South African Statistical Journal
Vol. 57, No. 1, pp. 27–44
https://doi.org/10.37920/sasj.2023.57.1.3

Variable selection in multivariate linear regression with random predictors

Alban Mbina Mbina[1], *Guy Martial Nkiet*[1] *and Assi N'guessan*[2]

[1]Département de Mathématiques et Informatique, Université des Sciences et Techniques de Masuku, Franceville, Gabon
[2]Laboratoire Paul Painlevé UMR CNRS 8524, Université de Lille, 59655, Villeneuve d'Ascq, France

We propose a method for variable selection in multivariate regression with random predictors. This method is based on a criterion that permits to reduce the variable selection problem to a problem of estimating a suitable set. Then, an estimator for this set is proposed and the resulting method for selecting variables is shown to be consistent. A simulation study that permits to study several properties of the proposed approach and to compare it with existing methods is given.

Keywords: Multivariate linear regression, Selection criterion, Variable selection.

1. Introduction

The problem of variable selection is an old and important problem in statistics, and several approaches have been proposed to deal with it for various methods of multivariate statistical analysis, including linear regression analysis. Surveys on earlier works in this field may be found in Hocking (1976), Thomson (1978a), Thomson (1978b), Breiman and Spector (1992), and some monographs on this topic are avalaible (e.g., Linhart and Zucchini, 1986; Miller, 1990). Most of the methods that have been proposed for variable selection in linear regression deal with the case of fixed design, where the covariates are assumed to be nonrandom. For this case, many selection criteria have been introduced in the literature. These include the final prediction error (FPE) criterion (Thomson, 1978a,b; Shibata, 1984; Shao, 1993; Zhang, 1993), cross-validation (Shao, 1993; Zhang, 1993), Akaike information criterion (AIC) and Mallows's C_p type criterion (e.g., Fujikoshi and Sato, 1997), the prediction error criterion (Fujikoshi et al., 2011). There is just a few works dealing with the case where the covariates are random which arises in many regression applications where the covariates values can only be observed and are not controllable. Differences between the two cases have been highlighted in Thomson (1978a) and Breiman and Spector (1992) where it is recognised that methods for the fixed design case do not perform in the same way in the random design case. That is why they introduced modifications of criteria used for fixed design in order to deal with the case of random design. Later, variable selection for linear regression with random design was considered in Zheng and Loh (1997) and Nkiet (2001) but only for univariate models in which the response is a real random variable.

Corresponding author: Guy Martial Nkiet (guymartial.nkiet@univ-masuku.org)
MSC2020 subject classifications: 62J05, 62H12

However, it often occurs in applications that one has to consider a set of response variables for the same predictors. For such situations, multivariate linear regression offers an adequate framework considered in An et al. (2013) which introduced a method for selecting both predictors and responses based on re-casting the multivariate regression problem as a classical canonical correlation analysis (CCA) problem for which a least squares type formulation is constructed, and applying an adaptive LASSO type penalty together with a selection criterion of the type of Bayesian information criterion (BIC). Multivariate linear regression has also been considered for variable selection purposes in Yuan et al. (2007), Giraud (2011) and Lu et al. (2012).

In this paper we extend the approach of Nkiet (2001) to the case of multivariate linear regression. In Section 2, the multivariate regression model that is used is presented as well as a statement of the variable selection problem. Then, we introduce a criterion by means of which the variable selection problem reduces to that of estimating a suitable set. In Section 3, an estimator of this set is introduced, so achieving our method for selecting variables. It is interesting to note that the proposed approach does not require assuming some particular distribution for the random vector X of predictors nor for the error term ε; we only have to assume that the moments up to order four of these random vectors are bounded, and that the covariance matrix of X has full rank. In Section 4, we present a simulation study made in order to study several properties of the proposal and to compare it to the adaptive sparse canonical correlation analysis (ASCCA) method given in An et al. (2013) and to methods based on traditional criteria (see, e.g., Nishii, 1984). The first issue that was addressed concerns the impact of choosing penalty functions that are involved in our procedure. We show the importance of choosing an appropriate penalty function since it has significant impact on the performance of the method. But there is no theoretical investigation for this problem and it seems that more work is needed in this direction. Secondly, we studied the influence of tuning parameters on the performance of our method. The simulation results clearly show their impact on the performance, so it is of interest to apply a method defined in Section 3 for obtaining optimal values of these parameters. When using this approach, we obtain better results in the simulations than all the considered existing methods. The proofs of the main results of the paper are postponed to Section 5.

2. Model and criterion for selection

We consider the multivariate linear regression model given by:

$$Y = BX + \varepsilon, \tag{1}$$

where X and Y are random vectors with values in $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively with $p \geq 2$ and $q \geq 2$, B is a $q \times p$ matrix of real coefficients, and ε is a random vector with values in $\mathbb{R}^q$, which is independent of X and such that $\mathbb{E}(\varepsilon) = 0$. Writing

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_q \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_q \end{pmatrix}$$

and

$$B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ b_{q1} & b_{q2} & \cdots & b_{qp} \end{pmatrix},$$

it is easily seen that model (1) is equivalent to having a set of q univariate regression models given by

$$Y_i = \sum_{j=1}^{p} b_{ij} X_j + \varepsilon_i, \quad i = 1, \cdots, q,$$

and can also be writen as

$$Y = \sum_{j=1}^{p} X_j \mathbf{b}_{\bullet j} + \varepsilon, \tag{2}$$

where

$$\mathbf{b}_{\bullet j} = \begin{pmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{qj} \end{pmatrix}.$$

We are interested with the variable selection problem, that is identifying the X_j that are not relevant in model (2), on the basis of an i.i.d. sample $(X^{(k)}, Y^{(k)})_{1 \leq k \leq n}$ of (X, Y). We say that a variable X_j is not relevant if the corresponding coefficients vector $\mathbf{b}_{\bullet j}$ equals 0. So, putting $I = \{1, \cdots, p\}$, and denoting by $\|\cdot\|_{\mathbb{R}^k}$ the usual Euclidean norm of $\mathbb{R}^k$, we consider the subset $I_0 = \{j \in I \,/\, \|\mathbf{b}_{\bullet j}\|_{\mathbb{R}^q} = 0\}$ which is assumed to be non-empty, and we tackle the variable selection problem in model (1) as a problem of estimating the set $I_1 = I - I_0$. In order to simplify the estimation of I_1 we will first characterise it by means of a criterion which introduced below. We assume that $\mathbb{E}\left(\|X\|^2_{\mathbb{R}^p}\right) < +\infty$ and $\mathbb{E}\left(\|Y\|^2_{\mathbb{R}^q}\right) < +\infty$; then, it is possible to define the covariance matrices

$$V_1 = \mathbb{E}\left((X-\mu)(X-\mu)^\top\right) \quad \text{and} \quad V_{12} = \mathbb{E}\left((X-\mu)\,(Y-\nu)^\top\right),$$

where $\mu = \mathbb{E}(X)$, $\nu = \mathbb{E}(Y)$ and $u^\top$ denotes the transposed of u. In all of the paper, the matrix V_1 is assumed to be invertible. For any subset $J := \{i_1, \cdots, i_k\}$ of I, consider the $k \times p$ matrix defined by

$$A_J = \begin{pmatrix} a^{(J)}_{11} & a^{(J)}_{12} & \cdots & a^{(J)}_{1p} \\ a^{(J)}_{21} & a^{(J)}_{22} & \cdots & a^{(J)}_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ a^{(J)}_{k1} & a^{(J)}_{k2} & \cdots & a^{(J)}_{kp} \end{pmatrix},$$

where

$$a^{(J)}_{\ell j} = \begin{cases} 1 & \text{if } j = i_\ell \\ 0 & \text{if } j \neq i_\ell \end{cases}, \; 1 \leq \ell \leq k, \, 1 \leq j \leq p.$$

This matrix transforms any vector $x = (x_1, \cdots, x_p)^\top$ to the vector $A_J x = (x_{i_1}, \cdots, x_{i_k})^\top$ of lower dimension, whose components are selected from the initial vector x by taking only the components

x_i such that $i \in J$. We are seeking for a criterion allowing to measure the degree of relevance of variables whose indices belong to J. From (1), we obtain $\nu = B\mu$ and, therefore,

$$\begin{aligned} V_{12} &= \mathbb{E}\left((X-\mu)\,(B(X-\mu))^\top\right) + \mathbb{E}\left((X-\mu)\,\varepsilon^\top\right) \\ &= \mathbb{E}\left((X-\mu)\,(X-\mu)^\top\right) B^\top + \mathbb{E}\,(X-\mu)\,\mathbb{E}\,(\varepsilon)^\top \\ &= V_1 B^\top, \end{aligned}$$

so $B^\top = V_1^{-1} V_{12}$. If we only consider in (1) the selected variables, that is the variables whose indices belong to J, we just have to replace X by $A_J X$. Then, since the related covariance and cross-covariance matrices are

$$V_1^J = \mathbb{E}\left((A_J(X-\mu))(A_J(X-\mu))^\top\right) = A_J \mathbb{E}\left((X-\mu)(X-\mu)^\top\right) A_J^\top = A_J V_1 A_J^\top$$

and

$$V_{12}^J = \mathbb{E}\left((A_J(X-\mu))(Y-\nu)^\top\right) = A_J \mathbb{E}\left((X-\mu)(Y-\nu)^\top\right) = A_J V_{12},$$

we deduce from a reasoning similar to that given before that the corresponding matrix of model coefficients is the $q \times k$ matrix B_J given by

$$B_J^\top = \left(V_1^J\right)^{-1} V_{12}^J = \left(A_J V_1 A_J^\top\right)^{-1} A_J V_{12}.$$

This matrix contains the coefficients of model (1) when only the variables corresponding to J are used; these variables are relevant if B_J is closed to B. But B_J can not be compared to B because it has different dimensions; we consider rather the $q \times p$ matrix $\widetilde{B} = B_J A_J$ which contains the same terms as B_J to which are added the same null columns as B. Then, a measure of the degree of relevance of variables whose indices belong to J is given by a distance between B and $\widetilde{B}$. Using the matrix norm $\|\cdot\|_{V_1}$ given by

$$\|A\|_{V_1}^2 = \operatorname{tr}\left(V_1 A^\top A V_1\right) = \|V_1 A^\top\|,$$

where $\|\cdot\|$ denotes the usual matrix norm defined by $\|A\|^2 = \operatorname{tr}(AA^\top)$, leads to the criterion

$$\xi_J = \|B - \widetilde{B}\|_{V_1}^2 = \|V_1(B^\top - \widetilde{B}^\top)\| = \|V_{12} - V_1 \Pi_J V_{12}\|,$$

where $\Pi_J = A_J^\top \left(A_J V_1 A_J^\top\right)^{-1} A_J$. Naturally, the required set I_1 may be obtained by minimising ξ_J over all subsets J of I. But such an approach will not be used since it may lead to high computational cost. We will rather use another strategy that allows a faster procedure. This is made possible thanks to the following lemma:

Lemma 1. *We have $I_1 \subset J$ if, and only if, $\xi_J = 0$.*

This lemma shows that I_1 is the minimal subset of I for which the above criterion vanishes. Furthermore, an index i belongs to I_0 if and only if $\xi_{K_i} = 0$, where K_i is the subset of I obtained by removing i from I, that is $K_i = I - \{i\}$. Indeed,

$$i \in I_0 \Leftrightarrow \{i\} \subset I_0 \Leftrightarrow I - I_0 \subset I - \{i\} \Leftrightarrow I_1 \subset K_i \Leftrightarrow \xi_{K_i} = 0,$$

the last equivalence coming from Lemma 1. Consequently, I_1 is the set of indices of I for which we have $\xi_{K_i} > 0$. So it can be described from sorting the ξ_{K_i} in decreasing order. Indeed, this approach allows to highlight the non-zero tems and those which are zero, and I_1 is made up of the indices corresponding to the non-zero terms. More specically, there exist integers $\nu_1, \cdots, \nu_p$ and an integer $d \in \{1, \cdots, p-1\}$ such that:

$$\xi_{K_{\nu_1}} \geq \xi_{K_{\nu_2}} \geq \cdots \geq \xi_{K_{\nu_d}} > 0 = \xi_{K_{\nu_{d+1}}} = \cdots = \xi_{K_{\nu_p}}, \tag{3}$$

with

$$\nu_i < \nu_\ell \quad \text{if} \quad \xi_{K_i} = \xi_{K_\ell} \quad \text{and} \quad i < \ell. \tag{4}$$

Note that if some of the ξ_{K_i} are tied there are several ways to order them in nonincreasing order, but there is a unique way for which (4) is satisfied. So (4) ensures the unicity of the ν_ℓ. From (3) we get

$$I_1 = \{\nu_1, \cdots, \nu_d\}, \tag{5}$$

so this set is completely determined if one determines the ν_ℓ and d. The ν_ℓ are determined by ordering the ξ_{K_i} in nonincreasing order so that (4) holds. On the other hand, putting $J_\ell = \{\nu_1, \cdots, \nu_\ell\}$, we see that $I_1 \subset J_\ell$ for all $\ell \in \{d, \cdots, p\}$. Then, from Lemma 1, we deduce that $\xi_{J_\ell} = 0$ if $\ell \in \{d, \cdots, p\}$, and $\xi_{J_\ell} > 0$ if $\ell \in \{1, \cdots, d-1\}$. Hence

$$d = \min\left\{\ell \in I \,/\, \xi_{J_\ell} = \min_{i \in I}\left(\xi_{J_i}\right)\right\}. \tag{6}$$

Selection of variables in (1) is reduced to the estimation of the subset I_1 from an i.i.d. sample $\left\{(X^{(k)}, Y^{(k)})\right\}_{1 \leq k \leq n}$ of (X, Y), which, according to (5), amounts to estimating the ν_ℓ and d. For doing that, we first have to estimate the above criterion. We consider the sample means

$$\overline{X}^{(n)} = n^{-1} \sum_{k=1}^{n} X^{(k)}, \quad \overline{Y}^{(n)} = n^{-1} \sum_{k=1}^{n} Y^{(k)},$$

and the empirical covariance matrices

$$\widehat{V}_1^{(n)} = n^{-1} \sum_{k=1}^{n} (X^{(k)} - \overline{X}^{(n)})\,(X^{(k)} - \overline{X}^{(n)})^\top,$$

and

$$\widehat{V}_{12}^{(n)} = n^{-1} \sum_{k=1}^{n} (X^{(k)} - \overline{X}^{(n)})(Y^{(k)} - \overline{Y}^{(n)})^\top.$$

Then, for any $J \subset I$, an estimator of ξ_J is given by

$$\widehat{\xi}_J^{(n)} = \|\widehat{V}_{12}^{(n)} - \widehat{V}_1^{(n)} \widehat{\Pi}_J^{(n)} \widehat{V}_{12}^{(n)}\|, \tag{7}$$

where $\widehat{\Pi}_K^{(n)} = A_J^\top (A_J \widehat{V}_1^{(n)} A_J^\top)^{-1} A_J$. Since the above empirical covariance matrices are consistent estimators of the related covariances matrices, it is easy to see that $\widehat{\xi}_J^{(n)}$ is a consistent estimator of ξ_J. We will consider this estimator for estimating I_1.

3. Selection of variables

In this section we propose a new method for selecting variables in the model (1). More precisely, the estimator given in (7) is used in order to obtain an estimator of I_1, so achieving variable selection. The latter estimator depends on two tuning parameters wich are introduced below, so a procedure for choosing optimal values for these parameters, based on K-fold cross-validation, is introduced.

3.1 Estimation of I_1

The subset I_1 has been characterised in (5) as depending on the ν_ℓ and d, so it would be natural to obtain an estimator of this subset by estimating the ν_ℓ and d from a use of the $\widehat{\xi}_{K_i}^{(n)}$ as it was done in (3), (4) and (6) with the ξ_{K_i}. But we found that such an approach does not make it possible to obtain consistency of the resulting estimators because of possible ties in the values of the $\widehat{\xi}_{K_i}^{(n)}$. This is why we rather adopt an approach consisting of adding to the $\widehat{\xi}_{K_i}^{(n)}$ positive penalisation terms tending towards 0 as $n \to +\infty$, in order to obtain consistent estimators of the ξ_{K_i} for which there are no ties. More precisely, by putting

$$\widehat{\phi}_i^{(n)} = \widehat{\xi}_{K_i}^{(n)} + \frac{f\,(i)}{n^\alpha}, \tag{8}$$

where $\alpha \in \,]0, 1/2[$ and f is a strictly decreasing function from I to $\mathbb{R}_+$, we obtain a consistent estimator $\widehat{\phi}_i^{(n)}$ of ξ_{K_i} for which the above required property is satisfied. Indeed, if $\widehat{\xi}_{K_i}^{(n)} = \widehat{\xi}_{K_j}^{(n)}$ for $i \neq j$ then $\widehat{\phi}_i^{(n)} \neq \widehat{\phi}_j^{(n)}$. By ordering the $\widehat{\phi}_i^{(n)}$ in nonincreasing order we obtain estimates $\widehat{\nu}_1, \cdots, \widehat{\nu}_p$ of $\nu_1, \cdots, \nu_p$ respectively; they satisfy

$$\widehat{\phi}_{\widehat{\nu}_1} \geq \widehat{\phi}_{\widehat{\nu}_2} \geq \cdots \geq \widehat{\phi}_{\widehat{\nu}_p}.$$

In order to estimate d, we will consider a consistent estimator of ξ_{J_ℓ} obtained, for the same reasons as above, from an appropriate penalisation of $\widehat{\xi}_{\widehat{J}_i}^{(n)}$ by a term which tends towars 0 as $n \to +\infty$, where $\widehat{J}_i = \{\widehat{\nu}_1, \cdots, \widehat{\nu}_i\}$. More specifically, we put

$$\widehat{\psi}_i^{(n)} = \widehat{\xi}_{\widehat{J}_i}^{(n)} + \frac{g\,(\widehat{\nu}_i)}{n^\beta}, \tag{9}$$

where $\beta \in \,]0, 1[$ and g is a strictly increasing function from I to $\mathbb{R}_+$. Then, copying (6), we define an estimate $\widehat{d}$ of d by

$$\widehat{d} = \arg\min_{i \in I} \left(\widehat{\psi}_i^{(n)}\right),$$

and we estimate I_1 by the set

$$\widehat{I}_1^{(n)} = \left\{\widehat{\nu}_1, \widehat{\nu}_2, \cdots, \widehat{\nu}_{\widehat{d}}\right\},$$

so achieving our variable selection procedure. The following theorem establishes consistency for this procedure:

Theorem 1. *We assume that* $I_0 \neq \emptyset$, $\mathbb{E}\left(\|X\|_{\mathbb{R}^p}^4\right) < \infty$ *and* $\mathbb{E}\left(\|Y\|_{\mathbb{R}^q}^4\right) < \infty$.

$$\lim_{n \to +\infty} P\left(\widehat{I}_1^{(n)} = I_1\right) = 1.$$

Remark 1. Technical arguments for the proofs explain the choice of f, g, α and β with the related properties. Indeed:

(i) in the proof of Lemma 3 we have, for instance, an inequality of the form

$$\left| n^{\alpha} \left(\widehat{\xi}^{(n)}_{K_{\nu_i}} - \widehat{\xi}^{(n)}_{K_{\nu_{i+1}}} \right) \right| \quad \leq \quad n^{\alpha - \frac{1}{2}} U_n,$$

from which we want to prove that $|n^{\alpha}(\widehat{\xi}^{(n)}_{K_{\nu_i}} - \widehat{\xi}^{(n)}_{K_{\nu_{i+1}}})|$ converges in probability to 0 as $n \longrightarrow +\infty$. So we have to take $\alpha < 1/2$ since it is proved that U_n converges in distribution to an appropriate random variable.

(ii) For proving Theorem 1, a similar argument leads to $0 < \beta < 1$ (see the proof of Theorem 3.1 in Nkiet, 2012). Further, we want to obtain $f(\nu_i) - f(\nu_{i+1}) > 0$ and $g(\nu_i) - g(\nu_s) > 0$ for $(i, s) \in I^2$ satisfying $\nu_i < \nu_{i+1}$ and $\nu_i > \nu_s$. That is why f (resp. g) is taken as a strictly decreasing (resp. increasing) function.

Remark 2. Theorem 1 holds for any numbers α and β satisfying $0 < \alpha < 1/2$ and $0 < \beta < 1$. So our procedure depends on these two parameters and its performance could vary depending on their values. The simulation results presented in the following section show the impact of these values. This brings us back to the problem of choosing optimal values for α and β. A method for doing that, based on cross-validation procedure, is proposed in the following subsection.

Remark 3. The choice of the penalty functions f and g may also have an impact on the performance of our method. Following Kundu and Murali (1976), we studied in the simulations this impact by varying these functions from a given list.

3.2 Choosing optimal tuning parameters

The procedure for variable selection introduced in the preceding section depends on two tuning parameters α and β. These parameters may have influence on the performance of our method; then the problem of choosing these parameters naturally occurs. We propose a method for obtaining an optimal choice of (α, β) based on K-fold cross-validation (with $K \in \mathbb{N}^*$) used in order to minimise prediction loss. More precisely, consider a partition $\{\mathcal{S}_1, \cdots, \mathcal{S}_K\}$ of the set $\mathcal{S} = \{1, \cdots, n\}$, each $\mathcal{S}_\ell$ having the same size $m \in \mathbb{N}^*$ (then, $n = K\,m$). For each ℓ in $\{1, \cdots, K\}$, after removing the ℓth subset $\mathcal{S}_\ell$ from $\mathcal{S}$, we apply our method for selecting variable on the remaining sample $\{(X^{(k)}, Y^{(k)});\ k \in \mathcal{S} - \mathcal{S}_\ell\}$ with a given value for (α, β); this leads to estimates $\widehat{d}$ and $\widehat{I}_1$. Putting $\mathcal{S}_\ell = \{s_1, \cdots, s_m\}$, we consider the $m \times \widehat{d}$ and $m \times q$ matrices given by

$$\mathbf{X}^{(\ell)}_{s,\alpha,\beta} = \begin{pmatrix} (A_{\widehat{I}_1} X^{(s_1)})^\top \\ \vdots \\ (A_{\widehat{I}_1} X^{(s_m)})^\top \end{pmatrix} \quad \text{and} \quad \mathbf{Y}^{(\ell)} = \begin{pmatrix} (Y^{(s_1)})^\top \\ \vdots \\ (Y^{(s_m)})^\top \end{pmatrix}$$

that contain respectively the observations of the variables of X that have been selected on the units belonging to $\mathcal{S}_\ell$, and the observations of the variables of Y on the units belonging to $\mathcal{S}_\ell$. Then, we consider the linear prediction

$$\widehat{\mathbf{Y}}^{(\ell)}_{s,\alpha,\beta} = \mathbf{X}^{(\ell)}_{s,\alpha,\beta} \left((\mathbf{X}^{(\ell)}_{s,\alpha,\beta})^\top \mathbf{X}^{(\ell)}_{s,\alpha,\beta} \right)^{-1} (\mathbf{X}^{(\ell)}_{s,\alpha,\beta})^\top \mathbf{Y}^{(\ell)},$$

and the ℓth prediction loss

$$\mathrm{PL}^{(\ell)}(\alpha,\beta) = \frac{1}{m}\|\mathbf{Y}^{(\ell)} - \widehat{\mathbf{Y}}^{(\ell)}_{s,\alpha,\beta}\|^2.$$

This permits to define the cross-validation index

$$CV(\alpha,\beta) = \frac{1}{K}\sum_{\ell=1}^{K}\mathrm{PL}^{(\ell)}(\alpha,\beta),$$

which has to be minimised in order to obtain optimal values $(\alpha_{opt},\beta_{opt})$ for (α,β) by

$$(\alpha_{opt},\beta_{opt}) = \underset{(\alpha,\beta)\in]0,1/2[\times]0,1[}{\operatorname{argmin}} CV(\alpha,\beta). \tag{10}$$

Note that, if we take $K = n$ and $\mathcal{S}_\ell = \{\ell\}$, for $\ell = 1,\cdots,n$, then we obtain the usual leave-one-out cross-validation. Nevertheless, K-fold cross-validation with $K < n$ is generally prefered since it could give better results than leave-one-out cross-validation.

3.3 Algorithm of the proposed method

Our proposal is to achieve variable selection in the model (1) from the following steps:

(1) Partition the sample $\{(X^{(k)},Y^{(k)})\}_{1\le k\le n}$ into two subsamples: a training sample $\{(X^{(k)},Y^{(k)})\}_{k\in\mathcal{S}_1}$ and a test sample $\{(X^{(k)},Y^{(k)})\}_{k\in\mathcal{S}_2}$, where $\mathcal{S}_1\cap\mathcal{S}_1 = \emptyset$ and $\mathcal{S}_1\cup\mathcal{S}_1 = \{1,\cdots,n\}$.

(2) Use the training sample for choosing optimal pair of tuning parameters $(\alpha_{opt},\beta_{opt})$ as defined in (10), from K-fold cross-validation as defined in Section 3.2.

(3) Use the test sample for computing the estimate $\widehat{I_1}$ as defined in Section 3.1 with tuning parameters equal to $(\alpha_{opt},\beta_{opt})$ obtained in the previous step.

4. Simulations

In this section, we report results of simulations made for studying properties of the proposed method. Several issues are adressed: the influence of the penalty functions introduced in (8) and (9) and that of the parameters α and β on the performance of the proposed method, and comparison with the adaptive sparse canonical correlation analysis (ASCCA) method of An et al. (2013) and traditional methods based on AIC, BIC, Mallows's C_p, FPE criterion, prediction sum of squares (PSS) criterion and generalised information criterion (GIC) (for a review, see Nishii, 1984). Each data set was generated as follows: $X^{(k)}$ is generated from a multivariate normal distribution in $\mathbb{R}^7$ with mean 0 and covariance matrix $\Lambda = (\lambda_{ij})_{1\le i,j\le p}$ given by $\lambda_{ij} = \rho^{|i-j|}$ for any $1\le i,j\le 7$, where $\rho\in[0,1]$. The corresponding response $Y^{(k)}$ is generated according to (1) with

$$B = \begin{pmatrix} 3 & 0 & 0 & 1.5 & 0 & 0 & 2 \\ 4 & 0 & 0 & 2.5 & 0 & 0 & -1 \\ 5 & 0 & 0 & 0.5 & 0 & 0 & 3 \\ 6 & 0 & 0 & 3 & 0 & 0 & 1 \\ 7 & 0 & 0 & 6 & 0 & 0 & 4 \end{pmatrix}$$

and the related error term $\varepsilon^{(k)}$ is generated from a multivariate normal distribution in $\mathbb{R}^7$ with mean 0 and covariance matrix $\tau^2 I_7$, where I_7 denotes the 7×7 identity matrix. For this example, the true set of relevant variables is $I_1 = \{1, 4, 7\}$. We simulate 1000 independent replications of samples generated as indicated above, over which four criteria are calculated in order to assess performance of the methods: (i) the average of prediction loss (PL); (ii) the proportion of equality (PE) of the set of selected variables to the true set of relevant variables, that is $\widehat{I_1} = I_1$; (iii) the proportion of inclusion (PI) of the true set of relevant variables in the set of selected variables to the true set of relevant variables that is $I_1 \subset \widehat{I_1}$; (iv) the average of number of selected variables (NV). For computing PL two independent data sets are generated for each replication: training data and test data, each with sample size $n = 100, 200$. The training data is used for selecting variables and the test data is used for computing prediction loss (PL). This is done from the following steps:

- we simulate a training data set in a $n \times p$ matrix $\mathbf{X}_1$ whose rows follow the distribution $N(0, \Lambda)$, and a $n \times q$ matrix $\mathbf{E}_1$ whose rows follow the distribution $N(0, \tau^2 I_q)$. Then we take $\mathbf{Y}_1 = \mathbf{X}_1 B^\top + \mathbf{E}_1$, and we calculate $\widehat{d}$ and $\widehat{I_1} := \widehat{I}_1^{(n)}$ on the basis of $(\mathbf{Y}_1, \mathbf{X}_1)$;

- we simulate a test data set in matrices $\mathbf{X}_2$, $\mathbf{E}_2$ and $\mathbf{Y}_2$ as above. Letting $\widehat{\mathbf{X}}$ be the $n \times \widehat{d}$ matrix obtained from $\mathbf{X}_2$ by considering only the columns corresponding to $\widehat{I_1}$, the aforementioned prediction loss is
$$\mathrm{PL} = \frac{1}{n} \|\mathbf{Y}_2 - \widehat{\mathbf{X}}\widehat{\mathbf{B}}^\top\|^2,$$
where $\widehat{\mathbf{B}}^\top = (\widehat{\mathbf{X}}^\top \widehat{\mathbf{X}})^{-1} \widehat{\mathbf{X}}^\top \mathbf{Y}_2$.

4.1 Influence of penalty functions

In order to evaluate the impact of penalty functions introduced in (8) and (9) on the performance of our method we took $f_n(i) = n^{-1/4}/h_\ell(i)$ and $g_n(i) = n^{-1/4} h_\ell(i)$, $\ell = 1, \cdots, 13$, where the h_ℓ are functions used in Kundu and Murali (1996) for studying impact of penalty functions on model selection in linear regression. More precisely, $h_1(x) = x$, $h_2(x) = x^{0.1}$, $h_3(x) = x^{0.5}$, $h_4(x) = x^{0.9}$, $h_5(x) = x^{10}$, $h_6(x) = \ln(x)$, $h_7(x) = \ln(x)^{0.1}$, $h_8(x) = \ln(x)^{0.5}$, $h_9(x) = \ln(x)^{0.9}$, $h_{10}(x) = x \ln(x)$, $h_{11}(x) = (x \ln(x))^{0.1}$, $h_{12}(x) = (x \ln(x))^{0.5}$, $h_{13}(x) = (x \ln(x))^{0.9}$. For the variance of ε, we took $\tau = 1$, and we considered three cases for generating data: (i) $\rho = 0.0$, (ii) $\rho = 0.5$ and (iii) $\rho = 0.95$, and we took $n = 100$, $\alpha = 0.45$, $\beta = 0.6$. The results are given in Table 1. These results clearly indicate the importance of choosing an appropriate penalty function since it has an impact on the performance of the method. Indeed, the worst results are obtained with h_5. On the other hand, h_1, h_4, h_{10}, h_{12} and h_{13} give worse results than h_2, h_3, h_6, h_7, h_8 and h_{11} whereas these latter functions behave very similarly. Comparing the effects of correlation of the X_i, it is observed that, for most functions, when the correlation changes from $\rho = 0.0$ to $\rho = 0.5$ the effect is not significant but when the correlation is increased to $\rho = 0.95$ the performance drops and the worst affected are h_1, h_3, h_4, h_6, h_{12} and h_{13}. Finally, choosing the proper penalty functions appears as one of the most important problems in practice, and its seems that more work is needed in this direction. Although no theoretical justifications can be given, we recommend using h_2, h_7, h_8, h_9 or h_{11} since these functions gave the best results in our simulations.

Table 1. Average of prediction loss (PL) over 1000 replications with different penalty functions $f_n = n^{-\alpha}/h_\ell$ and $g_n = n^{-\beta}h_\ell$, $\ell = 1, \cdots, 13$ with $\alpha = 0.45$ and $\beta = 0.6$, $n = 100$.

	PL		
Function	$\rho = 0.0$	$\rho = 0.5$	$\rho = 0.95$
h_1	0.02657	0.19909	2.68593
h_2	0.00068	0.00095	0.00298
h_3	0.00035	0.00075	0.00940
h_4	0.00533	0.01949	2.65396
h_5	114.908	117.487	54.8840
h_6	0.00094	0.00070	0.00524
h_7	0.00059	0.00107	0.00413
h_8	0.00061	0.00066	0.00396
h_9	0.00055	0.00115	0.00458
h_{10}	2.58811	5.63562	2.71928
h_{11}	0.00072	0.00059	0.00451
h_{12}	0.00058	0.00205	1.68564
h_{13}	1.95441	2.38247	2.71770

4.2 Influence of parameters α and β

Since tuning parameters may have impact on the performance of a statistical procedure, it is important to study their influence. That is why numerical experiments were made in order to appreciate the influence of α and β on the performance of our method. For doing that, we made simulations by taking

$$f_n(i) = n^{-\alpha}/h_7(i) \quad \text{and} \quad g_n(i) = n^{-\beta}h_7(i), \tag{11}$$

with $\alpha = 0.1, 0.2, 0.3, 0.4, 0.45$, and β varying in $[0, 1[$. The choice of h_7 in these penalty functions is motivated by the fact that it was one of the functions that gave the best results for PL in the previous study. The results are reported in Fig. 1 (a)–(b) and suggest that the parameters α and β have an impact on the performance of the method. Indeed, the curves obtained for PL vary as α varies. Further, for a fixed α, PL decreases as β increases in $[0, 0.4[$. Finally, choosing the proper values for α and β is also an important issue to address in practice. So, choosing optimal values for these parameters by using the procedure indicated in Section 3.2 is crucial for ensuring a good performance of our method.

4.3 Comparison with existing methods

In this section, we compare our method (OM) to the ASCCA method of An et al. (2013) and the methods based on AIC, BIC, C_p, FPE and GIC criteria (see Nishii, 1984). These latter methods select the variables that minimise these criteria among all subsets of variables. The average of prediction loss (PL), the proportion of good selection (PE), the proportion of inclusion (PI) and the average of number of selected variables (NV), over 1000 replications, are then used as a measures of the performances of all the methods. For computing PL, the following approach is used: for each of

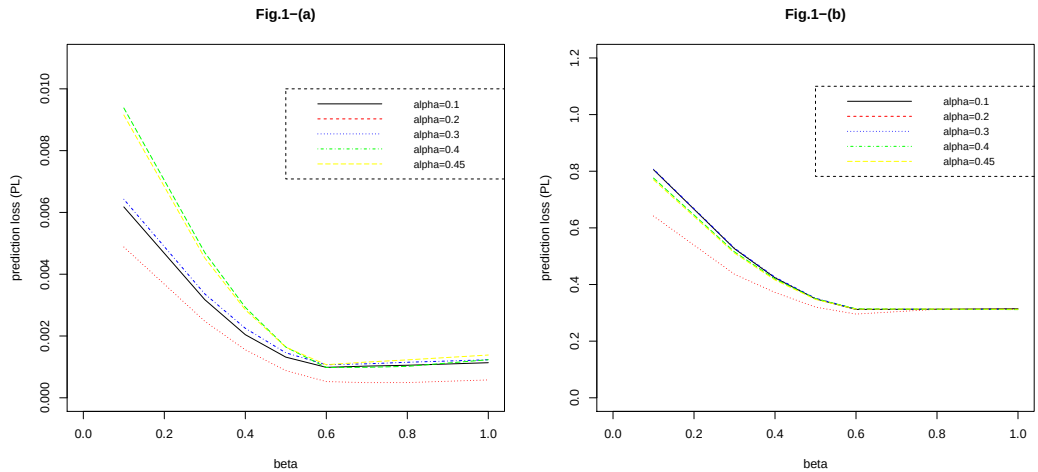

Figure 1. Average of PL over 1000 replications versus β for different values of α. (a) $n = 100$ and (b) $n = 300$.

the 1000 independent replications, the sample is partitioned into a training sample and a test sample, and:

(i) the training sample is used for selecting variables from all the methods; our method is used with penalty functions given in (11) with optimal (α, β) obtained by using K-fold cross-validation, with $K = 5$, as indicated in Section 3.3;

(ii) the test sample is then used for computing PL for all the methods.

Our method was performed by using the penalty functions given in (11). Table 2 reports the results for average of PL. It can be seen that our method gives results that are very close to those given by the ASCCA, C_p, FPE and PSS methods. They outperform the remaining methods that also give results that are very close to each other. Concerning PE, Table 3 shows that our method gives better results than all the others. ASCCA method is closest to ours, followed by FPE and PSS methods. The remaining methods give very bad results. The results for PI are reported in Table 4; it indicates that all methods except AIC, BIC and GIC based methods get the relevant variables among the variables that are selected. The results for NV, reported in Table 5, are better for our method since the obtained values are closer to the true value of number of relevant variables.

5. Proofs

In this section, we give the proof of Lemma 1 and Theorem 1 which are the main results of the paper. For proving Theorem 1 a preliminary lemma, that is given and proved, is required.

5.1 Proof of Lemma 1

It is easy to check that, putting

$$V_{12}^{(j)} = \mathbb{E}((Y_j - \nu_j)(X - \mu)), \quad \xi_J^{(j)} = \|V_{12}^{(j)} - V_1 \Pi_J V_{12}^{(j)}\|_{\mathbb{R}^p} \quad \text{and} \quad I_1^{(j)} = \{i \in I \,/\, b_{ji} \neq 0\},$$

Table 2. Average of PL for our method (OM) with $K = 5$, and ASCCA, AIC, BIC, CP, FPE, PSS and GIC methods over 1000 replications with $n = 100$, $\rho = 0.5$.

τ	OM	ASCCA	AIC	BIC	C_p	FPE	PSS	GIC
0.1	0.00011	0.00011	0.00766	0.00766	0.00011	0.00011	0.00011	0.00766
0.2	0.00021	0.00021	0.00732	0.00732	0.00021	0.00021	0.00022	0.00732
0.3	0.00036	0.00035	0.00648	0.00648	0.00035	0.00035	0.00035	0.00648
0.4	0.00042	0.00041	0.00780	0.00780	0.00041	0.00041	0.00041	0.00818
0.5	0.00052	0.00052	0.00860	0.00053	0.00053	0.00053	0.00053	0.00860

Table 3. PE for our method (OM) with $K = 5$, and ASCCA, AIC, BIC, CP, FPE, PSS and GIC methods over 1000 replications.

$n = 100$, $\rho = 0.5$								
τ	OM	ASCCA	AIC	BIC	C_p	FPE	PSS	GIC
0.1	0.92	0.62	0.00	0.00	0.03	0.60	0.67	0.00
0.2	1.00	0.73	0.00	0.00	0.10	0.59	0.68	0.00
0.3	0.93	0.62	0.00	0.00	0.09	0.66	0.73	0.00
0.4	1.00	0.64	0.00	0.00	0.11	0.61	0.74	0.00
0.5	0.90	0.69	0.00	0.00	0.19	0.67	0.72	0.00
$n = 200$, $\tau = 0.1$								
ρ	OM	ASCCA	AIC	BIC	C_p	FPE	PSS	GIC
0.1	0.97	0.79	0.00	0.00	0.03	0.68	0.69	0.00
0.2	1.00	0.86	0.00	0.00	0.01	0.74	0.82	0.00
0.3	0.98	0.77	0.00	0.00	0.03	0.67	0.77	0.00
0.4	1.00	0.78	0.00	0.00	0.02	0.60	0.68	0.00
0.5	0.98	0.78	0.00	0.00	0.02	0.67	0.70	0.00

one has $\xi_J^2 = \sum_{j=1}^q (\xi_J^{(j)})^2$ and $I_1 = \bigcup_{j=1}^q I_1^{(j)}$. Then, ξ_J equals 0 if, and only if, for all $j \in \{1, \cdots, q\}$, one has $\xi_J^{(j)} = 0$. From Lemma 1 in Nkiet (2001), this latter property is equivalent to having $I_1^{(j)} \subset J$ for any $j \in \{1, \cdots, q\}$, which is equivalent to $I_1 \subset J$.

5.2 A preliminary lemma

We denote by $\mathscr{M}_{m,r}(\mathbb{R})$ the space of $m \times r$ matrices with real terms; when $m = r$, we simply write $\mathscr{M}_m(\mathbb{R})$ instead of $\mathscr{M}_{m,m}(\mathbb{R})$. Each element A of $\mathscr{M}_{p+q}(\mathbb{R})$ can be writen as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ \\ A_{21} & A_{22} \end{pmatrix},$$

Table 4. PI for our method (OM) with $K = 5$, and ASCCA, AIC, BIC, CP, FPE, PSS and GIC methods over 1000 replications.

$n = 100, \rho = 0.5$								
τ	OM	ASCCA	AIC	BIC	C_p	FPE	PSS	GIC
0.1	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00
0.2	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00
0.3	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00
0.4	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00
0.5	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00
$n = 200, \tau = 0.1$								
ρ	OM	ASCCA	AIC	BIC	C_p	FPE	PSS	GIC
0.1	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00
0.2	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00
0.3	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00
0.4	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00
0.5	1.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00

where $A_{11} \in \mathscr{M}_p(\mathbb{R})$, $A_{12} \in \mathscr{M}_{p,q}(\mathbb{R})$, $A_{21} \in \mathscr{M}_{q,p}(\mathbb{R})$ and $A_{22} \in \mathscr{M}_q(\mathbb{R})$. Then we consider the projectors

$$P_1 \,:\, A \in \mathscr{M}_{p+q}(\mathbb{R}) \mapsto A_{11} \in \mathscr{M}_p(\mathbb{R}) \quad \text{and} \quad P_2 \,:\, A \in \mathscr{M}_{p+q}(\mathbb{R}) \mapsto A_{12} \in \mathscr{M}_{p,q}(\mathbb{R}),$$

and we have:

Lemma 2. *We have* $\sqrt{n}\,\widehat{\xi}_J^{(n)} = \|\widehat{\Psi}_J^{(n)}(\widehat{H}^{(n)}) + \sqrt{n}\,\delta_J\|$, *where* $\delta_J = V_{12} - V_1\Pi_J V_{12}$, $(\widehat{\Psi}_J^{(n)})_{n\in\mathbb{N}^*}$ *is a sequence of random linear maps from* $\mathscr{M}_{p+q}(\mathbb{R})$ *to* $\mathscr{M}_{p,q}(\mathbb{R})$ *which converges almost surely, as* $n \to +\infty$, *to the linear map* Ψ_J *given by*

$$\Psi_J(A) = P_2(A) - P_1(A)\Pi_J V_{12} + V_1\Pi_J P_1(A)\Pi_J V_{12} - V_1\Pi_J P_2(A),$$

and $(\widehat{H}^{(n)})_{n\in\mathbb{N}^*}$ *is a sequence of random variables with values in* $\mathscr{M}_{p+q}(\mathbb{R})$ *which converges in distribution to random variable* H *having a centred normal distribution in* $\mathscr{M}_{p+q}(\mathbb{R})$.

Proof. We have

$$\sqrt{n}\widehat{\xi}_J^{(n)} = \|\sqrt{n}(\widehat{V}_{12}^{(n)} - V_{12}) - \sqrt{n}(\widehat{V}_1^{(n)} - V_1)\widehat{\Pi}_J^{(n)}\widehat{V}_{12}^{(n)} - V_1\left(\sqrt{n}(\widehat{\Pi}_J^{(n)} - \Pi_J)\right)\widehat{V}_{12}^{(n)} - V_1\Pi_J\left(\sqrt{n}(\widehat{V}_{12}^{(n)} - V_{12})\right) + \sqrt{n}\delta_J\|,$$

and since

$$\begin{aligned}\widehat{\Pi}_J^{(n)} - \Pi_J &= A_J^\top\left((A_J\widehat{V}_1^{(n)}A_J^\top)^{-1} - (A_JV_1A_J^\top)^{-1}\right)A_J \\ &= A_J^\top\left(-(A_J\widehat{V}_1^{(n)}A_J^\top)^{-1}\left(A_J\widehat{V}_1^{(n)}A_J^\top - A_JV_1A_J^\top\right)(A_JV_1A_J^\top)^{-1}\right)A_J \\ &= -\widehat{\Pi}_J^{(n)}\left(\widehat{V}_1^{(n)} - V_1\right)\Pi_J,\end{aligned}$$

Table 5. Average of NV for our method (OM) with $K = 5$, and ASCCA, AIC, BIC, CP, FPE, PSS and GIC methods over 1000 replications.

$n = 100, \rho = 0.5$								
τ	OM	ASCCA	AIC	BIC	C_p	FPE	PSS	GIC
0.1	3.05	3.50	4.37	4.07	5.28	3.37	3.38	3.87
0.2	3.00	3.33	4.31	4.02	4.96	3.47	3.34	3.77
0.3	3.07	3.50	4.28	4.07	5.01	3.42	3.32	3.85
0.4	3.00	3.52	3.27	4.02	4.96	3.43	3.27	3.84
0.5	3.10	3.42	4.43	4.10	4.63	3.44	3.32	3.85
$n = 200, \tau = 0.1$								
ρ	OM	ASCCA	AIC	BIC	C_p	FPE	PSS	GIC
0.1	3.03	3.25	3.20	2.28	5.69	3.37	3.35	1.99
0.2	3.00	3.15	4.15	3.47	5.72	3.31	3.20	2.29
0.3	3.02	3.37	4.32	3.96	5.69	3.36	3.26	3.26
0.4	3.00	3.26	4.26	4.06	5.46	3.41	3.32	3.96
0.5	3.03	3.27	4.37	4.01	5.58	3.41	3.36	3.98

it follows that

$$\sqrt{n}\widehat{\xi}_J^{(n)} = \|\sqrt{n}(\widehat{V}_{12}^{(n)} - V_{12}) - \sqrt{n}(\widehat{V}_1^{(n)} - V_1)\widehat{\Pi}_J^{(n)}\widehat{V}_{12}^{(n)} + V_1\widehat{\Pi}_J^{(n)}\left(\sqrt{n}\left(\widehat{V}_1^{(n)} - V_1\right)\right)\Pi_J\widehat{V}_{12}^{(n)} - V_1\Pi_J\left(\sqrt{n}(\widehat{V}_{12}^{(n)} - V_{12})\right) + \sqrt{n}\delta_J\|. \quad (12)$$

Let us consider the $\mathbb{R}^{p+q}$-valued random vectors

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix}, \quad Z^{(k)} = \begin{pmatrix} X^{(k)} \\ Y^{(k)} \end{pmatrix}, \quad k = 1, \cdots, n.$$

The covariance matrix of Z is given by

$$V = \mathbb{E}((Z - \eta)(Z - \eta)^\top) = \begin{pmatrix} V_1 & V_{12} \\ V_{21} & V_2 \end{pmatrix}, \quad (13)$$

where $\eta = \mathbb{E}(Z)$, $V_2 = \mathbb{E}((Y - \nu)(Y - \nu)^\top)$ and $V_{21} = V_{12}^\top$. Further, putting

$$\overline{Z}^{(n)} = n^{-1}\sum_{k=1}^{n} Z^{(k)} \quad \text{and} \quad \widehat{V}^{(n)} = n^{-1}\sum_{k=1}^{n}(Z^{(k)} - \overline{Z}^{(n)})(Z^{(k)} - \overline{Z}^{(n)})^\top,$$

we can write

$$\widehat{V}^{(n)} = \begin{pmatrix} \widehat{V}_1^{(n)} & \widehat{V}_{12}^{(n)} \\ \widehat{V}_{21}^{(n)} & \widehat{V}_2^{(n)} \end{pmatrix}, \quad (14)$$

where $\widehat{V}_2^{(n)} = n^{-1}\sum_{k=1}^n (Y^{(k)} - \overline{Y}^{(n)})\,(Y^{(k)} - \overline{Y}^{(n)})^\top$ and $\widehat{V}_{21}^{(n)} = (\widehat{V}_{12}^{(n)})^\top$. Then we deduce from (12), (13) and (14) that $\sqrt{n}\widehat{\xi}_J^{(n)} = \|\widehat{\Psi}_J^{(n)}(\widehat{H}^{(n)}) + \sqrt{n}\delta_J\|$, where $\widehat{H}^{(n)} = \sqrt{n}(\widehat{V}^{(n)} - V)$ and $\widehat{\Psi}_J^{(n)}$ is defined by

$$\forall A \in \mathcal{L}(\mathbb{R}^{p+q}),\ \ \widehat{\Psi}_J^{(n)}(A) = P_2(A) - P_1(A)\widehat{\Pi}_J^{(n)}\widehat{V}_{12}^{(n)} + V_1\widehat{\Pi}_J^{(n)}P_1(A)\Pi_A\widehat{V}_{12}^{(n)} - V_1\Pi_J P_2(A).$$

Considering the usual matrices norm $\|\cdot\|_\infty$ defined in $\mathscr{M}_{p,q}(\mathbb{R})$ by $\|A\|_\infty = \sup_{x\in\mathbb{R}^q-\{0\}} \|Ax\|_{\mathbb{R}^p}/\|x\|_{\mathbb{R}^q}$ and recalling that, for two matrices A and B, one has $\|AB\|_\infty \leq \|A\|_\infty\|B\|_\infty$, we obtain

$$\begin{aligned}
\|\widehat{\Psi}_J^{(n)}(A) - \Psi_J(A)\|_\infty = \Big\| & -P_1(A)\left(\widehat{\Pi}_J^{(n)} - \Pi_J\right)\widehat{V}_{12}^{(n)} - P_1(A)\Pi_J\left(\widehat{V}_{12}^{(n)} - V_{12}\right) \\
& + V_1\left(\widehat{\Pi}_J^{(n)} - \Pi_J\right)P_1(A)\Pi_J\widehat{V}_{12}^{(n)} + V_1\Pi_J P_1(A)\Pi_J\left(\widehat{V}_{12}^{(n)} - V_{12}\right)\Big\|_\infty \\
\leq\ & \|P_1(A)\|_\infty\Big[\|\widehat{\Pi}_J^{(n)} - \Pi_J\|_\infty\|\widehat{V}_{12}^{(n)}\|_\infty + \|\Pi_J\|_\infty\|\widehat{V}_{12}^{(n)} - V_{12}\|_\infty \\
& + \|V_1\|_\infty\|\Pi_J\|_\infty\|\widehat{\Pi}_J^{(n)} - \Pi_J\|_\infty\|\widehat{V}_{12}^{(n)}\|_\infty + \|V_1\|_\infty\|\Pi_J\|_\infty^2\|\widehat{V}_{12}^{(n)} - V_{12}\|_\infty\Big] \\
\leq\ & \Big[\|\widehat{\Pi}_J^{(n)} - \Pi_J\|_\infty\|\widehat{V}_{12}^{(n)}\|_\infty + \|\Pi_J\|_\infty\|\widehat{V}_{12}^{(n)} - V_{12}\|_\infty \\
& + \|V_1\|_\infty\|\Pi_J\|_\infty\|\widehat{\Pi}_J^{(n)} - \Pi_J\|_\infty\|\widehat{V}_{12}^{(n)}\|_\infty \\
& + \|V_1\|_\infty\|\Pi_J\|_\infty^2\|\widehat{V}_{12}^{(n)} - V_{12}\|_\infty\Big]\,\|P_1\|_{\infty,\infty}\|A\|_\infty,
\end{aligned}$$

where $\|T\|_{\infty,\infty} := \sup_{A\in\mathscr{M}_{p,q}(\mathbb{R})-\{0\}} \|T(A)\|_\infty/\|A\|_\infty$. Hence

$$\begin{aligned}
\|\widehat{\Psi}_J^{(n)} - \Psi_J\|_{\infty,\infty} \leq\ & [\|1 + \|V_1\|_\infty\|\Pi_J\|_\infty\|]\,\|\widehat{V}_{12}^{(n)}\|_\infty\|\widehat{\Pi}_J^{(n)} - \Pi_J\|_\infty\|P_1\|_{\infty,\infty} \\
& + [1 + \|V_1\|_\infty\|\Pi_J\|_\infty]\,\|\Pi_J\|_\infty\|\widehat{V}_{12}^{(n)} - V_{12}\|_\infty\|P_1\|_{\infty,\infty}. \quad (15)
\end{aligned}$$

From the strong law of large numbers it is easily seen that $\widehat{V}_1^{(n)}$ (resp. $\widehat{V}_{12}^{(n)}$ converges almost surely, as $n \to +\infty$ to V_1 (resp. V_{12}). Therefore, $\widehat{\Pi}_J^{(n)}$ converges almost surely, as $n \to +\infty$ to Π_J, and from (15) we deduce that $\widehat{\Psi}_J^{(n)}$ converges almost surely, as $n \to +\infty$ to Ψ_J. It remains to obtain the asymptotic distribution of $\widehat{H}^{(n)}$. We have $\widehat{H}^{(n)} = \widehat{H}_1^{(n)} - \widehat{H}_2^{(n)}$ where

$$\widehat{H}_1^{(n)} = \sqrt{n}\left(\frac{1}{n}\sum_{k=1}^n Z_k\,Z_k^\top - V\right) \quad \text{and} \quad \widehat{H}_2^{(n)} = \frac{1}{\sqrt{n}}\left((\sqrt{n}\,\overline{Z}^{(n)})\,(\sqrt{n}\,\overline{Z}^{(n)})^\top\right).$$

The central limit theorem ensures that $\widehat{H}_1^{(n)}$ (resp. $\sqrt{n}\,\overline{Z}^{(n)}$) converges in distribution, as $n \to +\infty$, to a random variable H (resp. U) having a centred normal distribution. Hence, $\widehat{H}_2^{(n)}$ converges in probability, as $n \to +\infty$, to 0 and Slutsky theorem permits to conclude that $\widehat{H}^{(n)}$ converges in distribution, as $n \to +\infty$, to H.

5.3 Proof of Theorem 1

We just need to prove the lemma which is given below. Then the proof of Theorem 1 is similar to that of Theorem 3.1 in Nkiet (2012). Let $r \in \mathbb{N}^*$ and $(m_1, \cdots, m_r) \in (\mathbb{N}^*)^r$ such that $\sum_{\ell=1}^r m_\ell = p$

and

$$\xi_{K_{\nu_1}} = \cdots = \xi_{K_{\nu_{m_1}}} > \xi_{K_{\nu_{m_1+1}}} = \cdots = \xi_{K_{\nu_{m_1+m_2}}} > \cdots > \xi_{K_{\nu_{m_1+m_2+\cdots+m_{r-1}+1}}} = \cdots = \xi_{K_{\nu_{m_1+m_2+\cdots+m_r}}}.$$

Then, putting

$$E = \{\ell \in \mathbb{N}^* \,/\, 1 \le \ell \le r,\ m_\ell \ge 2\}$$

and

$$F_\ell := \left\{\left(\sum_{k=0}^{\ell-1} m_k\right) + 1, \cdots, \left(\sum_{k=0}^{\ell} m_k\right) - 1\right\},$$

with $m_0 = 0$, we have:

Lemma 3. *If $E \neq \emptyset$, then for all $\ell \in E$ and all $i \in F_\ell$, the sequence $n^\alpha(\widehat{\xi}^{(n)}_{K_{\nu_i}} - \widehat{\xi}^{(n)}_{K_{\nu_{i+1}}})$ converges in probability to 0 as $n \to +\infty$.*

Proof. Let us put $\gamma_\ell = \xi_{K_{\nu_i}} = \xi_{K_{\nu_{i+1}}}$; if $\gamma_\ell = 0$, then

$$\begin{aligned}\left|n^\alpha\left(\widehat{\xi}^{(n)}_{K_{\nu_i}} - \widehat{\xi}^{(n)}_{K_{\nu_{i+1}}}\right)\right| &= n^{\alpha-\frac{1}{2}}\left|\|\widehat{\xi}^{(n)}_{K_{\nu_i}}(\widehat{H}^{(n)})\| - \|\widehat{\xi}^{(n)}_{K_{\nu_{i+1}}}(\widehat{H}^{(n)})\|\right| \\ &\le n^{\alpha-\frac{1}{2}}\|\left(\widehat{\Psi}^{(n)}_{K_{\nu_i}} - \widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}\right)\left(\widehat{H}^{(n)}\right)\| \\ &\le n^{\alpha-\frac{1}{2}}\|\widehat{\Psi}^{(n)}_{K_{\nu_i}} - \widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}\|_\infty\|\widehat{H}^{(n)}\|.\end{aligned}$$

Since $\widehat{\Psi}^{(n)}_{K_{\nu_i}}$ and $\widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}$ converge almost surely, as $n \to +\infty$, to $\Psi_{K_{\nu_i}}$ and $\Psi_{K_{\nu_{i+1}}}$ respectively, and since $\widehat{H}^{(n)}$ converges in distribution, as $n \to +\infty$, to H, it follows from the preceding inequality and from $\alpha < 1/2$ that $n^\alpha(\widehat{\xi}^{(n)}_{K_{\nu_i}} - \widehat{\xi}^{(n)}_{K_{\nu_{i+1}}})$ converges in probability to 0 as $n \to +\infty$. If $\gamma_\ell \neq 0$, we have

$$\begin{aligned}&n^\alpha\left(\widehat{\xi}^{(n)}_{K_{\nu_i}} - \widehat{\xi}^{(n)}_{K_{\nu_{i+1}}}\right) \\ &= n^{\alpha-\frac{1}{2}}\left(\|\widehat{\Psi}^{(n)}_{K_{\nu_i}}(\widehat{H}^{(n)}) + \sqrt{n}\delta_{K_{\nu_i}}\| - \|\widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}(\widehat{H}^{(n)}) + \sqrt{n}\delta_{K_{\nu_{i+1}}}\|\right) \\ &= \frac{n^{\alpha-\frac{1}{2}}\left(\|\widehat{\Psi}^{(n)}_{K_{\nu_i}}(\widehat{H}^{(n)})\|^2 - \|\widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}(\widehat{H}^{(n)})\|^2\right)}{\|\widehat{\Psi}^{(n)}_{K_{\nu_i}}(\widehat{H}^{(n)}) + \sqrt{n}\delta_{K_{\nu_i}}\| + \|\widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}(\widehat{H}^{(n)}) + \sqrt{n}\delta_{K_{\nu_{i+1}}}\|} \\ &\quad + \frac{2n^\alpha\left(\left\langle\delta_{K_{\nu_i}}, \widehat{\Psi}^{(n)}_{K_{\nu_i}}(\widehat{H}^{(n)})\right\rangle - \left\langle\delta_{K_{\nu_{i+1}}}, \widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}(\widehat{H}^{(n)})\right\rangle\right)}{\|\widehat{\Psi}^{(n)}_{K_{\nu_i}}(\widehat{H}^{(n)}) + \sqrt{n}\delta_{K_{\nu_i}}\| + \|\widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}(\widehat{H}^{(n)}) + \sqrt{n}\delta_{K_{\nu_{i+1}}}\|} \\ &= \frac{n^{\alpha-1}\left(\|\widehat{\Psi}^{(n)}_{K_{\nu_i}}(\widehat{H}^{(n)})\|^2 - \|\widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}(\widehat{H}^{(n)})\|^2\right)}{\|n^{-\frac{1}{2}}\widehat{\Psi}^{(n)}_{K_{\nu_i}}(\widehat{H}^{(n)}) + \delta_{K_{\nu_i}}\| + \|n^{-\frac{1}{2}}\widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}(\widehat{H}^{(n)}) + \delta_{K_{\nu_{i+1}}}\|} \\ &\quad + \frac{2n^{\alpha-\frac{1}{2}}\left(\left\langle\delta_{K_{\nu_i}}, \widehat{\Psi}^{(n)}_{K_{\nu_i}}(\widehat{H}^{(n)})\right\rangle - \left\langle\delta_{K_{\nu_{i+1}}}, \widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}(\widehat{H}^{(n)})\right\rangle\right)}{\|n^{-\frac{1}{2}}\widehat{\Psi}^{(n)}_{K_{\nu_i}}(\widehat{H}^{(n)}) + \delta_{K_{\nu_i}}\| + \|n^{-\frac{1}{2}}\widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}(\widehat{H}^{(n)}) + \delta_{K_{\nu_{i+1}}}\|},\end{aligned}$$

where $< \cdot, \cdot >$ is given by $< A, B >= \mathrm{tr}(AB^\top)$. First,

$$\left|n^{\alpha-1}\left(\|\widehat{\Psi}^{(n)}_{K_{\nu_i}}(\widehat{H}^{(n)})\|^2 - \|\widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}(\widehat{H}^{(n)})\|^2\right)\right| \le n^{\alpha-1}\left(\|\widehat{\Psi}^{(n)}_{K_{\nu_i}}\|^2_\infty + \|\widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}\|^2_\infty\right)\|\widehat{H}^{(n)}\|^2 \qquad (16)$$

and, further,

$$\begin{aligned}
&\left|2n^{\alpha-\frac{1}{2}}\left(\left\langle \delta_{K_{\nu_i}}, \widehat{\Psi}^{(n)}_{K_{\nu_i}}(\widehat{H}^{(n)})\right\rangle - \left\langle \delta_{K_{\nu_{i+1}}}, \widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}(\widehat{H}^{(n)})\right\rangle\right)\right| \\
&\leq 2n^{\alpha-\frac{1}{2}}\left(\|\delta_{K_{\nu_i}}\|\|\widehat{\Psi}^{(n)}_{K_{\nu_i}}(\widehat{H}^{(n)})\| + \|\delta_{K_{\nu_{i+1}}}\|\|\widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}(\widehat{H}^{(n)})\|\right) \qquad (17)\\
&\leq 2n^{\alpha-\frac{1}{2}}\gamma_\ell\left(\|\widehat{\Psi}^{(n)}_{K_{\nu_i}}\|_\infty + \|\widehat{\Psi}^{(n)}_{K_{\nu_{i+1}}}\|_\infty\right)\|\widehat{H}^{(n)}\|.
\end{aligned}$$

Equations (16) and (17), and the above recalled convergence properties permit to conclude that the sequence $n^\alpha(\widehat{\xi}^{(n)}_{K_{\nu_i}} - \widehat{\xi}^{(n)}_{K_{\nu_{i+1}}})$ converges in probability to 0, as $n \to +\infty$.

References

An, B., Guo, J., and Wang, H. (2013). Multivariate regression shrinkage and selection by canonical correlation analysis. *Computational Statistics & Data Analysis*, **62**, 93–107.

Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression. The X-random case. *International Statistical Review*, **60**, 291–319.

Fujikoshi, Y., Kan, T., Takahashi, S., and Sakurai, T. (2011). Prediction error criterion for selecting variables in a linear regression model. *Annals of the Institute of Statistical Mathematics*, **63**, 387–403.

Fujikoshi, Y. and Sato, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika*, **84**, 707–716.

Giraud, C. (2011). Low rank multivariate regression. *Electronic Journal of Statistics*, **5**, 775–799.

Hocking, R. R. (1976). The analysis and selection in linear regression. *Biometrics*, **32**, 1–49.

Kundu, D. and Murali, G. (1976). Model selection in linear regression. *Computational Statistics and Data Analysis*, **22**, 461–469.

Linhart, H. and Zucchini, W. (1986). *Model Selection*. Wiley, New York, NY.

Lu, Z., Monteiro, R. D. C., and Yuan, M. (2012). Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Mathematical Programming*, **131**, 163–194.

Miller, A. J. (1990). *Subset Selection in Regression*. Chapman and Hall, London.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, **12**, 758–765.

Nkiet, G. M. (2001). Sélection des variables dans un modèle structurel de régression linéaire. *Comptes Rendus de l'Académie des Sciences de Paris. Série I*, **333**, 1105–1110.

Nkiet, G. M. (2012). Direct variable selection for discrimination among several groups. *Journal of Multivariate Analysis*, **105**, 151–163.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, **88**, 488–494.

Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika*, **71**, 43–49.

Thomson, M. L. (1978a). Selection of variables in multiple regression. Part I. A review and evaluation. *International Statistical Review*, **46**, 1–19.

Thomson, M. L. (1978b). Selection of variables in multiple regression. Part II. Chosen procedures, computations and examples. *International Statistical Review*, **46**, 129–145.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. D. C. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Association*, **69**, 329–346.

Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, **21**, 299–313.

Zheng, X. and Loh, X. Y. (1997). A consistent variable selection criterion for linear models with high-dimensional covariates. *Statistica Sinica*, **7**, 311–325.

Manuscript received 2022-03-18, revised 2022-11-04, accepted 2023-01-29.