the largest variance among the values of $\mathbf{X}$. Our goal is to maximize the variance of the projection $\alpha'\mathbf{X}$, *i.e.*, to choose $\alpha \in \mathbb{R}^p$ according to,

$$\alpha_1 = \underset{\{\alpha:\|\alpha\|=1\}}{\text{argmax}}\, Var(\alpha'\mathbf{X}) = \underset{\{\alpha:\|\alpha\|=1\}}{\text{argmax}}\, \alpha'\Sigma\alpha,$$

where Σ is $\mathbf{X}$'s covariance matrix. In this way, we obtain the first principal component, which is the direction of the eigenvector corresponding to the largest eigenvalue. We assume that Σ is positive definite and that all the eigenvalues, $\lambda_1, \ldots, \lambda_p$ are different. The next principal component directions are defined by

$$\alpha_k = \underset{\{\|\alpha\|=1,\ \alpha\perp[\alpha_1,\ldots,\alpha_{k-1}]\}}{\text{argmax}} Var\left(\alpha'\mathbf{X}\right) = \underset{\{\|\alpha\|=1,\ \alpha\perp[\alpha_1,\ldots,\alpha_{k-1}]\}}{\text{argmax}} \alpha'\Sigma\alpha \quad \text{for} \quad k = 2, \ldots, p, \tag{2.6}$$

where $[\alpha_1, \ldots, \alpha_{k-1}]$ is the subspace spanned 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 eigenvalues of Σ, then the solutions to the principal components analysis are the corresponding eigenvectors $\alpha_k,\ \ k = 1, \ldots, p$. Therefore, $\mathbf{U}_k = \alpha_k'\mathbf{X}$ is the kth principal component. We assume that having performed PCA, the first $l \ll p$, principal components successfully represent the original data–set. Next, we define $I \in \mathcal{I}_d$ (typically $d \ll p$)

$$\mathbf{U}_k(I) = \alpha_k'\mathbf{Z}(I), \text{ for } k = 1, \ldots, l.$$

Lastly, we look for the subset $I \in \mathcal{I}_d$ that minimizes the objective function,

$$h(I) = \sum_{k=1}^{l} E\left(|\mathbf{U}_k - \mathbf{U}_k(I)|^2 \right). \tag{2.7}$$

This function measures the sum of the mean square distances between the projections in the direction of the l first principal components considering the original variables and the blinded ones.

Remark 2.1. The proposal can be extended to the case of eigenvalues with multiplicity greater than one. Let λ_j be an eigenvalue with multiplicity greater than one, denote by E_j the eigenmanifold of Σ corresponding to λ_j and P_j the orthogonal projection operator from $\mathbb{R}^p$ to E_j. Then write $U_j = P_jX$ and $U_j(I) = P_jZ(I)$, hence equation (2.7) remains well defined.

Remark 2.2. Extensions of these procedures to generalized linear models and canonical correlation analysis can be found in [9].

3. Empirical versions

In this section, we will define the empirical versions of the models described above and present the consistency results for each of the models studied. We require consistent estimates of the set I_0, $I_0 \subseteq \mathcal{I}_d$ based on a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of iid random p dimensional vectors, with a distribution $\mathcal{P}$.

Given a subset $I \in \mathcal{I}_d$, the first step is to obtain the blinded version of the sample of random vectors in $\mathbb{R}^p$, $\hat{\mathbf{X}}_1(I), \dots, \hat{\mathbf{X}}_n(I)$, that only depend on $\mathbf{X}(I)$, and estimate the conditional expectation (the regression function) non–parametrically. The following assumption is needed for all the models considered here.

H1. For all $i \notin I$, $\hat{\eta}^i(z)$ is a strongly consistent estimator of $\eta^i(z) = E(X[i]|\mathbf{X}(I) = z)$ for almost all z $(\mathcal{P})$ uniformly. Conditions under which **H1** holds can be found in [10].

First, we define the empirical version of the *blinded* observations. As an example, we consider the r nearest neighbour (r-NN) estimates. We fix an integer r (the number of nearest neighbours used) and calculate the Euclidean distance restricted to the coordinates I among the observations $\mathbf{X}_1(I), \dots, \mathbf{X}_n(I)$. For each $j \in \{1, \dots, n\}$, we find the set of indices C_j of the r nearest neighbours of $\mathbf{X}_j(I)$ among $\{\mathbf{X}_1(I), \dots, \mathbf{X}_n(I)\}$, where $\mathbf{X}_j(I) = \{\mathbf{X}_j[i], i \in I\}$.

Next we define the random vectors $\hat{\mathbf{X}}_j(I), 1 \leq j \leq n$ satisfying

$$\hat{X}_j(I)[i] = \begin{cases} X_j[i] & \text{if } i \in I \\ \frac{1}{r}\sum_{m \in C_j} X_m[i] & \text{otherwise,} \end{cases} \tag{3.1}$$

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

$Q_n(I)$ stands for the empirical distribution of $\{\hat{\mathbf{X}}_j(I), 1 \leq j \leq n\}$.

Given a subset of indices $I \in \mathcal{I}_d$, we define the empirical version of the objective function $h_n(I)$, and we seek the optimal subsets of variables $I_0 \subset I_d$, which are the family of subsets in which the minimum of $h_n(I)$ is reached, *i.e.*

$$\hat{\mathcal{I}}_0 = \text{argmin}_{I \in \mathcal{I}_d} h_n(I). \tag{3.2}$$

Thus, $h_n(I)$ and $h(I)$ must be explicitly determined for each statistical procedure.

For instance, in regression models, the empirical counterpart of the objective function (2.5) is given by

$$h_n(I) = \frac{1}{n}\sum_{j=1}^{n}\left(g\left(\mathbf{X}_j, \beta_n\right) - g\left(\hat{\mathbf{X}}_j(I), \beta_n\right)\right)^2, \tag{3.3}$$

where β_n is an almost surely consistent estimator of β_0. To obtain our main results, some additional assumptions may be needed for each model. For instance in a regression problem, we will require the following assumptions.

HR1. The estimate β_n converges a.s. to the true value β_0, and the regression function g is uniformly continuous.

Remark 3.1. In order to include the linear regression case, we can replace the uniform continuity of the regression function g by the following conditions:

$$|g^2(x,a) - g^2(x,b)| \leq C\|a-b\|^2\|x\|^2 \quad \text{and,} \quad |g^2(x,a) - g^2(y,a)| \leq C\|x-y\|^2\|a\|^2.$$

HR2. $E\left(g^2\left(\mathbf{X}, \beta_0\right)\right) < \infty$.

Theorem 3.2. *Let $\{(\mathbf{X}_j, Y_j), j \geq 1\}$ be iid $p{+}1$ dimensional random vectors satisfying (2.1). Given $d, 1 \leq d \leq p$, let I_d be the family of all subsets of $\{1, \dots, p\}$ with cardinality d and let $I_{d,0} \subset I_d$ be the family of subsets in which the minimum of the right–hand side of equation (2.5) is reached. Then, under* **H1**, **HR1** *and* **HR2**, *we have that $\hat{I}_n \in \mathcal{I}_0$ eventually almost surely, i.e., $\hat{I}_n = I_0$ with $I_0 \in \mathcal{I}_0$ $\forall n > n_0(\omega)$, with probability one.*

The proof can be found in the Appendix.

For the PCA, the empirical counterpart of equation (2.7) is given by

$$h_n(I) = \sum_{k=1}^{l}\frac{1}{n}\sum_{j=1}^{n}\left|\alpha_k'^n\mathbf{X}_j - \alpha_k'^n\hat{\mathbf{X}}_j(I)\right|^2. \tag{3.4}$$

where $\alpha_k'^n$ denotes an almost surely consistent estimator of α_k', for $k = 1, \dots, l$ and we require the following additional assumptions.

HP1. The covariance matrix Σ is positive definite and all of the eigenvalues are different.
HP2. $E\left((Z(I)[j]-X(I)[j])^2\right)<\infty$.

Theorem 3.3. *Let* $\{\mathbf{X}_j, j\geq 1\}$ *be iid* p *dimensional random vectors. Given* $d, 1\leq d\leq p$*, let* I_d *be the family of all the subsets of* $\{1,\ldots,p\}$ *with cardinality* d *and let* $I_{d,0}\subset I_d$ *be the family of subsets in which the minimum of the right–hand side of equation* (2.7) *is reached. Then, under* **H1**, **HP1** *and* **HP2** *we have that* $\hat{I}_n\in\mathcal{I}_0$ *eventually almost surely, i.e.* $\hat{I}_n=I_0$ *with* $I_0\in\mathcal{I}_0$ $\forall n>n_0(\omega)$*, with probability one.*

The proof can be found in the Appendix.

Remark 3.4. This result can be extended to the case where there are eigenvalues with multiplicity greater than one. Let $\lambda_{j,n}$ be a consistent estimator of λ_j an eigenvalue with multiplicity greater than one, and $P_{n,j}$ be the corresponding empirical orthogonal projection operator from $\mathbb{R}^p$ to E_n followed by estimating *via* plugging in U_j and $U_j(I)$. The proof of the consistency is similar to the proof of Theorem 3.3, the strong consistency of $P_{n,j}$ to P_j is stated in [2] Proposition 3.

Remark 3.5. If d is not sufficiently small the algorithm may become greedy. In that case, it may be convenient to use semiparametric models on the conditional mean or to use robust alternatives as more practical methods. The consistency results given in the previous theorem will be valid as long as the semiparametric estimates verify the consistency assumptions required for the purely nonparametric estimates. For instance, we can use the results presented by He and Shi (1996) to validate our consistency results.

4. Practical considerations

In this section, we present some practical considerations that will be useful for implementing the method, and will be helpful in the analysis of real data examples. The first point explains how the nonparametric regression estimation should be implemented, and the last point provides two heuristic rules for choosing the number of variables that should be retained.

Let $I\subseteq\{1,\ldots,p\}$, our aim is to perform a regression estimate of $X[i]$ based on $\mathbf{X}(I)$ for $i\in I^c$,

$$\widehat{X}[i]=E_{P_n}\left(X[i]|\mathbf{X}(I)\right).$$

For the sake of simplicity, as mentioned in Section 3, we choose an r-NN estimate. We suggest that the generalized cross validation procedure should be used, so that the optimal number of neighbours is estimated consistently, as described by Li (1987),

$$\widehat{X}_k[i]=\frac{1}{r}\sum_{j=1}^{n}X_j[i]I_{\{\|\mathbf{X}_j(I)-\mathbf{X}_k(I)\|\leq R_k(I)\}},$$

where $R_k(I)$ is the distance from $\mathbf{X}_k(I)$ to its r nearest neighbour.

The number of nearest neighbours $r=r(i,I)$ must be chosen for each subset I and coordinate $i\in I^c$. The optimal r will be chosen by generalized cross–validation:

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

Another important issue is how to select the number of variables that should be kept. In the procedures introduced in this paper the objective function is scale dependent, and so it is not straightforward to fix a threshold exogenously. We propose two heuristic procedures for choosing the number of variables that should be retained. In a *scree plot* approach, for each $d=\sharp I$, the subset $I_0(d)$ with the best predictive ability relative to the statistical procedure under consideration (*i.e.*, that minimizes the objective function) is kept, as well as the value that the objective function reaches, $h_n(I_0(d))$. It is clear that $h_n(I_0(d))$ decreases when the number of variables increases. Our proposal is to inspect the graph $(d, h_n(I_0(d)))$ and decide at which value of d the slope of the lines through the points are very steep to the left of d but not to the right. One can visualize an *elbow* at this value. We suggest that this number, d of variables should be used. Of course this procedure is greedy and computationally expensive if the number of variables in the original data–set and/or the number of retained variables is large. There are criteria that share the spirit of our proposal in problems concerning cluster and principal components analysis.

In a *penalized minimization* criterion, we propose to minimize h_n introducing a penalty for the cardinality of I, $\sharp I = d$. Specifically, our aim is to find a subset I_0 such that

$$I_0 = \underset{I \in \mathcal{I}}{\arg\min}\, h_n(I) + d^2 \sqrt{h_{n,0}}, \tag{4.1}$$

where $\mathcal{I}$ is the family of all the nonempty subsets of $\{1, \ldots, p\}$ and $h_{n,0} = \frac{1}{n}\sum_{j=1}^{n} g(\mathbf{X}_j, \beta_n)$ (respectively, $h_{n,0} = \sum_{k=1}^{l} \frac{1}{n}\sum_{j=1}^{n}(\alpha_k'^n \mathbf{X}_j))$ for the regression models (respectively, for PCA). The main advantage of this criterion is that it may be combined with a genetic algorithm to find a set of initial solutions, and then one may carry out an exhaustive search among the subsets of those variables. Another alternative is to implement a stepwise algorithm to find a solution. In either cases, there is no guarantee that an optimal solution will be attained, but these strategies allow avoiding the examination of all $2^p - 1$ possible subsets.

5. Some simulated and real data examples

In this section we conduct some simulation experiments to analyse the behaviour of our procedure and compare it with several well known variable selection procedures. We also challenge our proposal with a well known data–set.

5.1. Synthetic data–sets

5.1.1. Regression: The classical linear model

The data are generated fulfilling the following linear model:

$$Y_j = 2 + 3X_{1j} - 4X_{2j} + 2X_{3j} - \frac{3}{2}X_{4j} - X_{5j} + \frac{1}{50}X_{6j} + 3X_{7j} + 4X_{8j} - X_{9j} + e_j,$$

for $j = 1, \ldots, n$ where $X_{1j}, X_{2j} \sim N(0,4)$ are independent random variables, $X_{3j} = (X_{1j}X_{2j})^2$, $X_{4j} = (X_{1j}X_{2j})^3 Z_{1j}$, $X_{5j} = X_{1j}^4 Z_{2j}$, $X_{6j} = (X_{1j}X_{2j})^5 Z_{3j}$, $X_{7j} = \exp(X_{1j}X_{2j}), X_{8j} = \sqrt{|X_{1j}X_{2j}|}$ and $X_{9j} = \exp(X_{2j})Z_{4j}$, Z_{ij} for $i = 1, \ldots, 4$ are independent random variables normally distributed centred at the origin with variance 0.25, and the errors e_j are iid random variables with distribution $N(0,1)$. The model basically depends on two variables X_{1j} and X_{2j}, while variables $X_{3j}, \ldots, X_{8j}$ are nonlinear transformations of X_{1j} and X_{2j} or of the interaction between them.

Samples of size $n = 200$ and 300 were drawn repeatedly to attain 1000 replicates. For the sake of simplicity we did not find the optimal number of nearest neighbours for each variable and established 10 or 20 nearest neighbours for the nonparametric estimations. As suggested in Section 4, d, the number of variables retained, was selected by inspecting the graph $(d, h_n(I_0(d)))$ until the slope of the lines through the points is steepest at the left of d, blinded procedure scree plot (BPSP). We also implemented the blinded procedure penalized minimization criterion (BPPM) introduced in Section 4.

We challenged our method with several variable selection procedures. Four of them are classical variable selection criteria, namely: Akaike (AIC), Bayesian Information Criteria (BIC), Adjust R^2 (ADJR2) and the p-value for an F-test of the change in the sum of squared errors by adding or removing the term (SSE). A detailed reference for these criteria can be found in Seber and Lee [19] and almost every statistical software

Table 1. Summary statistics. *First line:* Mean number of variables selected. *Second line:* Median number of variables selected. *Third line:* % of hits = % of times that two variables have been chosen. *Fourth line:* $NOVS \leq 5$ = % of times that up to 5 variables have been chosen.

		Variable Selection Criteria										
Number of Observations		BPSP		BPPM		AIC	BIC	ADJR2	SSE	LARS	SCAD	MCP
		$r=10$	$r=20$	$r=10$	$r=20$							
$n=200$	Mean	2.05	2.05	2.22	2.22	8.39	8.28	8.19	8.24	4.64	3.42	4.31
	Median	2	2	2	2	9	9	9	9	4	3	4
	% of hits	94.50	94.70	79.40	79.40	0.40	1.00	2.10	1.90	14.2	19.7	13.5
	$NOVS \leq 5$	100	100	100	100	5.30	7.80	10.60	10.30	58.4	75.9	52.6
$n=300$	Mean	2.04	2.04	2.10	2.10	8.25	8.12	7.87	7.92	4.03	3.58	4.36
	Median	2	2	2	2	9	9	9	9	4	3	4
	% of hits	96.30	96.30	89.80	89.80	0.20	0.40	1.40	1.50	14	19	13
	$NOVS \leq 5$	100	100	100	100	5.10	8.60	14.10	13.70	68.5	70.5	49.8

package has built-in routines for them. We also ran several newer proposals, Least Angle Regression LARS (the regularization parameter was estimated by 10-fold cross–validation) [4], Smoothly Clipped Absolute deviation (SCAD) [5], and minimax concave penalty (MCP) [25]. The results are presented in Table 1.

We see from Table 1 that the number of variables selected by our criterion is on average very close to the number of variables that generated the model. It chooses two variables in more than 94% (respectively, 79%) of the replicates for the BPSP (respectively, BPPM), detecting the true underlying dependence structure. Moreover, whenever it does not choose two variables, it chooses at most 4. In contrast, the classical variable selection algorithms tend to keep almost all the variables, and if not, they rarely discard more than one variable. LARS, SCAD and MCP perform better: these methods keep fewer variables. LARS kept two variables in 14% of the replicates and the median number of variables is 4, however in 4.9% (respectively, 6.6%) of the replicates, only one variable has been selected for $n = 200$ (respectively, $n = 300$). MCP presents similar results: it kept two variables in 13% of the replicates, the median number of variables is again 4, and it chose one variable in 8.8% (respectively, 8.5%) of the replicates for $n = 200$ (respectively, $n = 300$). SCAD is the procedure that has the best performance among our competitors, achieving the correct number of variables in more than 19% of the replicates, however: it did not converge 10.4% (respectively, 16.5%) of the time, and in 30.7% (respectively, 28.7%) of the cases, it chose one or no variables (except for the intercept) for $n = 200$ (respectively, $n = 300$). In Figure 2, we exhibit the histograms for the different variable selection procedures for $n = 300$. It is clear that the blinding procedure is the most accurate one and that it is not sensitive to the number of neighbours chosen. SCAD, MCP and LARS keep a number of variables close to the optimal, while the classical procedures are not able to detect nonlinear dependence.

Finally, we measure the mean relative square error (MRSE)

$$MRSE = \frac{1}{n}\sum_{i=1}^{n}\frac{(\widehat{\mathbf{Y_i}} - \widehat{\mathbf{Y_i^*}})^2}{\widehat{\mathbf{Y_i}}^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{(\mathbf{X_i'}\widehat{\beta} - \widehat{\mathbf{X_i}}'\widehat{\beta})^2}{(\mathbf{X_i'}\widehat{\beta})^2},$$

for PBSP, where $\widehat{\mathbf{Y_i^*}} = \widehat{\mathbf{X_i}}'\widehat{\beta}$. Our aim is to measure the mean relative distance of the predicted observation considering the original and the blinded variables for the subset of variables selected. This quantity is between 0 and 1, and small values of it represents a good fit of the blinded process. The results are exhibit in Table 2. In every case the MRSE is of order less than or equal to 10^{-5}, showing a good fit of the variables selected.

5.1.2. Principal components

In this section we analyze the performance of the proposals for three different models.

TABLE 2. MRSE for the optimal subset of variables for BPSP, for $n = 200$ and 300 with $r = 10$ and 20.

Number of observations	k	
	10	20
200	4.71×10^{-5}	4.90×10^{-5}
300	7.54×10^{-6}	7.52×10^{-6}

Model 1. This example was first analysed by Zou *et al.* [26]. The model has 10 variables, that depend either on two random variables with an additive noise or a linear combination of them with an additive noise. The variables Y_{1j} are independent normally distributed with zero mean and variance 290, Y_{2j} are also iid normally distributed random variables with zero mean and variance 300 and $Y_{3j} = -0.3Y_{1j} + 0.925Y_{2j} + e_j$, where e_j are iid errors with normal distribution with zero mean and unit variance. The observations $(X_{1j}, \ldots, X_{10j})$ for $j = 1, \ldots, 100$ are defined as follows,

$$X_{ij} = \begin{cases} Y_{1j} + e_{ij} & \text{for} \quad i = 1, \ldots, 4, \\ Y_{2j} + e_{ij} & \text{for} \quad i = 5, \ldots, 8, \\ Y_{3j} + e_{ij} & \text{for} \quad i = 9, 10, \end{cases}$$

while the errors e_{ij} have been generated as $N(0, \sigma^2)$ random variables, with $\sigma^2 = 1$ or 100. The first two principal components account for more than 99% (respectively, more than 75%) of the total variance for $\sigma^2 = 1$ (respectively, $\sigma^2 = 100$).

Model 2. We consider $Y_{1,j}, Y_{2,j}$ and $Y_{3,j}$ for $j = 1, \ldots, 100$ as in *Model 1,* and also generate variables that either depend on $Y_{1,j}$ or on $Y_{2,j}$ or on the interaction between them, $Y_{1,j}Y_{2,j}$, defined as follows:

$$X_{ij} = \begin{cases} Y_{1j} + e_{ij} & \text{for} \quad i = 1, 2, \\ Y_{2j} + e_{ij} & \text{for} \quad i = 3, 4, \\ Y_{3j} + e_{ij} & \text{for} \quad i = 5, \\ \sqrt{|Y_{1j}Y_{2j}|} + e_{ij} & \text{for} \quad i = 6, \\ \log Y_{1j}^2 + e_{ij} & \text{for} \quad i = 7, \\ 12 \log Y_{2j}^2 + e_{ij} & \text{for} \quad i = 8, \end{cases} \tag{5.1}$$

where the errors e_{ij} follow the same distribution as in *Model 1.* The proportion of the total variance explained for $\sigma^2 = 1$ (respectively, $\sigma^2 = 100$) is almost always between 70% and 80% (respectively, 60% and 70%).

Model 3. We consider variables $X_{1,j}, \ldots, X_{8,j}$ as in *Model 2* and add 60 independent noise variables normally distributed with zero mean and variance $\sigma^2 = 1$ or 100.

We perform 1000 replications generating data from *Models 1* and *Models 2* and 200 replicates for *Model 3.* The nonparametric estimation of the conditional expectation is calculated with 5, 10 and 20 neighbours.

In every case, the data has been analysed by the blinded procedure scree plot scheme (BPSP) and also by the blinded procedure penalized minimization criterion (BPPM).

The output must be analysed in two stages. First, it is important to report the number of variables selected, and in the second stage, whenever two variables are selected, it is informative to determine whether those two variables are informative for retrieving almost all the information of the model.

We now analyse the results for *Model 1.* When the data is analysed by BPSP, two variables are always chosen. Almost always one variable has information only about Y_{1j} and the other one about Y_{2j}. In very few cases, it keeps one of them and selects a linear combination of them for the other variable. If the data is analysed by means of BPPM for $\sigma^2 = 1$, the results are the same as in the previous analysis, but for $\sigma^2 = 100$ it chooses two variables between 93.70% and 97.20% of the replicates, and if not three variables are chosen. Whenever

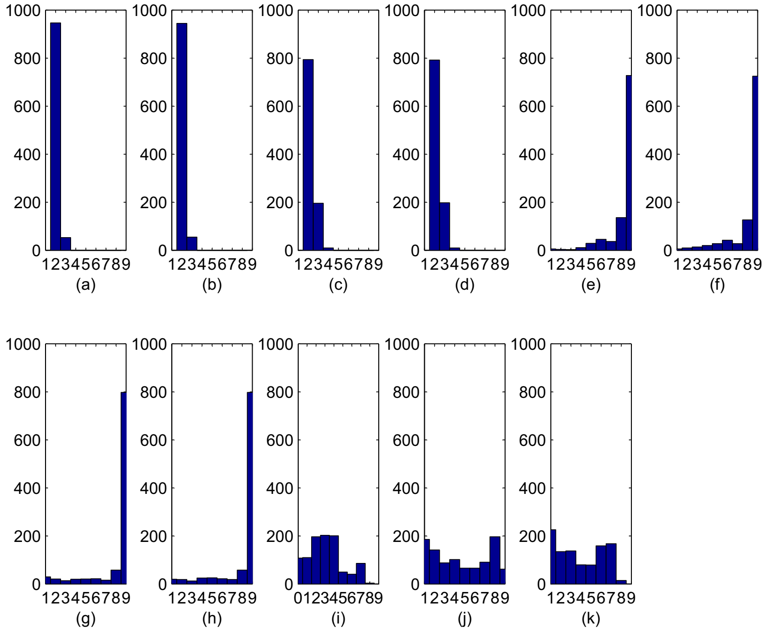

FIGURE 2. Number of times that each procedure selects d variables for Experiment 2 $n = 200$. (a) BPSP $r = 10$, (b) BPSP $r = 20$, (c) BPPM $r = 10$, (d) BPPM $r = 20$ (e) AIC, (f) BIC, (g) ADJR2, (h) SSE (i) LARS (j) SCAD (k) MCP.

two variables are chosen, the first has information about Y_1 and the second has information about Y_2 in almost every case, and if not, one of them is a linear combinations of Y_1 and Y_2.

It is important to note that *Model 1*, with $\sigma = 1$, has been previously analysed by Zou *et al.* [26]. They select the first 4 variables for the first principal component and $X_{5j}, \dots, X_{8j}$ for the second principal component. Is is clear that there is redundancy within each of those two groups.

For *Model 2*, BPSC keeps two variables in more than 90% of the replicates and if not, three variables are selected. The BPPM retains two variables in more than 80% (respectively, 73%) of the replicates for $\sigma = 1$ (respectively, $\sigma^2 = 100.$), and if not, it chooses either one or three variables. In Figure 3 we exhibit the optimal value for the objective function, for $1, 2, \dots, 5$ variables for each replication number of nearest neighbours, also the mean curve is plotted. It is clear that in almost every case the optimal number of variables is two. However it is important to retain the information provided by Y_1 and Y_2, hence keeping a set of variables that depends only either on Y_1 or on Y_2 it is not desirable. This happens, for the BPSP, only less than 6% (respectively, 31%) of the times for $\sigma^2 = 1$ (respectively, $\sigma^2 = 100$,) and for BPPM, only less than 9% (respectively, 34%) of the time for $\sigma^2 = 1$ (respectively, $\sigma^2 = 100$).

In addition, we compare our proposals with other variable selection procedures. There are methods where the number of variables must be given beforehand, while in other cases the number of variables is determined automatically. In the first category, we find several methods introduced by Jolliffe [15] such as $B2$ (to retain q variables, one associates one of the original variables to each of the $p - q$ last PCA vectors and deletes those variables) or $B4$ (which associates one of the original variables to each of the first q PCA vectors and retains those variables). There is also an algorithm introduced by McCabe [18]. From the second category, we consider the sparse PCA procedures, for instance BPSPA *via* penalized matrix decomposition (PMD) [24]. It is fairer to compare our results with those obtained applying BPSPA *via* PMD, because those methods do not need as an input the number of variables that should be retained. Hence we perform the same simulation (Model (5.1)) and find that even though they always keep all the variables, in more than 80% of the replicates, only 4 of them

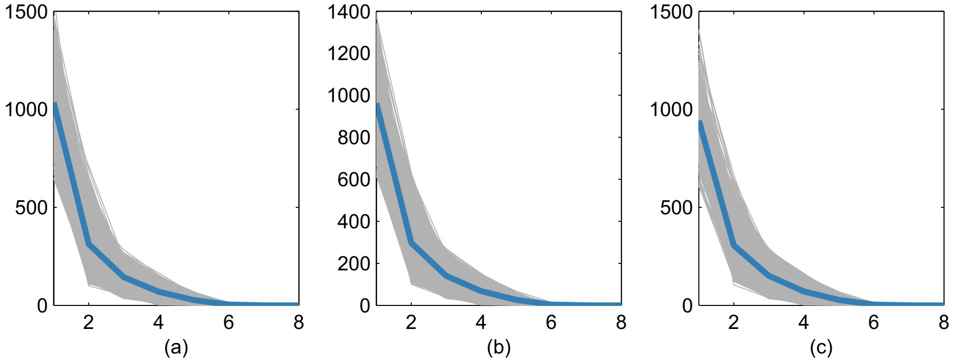

FIGURE 3. In each graph the optimal value for the objective function for each replicate is exhibited. The dark line is the mean curve. (a) 5-Nearest neighbours. (b) 10-Nearest neighbours. (c) 20-Nearest neighbours.

have a high loading on the first two PCAs, keeping redundant information. If we perform the same simulation study for $B2$, $B4$ and McCabe, always keeping two variables, the results of the different procedures are very similar, retaining a set of variables without information either about Y_1 or about Y_2 between 40% and 63% (respectively, 43% and 47%) of the time for $\sigma^2 = 1$ (respectively, $\sigma^2 = 100$). In every case our method clearly outperforms the results for the considered competitors.

Lastly, we analyse the results for *Model 3*. The BPSP chooses two variables and neither of them are noise, and only for 10% (respectively, 30%) of the replicates for $\sigma^2 = 1$ (respectively, $\sigma^2 = 100$) are the selected variables inadequate. For the BPPM, two variables are chosen in more than 77% of the replicates and in those cases neither of them were noise.

5.2. A real data example

5.2.1. Regression application: The diabetes data–set

Efron *et al.* [4] introduced the diabetes data–set for analysing least–angle regression (LARS) performance. LARS is a model selection technique for linear models that is less greedy than forward stepwise regression, and simple modifications of this method lead to considering all possible LASSO estimates or considering the "forward stagewise linear regression". The diabetes data–set consists of 442 observations of 10 predictors (age, gender, body mass, average blood pressure, and six blood serum measurements) and a response variable, glycaemia. Efron *et al.* [4] applied LARS to this data–set and identified only 4 important variables in the diabetes study: body mass, average blood pressure, and two blood serum measurements (variables 7 and 9). The authors suggested that the ordinary least squares model should be used rather than considering LARS, once the model has been chosen, both for the model selection procedure and for the parameter estimation, to make the output more familiar. Following their advice, our goal is to determine if the 4 variables are necessary for the complete interpretation of the model or if there is redundant information present. We conducted an exhaustive search among all the possible subsets of variables (by means of both criteria namely, scree plot and penalized minimization), because the number of variables is small. In Figure 4(a) we show the optimal value for the objective function, for different numbers of variables. It is clear that the biggest gain is obtained using two variables, where the graph exhibits an elbow. The optimal subset according to equation (3.3) is body mass and one of the blood serum measurements (variable 9). The number of nearest neighbours was cross–validated and found to be 86 for blood pressure and 107 for the other blood serum measurement (variable 7). In Figure 4(b), we present the values for the regression function with the original variables and with the blinded variables. The points should lie on the identity line if the estimation is errorless, and in our case it appears that the approximation is good because the points are in general not that far from the line.

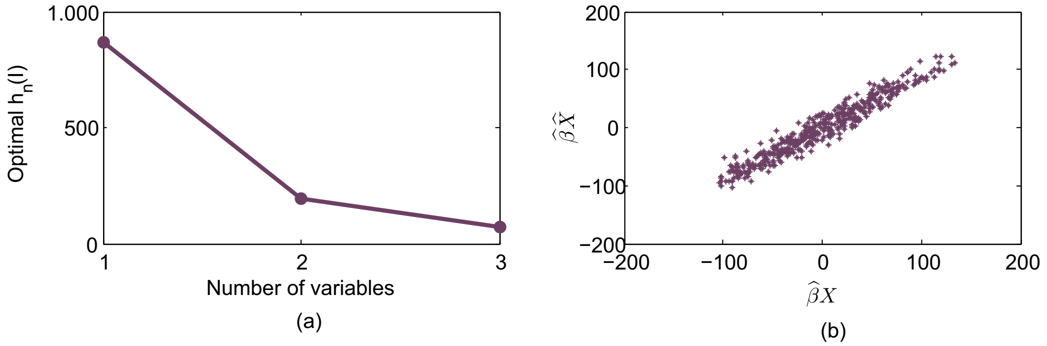

FIGURE 4. (a) Optimal values for the objective function for different numbers of variables. (b) Scatter plot with, on one axis the regression function for the complete data–set and on the other one the regression function when blood pressure and one of the blood serum measurements (variable 7) are predicted using body mass and the other blood serum measurement (variable 9).

6. Concluding remarks

We have provided a general procedure for finding the relevant variables for a family of different statistical multivariate models, including regression and principal components analysis. The same idea can be extended to other multivariate models, like GLIM, canonical correlation among others (see [9] for details).

The same idea can be extended to other multivariate models. The main idea is to *blind* some of the variables, replacing them by their predictions based on a small subset of the original variables.

From the computational point of view, it is well known that the variable selection problem is intrinsically NP–complete [3], since, as in every subset selection problem there is no way, without making further assumptions, of avoiding an exhaustive search: visiting all $2^p - 1$ possible subsets. The procedure is not therefore feasible if there is a large number of variables. This problem was previously described by Fraiman *et al.* [6], who proposed two ways of tackling it. One way is to use a fast optimization algorithm, genetic algorithm, for instance, retaining a small subset of variables, s, and search exhaustively among them. They also describe another method, a consistent forward-backward. It can be easily proved that the algorithm is consistent for all the cases that we studied. Another alternative for avoiding an exhaustive search is to implement a simulated annealing algorithm (Snell, 1995), but that approach goes beyond the scope of this paper.

Also, our proposal can be applied after having performed some classical selection of variables method, such as a LASSO based procedures. Hence our search begins on a lower dimensional space.

Appendix A.

Proof of Theorem 3.2. In order to prove our result it suffices to show that for each fixed subset I the empirical objective function (3.3) converges almost surely to the theoretical objective function (2.5), which will hold if we show that,

$$\frac{1}{n}\sum_{j=1}^{n} g^2\left(\mathbf{X}_j, \beta_n\right) \to E\left(g^2\left(\mathbf{X}, \beta_0\right)\right) \text{ a.s.,} \tag{A.1}$$

$$\frac{1}{n}\sum_{j=1}^{n} g^2\left(\hat{\mathbf{X}}_j(I), \beta_n\right) \to E\left(g^2\left(\mathbf{Z}(I), \beta_0\right)\right) \text{ a.s.} \tag{A.2}$$

and

$$\frac{1}{n}\sum_{j=1}^{n} g\left(\mathbf{X}_j, \beta_n\right) g\left(\hat{\mathbf{X}}_j(I), \beta_n\right) \to E\left(g\left(\mathbf{X}, \beta_0\right) g\left(\mathbf{Z}(I), \beta_0\right)\right) \text{ a.s} \tag{A.3}$$

In first place to prove (A.1). We observe that,

$$\frac{1}{n}\sum_{j=1}^{n} g^2\left(\mathbf{X}_j,\beta_n\right) = \frac{1}{n}\sum_{j=1}^{n}\left(g^2\left(\mathbf{X}_j,\beta_n\right) - g^2\left(\mathbf{X}_j,\beta_0\right)\right) \tag{A.4}$$

$$+\frac{1}{n}\sum_{j=1}^{n} g^2\left(\mathbf{X}_j,\beta_0\right) \tag{A.5}$$

From assumption **HR1** it follows that the right–hand side of (A.4) converges a.s. to 0. On the other hand, assumption **HR2**, $E\left(g^2\left(\mathbf{X},\beta_0\right)\right) < \infty$, implies that the almost surely convergence to $E\left(g^2\left(\mathbf{X},\beta_0\right)\right)$ of (A.5) follows from the Strong Law of Large Number (SLLN).

On a second stage, and in order to prove (A.2) we define the unobservable variables $\mathbf{Z}_1(I),\ldots,\mathbf{Z}_n(I)$, where

$$Z_j(I)[i] = \begin{cases} X_j[i] & \text{if } i \in I \\ E(X_j[i]|X_j(I)) & \text{otherwise,} \end{cases} \tag{A.6}$$

and denote by $Q_n^*(I)$ it's empirical distribution.

Hence, we have

$$\frac{1}{n}\sum_{j=1}^{n} g^2\left(\hat{\mathbf{X}}_j(I),\beta_n\right) = \frac{1}{n}\sum_{j=1}^{n}\left(g^2\left(\hat{\mathbf{X}}_j(I),\beta_n\right) - g^2\left(\hat{\mathbf{X}}_j(I),\beta_0\right)\right) \tag{A.7}$$

$$+\frac{1}{n}\sum_{j=1}^{n}\left(g^2\left(\hat{\mathbf{X}}_j(I),\beta_0\right) - g^2\left(\mathbf{Z}_j(I),\beta_0\right)\right) \tag{A.8}$$

$$+\frac{1}{n}\sum_{j=1}^{n} g^2\left(\mathbf{Z}_j(I),\beta_0\right). \tag{A.9}$$

The right–hand side of equation (A.7) converges a.s. to 0 since assumption **HR1** holds. The a.s. convergence to 0 of the term (A.8) follows from the uniform continuity of g stated in **HR1** and assumption **H1**. Moreover assumption **HR2**, $E\left(g^2\left(\mathbf{Z}(I),\beta_0\right)\right) < \infty$, implies that the almost sure convergence to $E\left(g^2\left(\mathbf{Z}(I),\beta_0\right)\right)$ of (A.9) follows from the SLLN.

On a third stage we are going to prove (A.3). We observe that,

$$\frac{1}{n}\sum_{j=1}^{n} g\left(\mathbf{X}_j,\beta_n\right) g\left(\hat{\mathbf{X}}_j(I),\beta_n\right) = \frac{1}{n}\sum_{j=1}^{n}\left(g\left(\mathbf{X}_j,\beta_n\right) g\left(\hat{\mathbf{X}}_j(I),\beta_n\right) - g\left(\mathbf{X}_j,\beta_0\right) g\left(\hat{\mathbf{X}}_j(I),\beta_0\right)\right)$$

$$+\frac{1}{n}\sum_{j=1}^{n}\left(g\left(\mathbf{X}_j,\beta_0\right) g\left(\hat{\mathbf{X}}_j(I),\beta_0\right) - g\left(\mathbf{X}_j,\beta_0\right) g\left(\mathbf{Z}_j(I),\beta_0\right)\right) \tag{A.10}$$

$$+\frac{1}{n}\sum_{j=1}^{n} g\left(\mathbf{X}_j,\beta_0\right) g\left(\mathbf{Z}_j(I),\beta_0\right). \tag{A.11}$$

From assumption **HR1** it follows that (A.10) converges a.s. to 0. The a.s. convergence to 0 of the term (A.10) follows from the uniform continuity of g stated in **HR1** and assumption **H1**. Moreover **HR2** implies that $E\left(\left|g\left(\mathbf{X},\beta_0\right) g\left(\mathbf{Z}(I),\beta_0\right)\right|\right) < \infty$, then from the SLLN we have that (A.11) converges to $E\left(\left|g\left(\mathbf{X},\beta_0\right) g\left(\mathbf{Z}(I),\beta_0\right)\right|\right)$. □

Proof of Theorem 3.3. In order to prove our statement it is enough to show that (3.4) converges to (2.7) a.s. To simplify notation, without loss of generality, we consider only one principal component (*i.e.*, $l = 1$), which will be denoted $\alpha = \alpha_1$. Then we have,

$$\begin{aligned} h_n(I) &= \frac{1}{n}\sum_{j=1}^{n}\left(\alpha'^n \mathbf{X}_j - \alpha'^n \hat{\mathbf{X}}_j(I)\right)^2 = \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i\notin I}\alpha^n[i](X_j[i]-\hat{X}_j(I)[i])\right)^2 \\ &= \sum_{i\notin I}(\alpha^n[i])^2\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-\hat{X}_j(I)[i]\right)^2 \\ &\quad +2\sum_{i,k\notin I, i<k}\alpha^n[i]\alpha^n[k]\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-\hat{X}_j(I)[i]\right)\left(X_j[k]-\hat{X}_j(I)[k]\right) \end{aligned}$$

and

$$\begin{aligned} h(I) &= E\left((\alpha'\mathbf{X}-\alpha'\mathbf{Z}(I))^2\right) = E\left(\left(\sum_{i\notin I}\alpha[i]\left(X[i]-Z(I)[i]\right)\right)^2\right) \\ &= \sum_{i\notin I}(\alpha[i])^2 E\left((X[i]-Z(I)[i])^2\right) \\ &\quad +2\sum_{i,k\notin I,i<k}\alpha[i]\alpha[k]E\left((X[i]-Z(I)[i])\left(X[k]-Z(I)[k]\right)\right). \end{aligned}$$

Dauxois *et al.* [2] show strong consistency of the eigenvalues and their associated eigenvectors, under mild regular conditions (see Prop. 2 and 4 therein). They establish that it is enough to show the convergence of the covariance matrix in the operator space norm. More specifically, they prove that if

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

then

$$\alpha_k^n \to \alpha_k \text{ a.s., for all } 1\le k\le p,$$

where α_k^n (respectively, α_k) are the eigenvectors of $\hat{\Sigma}_n$ (respectively, Σ), which is the empirical covariance matrix associated with P_n (respectively, $\mathcal{P}$).

The proof will be complete if we show that,

$$\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-\hat{X}_j(I)[i]\right)^2 \to E\left((X[i]-Z(I)[i])^2\right) \quad \text{a.s.,} \tag{A.12}$$

and

$$\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-\hat{X}_j(I)[i]\right)\left(X_j[k]-\hat{X}_j(I)[k]\right) \to E\left((X[i]-Z(I)[i])\left(X[k]-Z(I)[k]\right)\right) \quad \text{a.s.} \tag{A.13}$$

First, we prove (A.12). Recalling the definition given in (A.6) we have

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-\hat{X}_j(I)[i]\right)^2 &= \frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]+Z_j(I)[i]-\hat{X}_j(I)[i]\right)^2\\ &= \underbrace{\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]\right)^2}_{(a)}+\underbrace{\frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[i]-\hat{X}_j(I)[i]\right)^2}_{(b)}\\ &\quad \underbrace{+\frac{2}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]\right)\left(Z_j(I)[i]-\hat{X}_j(I)[i]\right)}_{(c)}.\end{aligned}$$

By the SLLN, since $\{X_j[i]-Z_j(I)[i],\ \text{for } j=1,\dots,n\}$ are iid random variables with finite second moment (assumption **HP2**), we have that,

$$\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]\right)^2 \to E\left(\left(X[i]-Z(I)[i]\right)^2\right) \quad \text{a.s.}$$

The a.s. convergence to 0 of (b) follows from assumption **H1**. Finally from Cauchy−Schwarz inequality, we have that,

$$\left|\frac{2}{n}\sum_{j=1}^{n}(X_j[i]-Z_j(I)[i])\left(Z_j(I)[i]-\hat{X}_j(I)[i]\right)\right| \le 2\left(\frac{1}{n}\sum_{j=1}^{n}(X_j[i]-Z_j(I)[i])^2\right)^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[i]-\hat{X}_j(I)[i]\right)^2\right)^{1/2}. \tag{A.14}$$

The first term on the right–hand side of (A.18) converges a.s. to $E((X[i]-Z[i])^2)$ by the SLLN, while the second one converges a.s. to 0 by assumption **H1**.

In second place, we are going to proof (A.13) following the same idea.

$$\begin{aligned}&\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-\hat{X}_j(I)[i]\right)\left(X_j[k]-\hat{X}_j(I)[k]\right)\\ &=\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]+Z_j(I)[i]-\hat{X}_j(I)[i]\right)\left(X_j[k]-Z_j(I)[k]+Z_j(I)[k]-\hat{X}_j(I)[k]\right)\\ &=\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]\right)\left(X_j[k]-Z_j(I)[k]\right)\\ &\quad+\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]\right)\left(Z_j(I)[k]-\hat{X}_j(I)[k]\right) &\text{(A.15)}\\ &\quad+\frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[i]-\hat{X}_j(I)[i]\right)\left(X_j[k]-Z_j(I)[k]\right) &\text{(A.16)}\\ &\quad+\frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[i]-\hat{X}_j(I)[i]\right)\left(Z_j(I)[k]-\hat{X}_j(I)[k]\right). &\text{(A.17)}\end{aligned}$$

For $i \in I^c$, since $\{X_j[i]-Z_j(I)[i],\ \text{for } j=1,\dots,n\}$ are iid random variables with finite second moment (assumption **HP2**), we have that, $\{(X_j[i]-Z_j(I)[i])(X_j[k]-Z_j(I)[k]),\ \text{for } j=1,\dots,n\}$ are iid random variables with finite first moment, and using the SLLN we have that

$$\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]\right)\left(X_j[k]-Z_j(I)[k]\right)\to E\left(\left(X[i]-Z(I)[i]\right)\left(X[k]-Z(I)[k]\right)\right)\quad \text{a.s.}$$

From Cauchy−Schwarz inequality, we have that,

$$\left|\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]\right)\left(Z_j(I)[k]-\hat{X}_j(I)[k]\right)\right|\leq$$

$$\left(\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]\right)^2\right)^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[k]-\hat{X}_j(I)[k]\right)^2\right)^{1/2}.$$

The first term on the right–hand side converges a.s. to $E((X[i]-Z[i])^2)$ by the SLLN, while the second one converges a.s. to 0 by assumption **H1**.

Following the same idea we have that (A.16) converges to 0 a.s.

Finally, to prove that (A.17) vanishes, we apply the Cauchy−Schwarz inequality again,

$$\left|\frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[i]-\hat{X}_j(I)[i]\right)\left(Z_j(I)[k]-\hat{X}_j(I)[k]\right)\right|\leq$$

$$\left(\frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[i]-\hat{X}_j(I)[i]\right)^2\right)^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[k]-\hat{X}_j(I)[k]\right)^2\right)^{1/2}.$$

Both terms converge a.s. to 0 by assumption **H1**. □

Acknowledgements. This work has been partially supported by Grant pict 2008-0921 from ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica), Argentina. The authors are grateful to two anonymous referees and the Associate Editors for the careful reading and insightful comments that substantially improved the paper.

References

[1] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981).

[2] J. Dauxois, A. Pousse and Y. Romain, Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *J. Multivariate Anal.* **12** (1982) 136–154.

[3] K.A. De Jong and W.M. Spears, Using genetic algorithms to solve NP-complete problems. In *Proc. of the Third International Conference on Genetic Algorithms*. Edited by J.D. Schaffer (1989) 124–132.

[4] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, Least angle regression. With discussion, and a rejoinder by the authors. *Ann. Stat.* **32** (2004) 407–499.

[5] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96** (2001) 1348–1361.

[6] R. Fraiman, A. Justel and M. Svarc, Selection of variables for cluster analysis and classification rules. *J. Am. Stat. Assoc.* **103** (2008) 1294–1303.

[7] C. Fraley and A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* **97** (2002) 611–631.

[8] C. Fraley and A.E. Raftery, *MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering.* Technical Report No. 504, Department of Statistics, University of Washington (2009).

[9] Y. Gimenez, *Selección de variables para datos multivariado y para datos funcionales.* Ph.D. thesis (2015). Available at http://cms.dm.uba.ar/academico/carreras/doctorado/TesisYaninaGimenez.pdf

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

[11] W.K. Härdle and L. Simar, *Applied Multivariate Statistical Analysis.* Springer Verlag, Berlin (2007).

[12] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning. Data mining, Inference and Prediction.* Springer Verlag, Berlin (2001).

[13] X. He and P. Shi, Bivariate tensor-product B-splines in partly linear models. *J. Multivariate Anal.* **58** (1996) 162–181.

[14] J. Hoeting, A.E. Raftrey and D. Madigan, Bayesian variable and transformation selection in linear regression. *J. Comput. Graph. Statist.* **11** (2002) 485–507.

[15] I.T. Jolliffe, *Principal Components Analysis*, 2nd edition. Springer Verlag, Berlin (2002).

[16] R. Li, and G. Gong, K-NN nonparametric estimation of regression functions in the presence of irrelevant variables. *Econom. J.* **11** (1987) 396–408.

[17] R.A. Marona, D.R. Martin and V.Y. Yohai, *Robust Statistics. Theory and Methods.* Wiley (2006).

[18] G.P. McCabe, Principal variables. *Technometrics* **26** (1984) 137–144.

[19] G.A.F. Seber and A.J. Lee, Linear regression analysis, Second edition. *Wiley series in probability and statistics* (2005).

[20] L.J. Snell, *Topics in Contemporary Probability and its Applications.* CRC Press (1995).

[21] R. Tibshirani, Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **58** (1996) 267–288.

[22] D.M. Witten and R. Tibshirani, Testing significance of features by lassoed principal components. *Ann. Appl. Stat.* **2** (2008) 986–1012.

[23] D.M. Witten and R. Tibshirani, A framework for feature selection in clustering. *J. Am. Stat. Assoc.* **105** (2010) 713–726.

[24] D.M. Witten, R. Tibshirani and T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** (2009) 515–534.

[25] C.H. Zhang, Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **38** (2010) 894–942.

[26] H. Zou, T. Hastie and R. Tibshirani, Sparse principal components analysis. *J. Comput. Graph. Stat.* **15** (2006) 265–286.