Model-Based Variable Decorrelation in Linear Regression

Clément Théry[1] & Christophe Biernacki[2] & Gaétan Loridant[3]

[1] *ArcelorMittal, Université Lille 1, CNRS, Inria, clement.thery@arcelormittal.com*
[2] *Université Lille 1, CNRS, Inria, christophe.biernacki@math.univ-lille1.fr*
[3] *Etudes Industrielles ArcelorMittal Dunkerque, gaetan.loridant@arcelormittal.com*

Abstract. Linear regression outcomes (estimates, prevision) are known to be damaged by highly correlated covariates. However most modern datasets are expected to mechanically convey more and more highly correlated covariates due to the global increase of the amount of variables they contain. We propose to explicitly model such correlations by a family of linear regressions between the covariates. The structure of correlations is found with an MCMC algorithm aiming at optimizing a specific BIC criterion. This hierarchical-like approach leads to a joint probability distribution on both the initial response variable and the linearly explained covariates. Then, marginalisation on the linearly explained covariates produces a parsimonious correlation-free regression model from which classical procedures for estimating regression coefficient, including any variable selection procedures, can be plugged. Both simulated and real-life datasets from steel industry, where correlated variables are frequent, highlight that this proposed covariates pretreatment-like method has two essential benefits: First, it offers a real readability of the linear links between covariates; Second, it improves significantly efficiency of classical estimation/selection methods which are performed after. An R package (CorReg), available on the CRAN, implements this new method.

Keywords. Regression, correlations, steel industry, variable selection, generative models, model selection

1 Introduction

Linear regression is a very standard and efficient method providing a predictive model with a good interpretability even for non-statisticians. Therefore, it is used in nearly all the fields where statistics are made [20]: Astronomy [11], Sociology [15], Industry (real datasets of the present paper), ... With the rise of informatics, datasets contain more and more covariates leading high variance estimates, misleading interpretations and poor prediction accuracy for two different, quite related, reasons. First, the number of covariates leads to more complex models. Second, the number of covariates mechanically increases the chance to have correlated ones. Both situations are quite intricate and appear technically in a similar manner in the estimate procedure through ill-conditioned matrices. As a consequence, explicit break-up between them is not very common, most proposed methods focusing on the seminal question of the number of covariates selection. Originality of this paper is to distinguish explicitly them by proposing a specific decorrelation step followed by any classical variable selection step chosen by the practitioner. The decorrelation step can be viewed as a variable selection step but it is focused only on correlations, not on the number of covariates.

Reducing the variance induced by the large number of covariates can be reached by targeting a better bias-variance trade off. Relating methods are numerous and continue to generate a lot of work. Most traditional directions are shrinkage, including variable selection, and also variable clustering.

Ridge regression [16] is a shrinkage method which proposes possibly biased estimator that can be written in terms of a parametric L_2 penalty. It does not select variables since coefficients tend to 0 but don't reach 0, leading to difficult interpretations for a large number of covariates. Since real datasets may imply many irrelevant variables, variable selection should be preferred for more interpretable models. Variable selection methods may add also some bias by deleting some relevant covariates but it may reduced drastically the variance by the dimension reduction. As an emblematic method, the Least Absolute Shrinkage and Selection Operator (LASSO [24]) consists in a shrinkage of the regression coefficients based on a parametric L_1 penalty to shrink some coefficients exactly to zero. But, like the ridge regression, the penalty does not distinguish correlated and independent covariates so there

is no guarantee to have less correlated covariates. It only produces a parsimonious model, that is a gain for interpretation but only half the way from our point of view. In particular, LASSO is also known to face consistency problems [29] when confronted with correlated covariates. So the quality of interpretation is compromised. Elastic net [30] is a method developed to reach a compromise between Ridge regression and the LASSO by a linear combination of L_1 and L_2 penalties. But, since it is based on the grouping effect, correlated covariates get similar coefficients and are selected together.

Another way for improving the conditioning and the understandability is to consider clusters of variables with the same coefficients, like the Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR [2]) to reduce dimension and also correlations if correlated covariates are in the same clusters. A possible bias is added by the dimension reduction inherent to the coefficients clustering. The CLusterwise Effect REgression (CLERE [26]) describes the regression coefficients no longer as fixed effect parameters but as unobserved independent random variables with grouped coefficients following a Gaussian Mixture distribution. The idea is that if the model has a small number of groups of covariates, the model will have a number of free parameters significantly lower than the number of covariates. In such a case, it improves interpretability and ability to yeld reliable prediction with a smaller variance on the coefficients estimator. Spike and Slab variable selection [10] also relies on a Gaussian mixture (the spike and the slab) hypothesis for the regression coefficients and gives a subset of covariates (not grouped) on which to compute the Ordinary Least Squares estimate (OLS) but has no specific protection against correlations issues.

None of the above methods takes explicitly the correlations into account, even if the clustering methods may group the correlated covariates together. However, modeling explicitly linear correlation between variables already exists in statistics. In Gaussian model-based clustering, [17] consider that some irrelevant covariates for clustering are in linear regression with some relevant ones. The algorithm used to find the structure is a stepwise-like algorithm [21] even if it is known to be often unstable [19]. We propose to transpose this method for linear regression with a specifically adapted algorithm to find the structure of sub-regression.

The idea of the present paper is that if we know explicitly the correlations, we could use this knowledge to avoid this specific problem. Correlations are thus new information to reduce the variance without adding any bias. More precisely, correlations are modeled through a system of linear sub-regressions between covariates. The set of covariates which are *never* at the place of a response variable in these sub-regressions is finally the greatest set of orthogonal covariates. Marginalizing over the dependent co-variables leads then to a linear regression (in relation to the initial response variable) with only orthogonal covariates. This marginalization step can be viewed also as a variable selection step but guided only by the correlations between covariates. Advantages of this approach is twofold. First, it improves interpretation through a good readability of dependency between covariates. Second, this marginal model is still a "true" model provided that both the initial regression model and all the sub-regressions are "true". As a consequence, the associated OLS will preserve an unbiased estimate but with a possibly reduced variance comparing to the OLS with the full regression model. The fact that the variance decreases depends on the residual variances involved in the sub-regressions: The more the sub-regressions are marked, the less will be the variance of associated OLS. In fact, any other estimation method than OLS can be plugged after the marginalization step. Indeed, it can be viewed as a pretreatment against correlation which can be chained after with dimension reduction methods, without no more suffering from correlations this time. The sub-linear structure is obtained by a MCMC algorithm optimizing a specific BIC criterion associated to the joint distribution on the covariates, regardless of the initial response variable. This algorithm is part of the R package CorReg accessible on CRAN.

This paper will first present in Section 2 the modelisation of the correlations between covariates by sub-regressions and the by-product marginal regression model. Section 3 is devoted to describe the MCMC random walk used to find the structure of sub-regressions. Some numerical results on simulated datasets (Section 4) and real industrial datasets (Section 5) are then conducted to quantify the added value of our approach. Concluding remarks, including perspectives, are then given in Section 6.

2 Model-based approach for selecting uncorrelated covariates

2.1 Sub-regressions between covariates

The classical linear regression model can be written

$$\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\beta},\sigma_Y^2 = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_Y \tag{1}$$

with $\boldsymbol{X} = (\boldsymbol{X}^1,\ldots,\boldsymbol{X}^d)$ the $n\times d$ matrix of d predictor variables $\boldsymbol{X}^j$ $(j=1,\ldots,d)$, $\boldsymbol{Y}$ the $n\times 1$ vector of the n response variables and $\boldsymbol{\varepsilon}_Y \sim \mathcal{N}_n(\boldsymbol{0},\sigma_Y^2\boldsymbol{I})$ the centered Gaussian noise of the regression with standard deviation $\sigma_Y > 0$, $\boldsymbol{I}$ denoting the identity matrix of suitable dimension. The $d\times 1$ vector $\boldsymbol{\beta}$ gathers the coefficients of the regression[1], that can be estimated by Ordinary Least Squares (OLS):

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{2}$$

It is an unbiased estimate with variance matrix

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma_Y^2\,(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{3}$$

This estimate requires the inversion of $\boldsymbol{X}'\boldsymbol{X}$ which can lead to great variance estimates, so unstable estimates, if it is ill-conditioned. Ill-conditioning increases when the number d of covariates grows and/or when correlations within the covariates grow (in absolute value) [9]. At the limit, when $d > n$ and/or when some correlations are maximum, $\boldsymbol{X}'\boldsymbol{X}$ becomes singular. Note that d and correlations are also two non unrelated factors since, in real applications, increasing d makes the risk to obtain correlated covariates higher.

We focus now on an original manner to solve the covariates correlation problem. The covariates number problem will be solved at a second stage by standard methods, once only decorrelated covariates will be identified. The proposed method relies on the two following hypotheses.

Hypothesis 1 *In order to take into account the covariates correlation problem, we make the hypothesis correlation between covariates is only the consequence that some covariates linearly depend on some other covariates. More precisely, there are $d_r \geq 0$ such "sub-regressions", each sub-regression $j = 1,\ldots,d_r$ having the covariate $\boldsymbol{X}^{J_r^j}$ as response variable ($J_r^j \in \{1,\ldots,d\}$ and $J_r^j \neq J_r^{j'}$ if $j \neq j'$) and having the $d_p^j > 0$ covariates $\boldsymbol{X}^{J_p^j}$ as predictor variables ($J_p^j \subset \{1,\ldots,d\}\backslash J_r^j$ and $d_p^j = |J_p^j|$ the cardinal of J_p^j):*

$$\boldsymbol{X}^{J_r^j}|\boldsymbol{X}^{J_p^j};\boldsymbol{\alpha}_j,\sigma_j^2 = \boldsymbol{X}^{J_p^j}\boldsymbol{\alpha}_j + \boldsymbol{\varepsilon}_j, \tag{4}$$

where $\boldsymbol{\alpha}_j \in \mathbb{R}^{d_r^j}$ ($\alpha_j^h \neq 0$ for all $j = 1,\ldots,d_r$ and $h = 1,\ldots,d_p^j$) and $\boldsymbol{\varepsilon}_j \sim \mathcal{N}_n(\boldsymbol{0},\sigma_j^2\boldsymbol{I})$.cf

Hypothesis 2 *In addition, we make the complementary hypothesis that the response covariates and the predictor covariates are totally disjoint: for any sub-regression $j = 1,\ldots,d_r$, $J_p^j \subset J_f$ where $J_r = \{J_r^1,\ldots,J_r^{d_r}\}$ is set of all response covariates and $J_f = \{1,\ldots,d\}\backslash J_r$ is the set of all non response covariates of cardinal $d_f = d - d_r = |J_f|$. This new assumption allows to obtain very simple sub-regressions sequences, discarding hierarchical ones, in particular uninteresting cyclic sub-regressions. However it is not too much restrictive since any hierarchical (but non-cyclic) sequence of sub-regressions can be agglomerated into a non-hierarchical sequence of sub-regressions, even if it may implies to partially loose information through variance increase in the new non-hierarchical sub-regressions.*

[1] Usually a constant is included as one of the regressors. For example we can take $\boldsymbol{X}^1 = (1,\ldots,1)'$. The corresponding element of $\boldsymbol{\beta}$ is then the intercept β_1.

Further notations In the following, we will note also $\boldsymbol{J}_r = (J_r^1, \ldots, J_r^{d_r})$ the d_r-uple of all the response variable (to be not confused with the corresponding set J_r previously defined), $\boldsymbol{J}_p = (J_p^1, \ldots, J_p^{d_r})$ the d_r-uple of all the predictors for all the sub-regressions, $\boldsymbol{d}_p = (d_p^1, \ldots, d_p^{d_r})$ the associated number of predictors and $\boldsymbol{S} = (\boldsymbol{J}_r, \boldsymbol{J}_p)$ the global *model* of all the sub-regressions. As more compact notations, we define also $\boldsymbol{X}_r = \boldsymbol{X}^{J_r}$ the whole set of response covariates and also $\boldsymbol{X}_f = \boldsymbol{X}^{J_f}$ the *all* other covariates, denominating now as *free* covariates, including those used as predictor covariates in $\boldsymbol{J}_p$. An illustration of all these notations is displayed through an example in Section 2.3. The parameters are also stacked together: $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{d_r})$ denotes the global coefficient of sub-regressions and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_{d_r}^2)$ denotes the corresponding global variance.

Remarks

- Sub-regressions defined in (4) are very easy to understand by any practitioner and, thus, will give a clear view of all the correlations present in the dataset at hand.
- We have considered correlations between the covariates of the main regression on $\boldsymbol{Y}$, not between the residuals. Thus $\boldsymbol{S}$ does not depend on $\boldsymbol{Y}$ and it can be estimated independently as we will see in Section 3.
- The model of sub-regressions $\boldsymbol{S}$ gives a system of linear regressions that can be viewed as a recursive Simultaneous Equation Model (SEM)[4, 25] or also as a Seemingly Unrelated Regression (SUR) [27].

2.2 Marginal regression with decorrelated covariates

The aim is now to use the model of linear sub-regressions $\boldsymbol{S}$ (that we assume to be known in this part) between some covariates of $\boldsymbol{X}$ to obtain a linear regression on $\boldsymbol{Y}$ relying only on uncorrelated variables $\boldsymbol{X}_f$. The way to proceed is to marginalize the joint distribution of $\{(\boldsymbol{Y}, \boldsymbol{X}_r)|\boldsymbol{X}_f, \boldsymbol{S}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_Y^2, \boldsymbol{\sigma^2}\}$ to obtain the distribution of $\{\boldsymbol{Y}|\boldsymbol{X}_f, \boldsymbol{S}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_Y^2, \boldsymbol{\sigma^2}\}$ depending only on uncorrelated variables $\boldsymbol{X}_f$:

$$\mathbb{P}(\boldsymbol{Y}|\boldsymbol{X}_f, \boldsymbol{S}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_Y^2, \boldsymbol{\sigma}^2) = \int_{\mathbb{R}^{d_r}} \mathbb{P}(\boldsymbol{Y}|\boldsymbol{X}_f, \boldsymbol{X}_r, \boldsymbol{S}; \boldsymbol{\beta}, \sigma_Y^2)\mathbb{P}(\boldsymbol{X}_r|\boldsymbol{X}_f, \boldsymbol{S}; \boldsymbol{\alpha}, \boldsymbol{\sigma}^2)d\boldsymbol{X}_r. \tag{5}$$

We need the following new hypothesis.

Hypothesis 3 *We assume that all errors $\boldsymbol{\varepsilon}_Y$ and $\boldsymbol{\varepsilon}_j$ ($j = 1, \ldots, d_r$) are mutually independent. It implies in particular that conditional response covariates $\{\boldsymbol{X}^{J_r^j}|\boldsymbol{X}^{J_p^j}, \boldsymbol{S}; \boldsymbol{\alpha}_j, \sigma_j^2\}$, with distribution defined in (4), are mutually independent:*

$$\mathbb{P}(\boldsymbol{X}_r|\boldsymbol{X}_f, \boldsymbol{S}; \boldsymbol{\alpha}, \boldsymbol{\sigma}^2) = \prod_{j=1}^{d_r} \mathbb{P}(\boldsymbol{X}^{J_r^j}|\boldsymbol{X}^{J_p^j}, \boldsymbol{S}; \boldsymbol{\alpha}_j, \sigma_j^2). \tag{6}$$

Noting $\boldsymbol{\beta}_r = \boldsymbol{\beta}_{J_r}$ and $\boldsymbol{\beta}_f = \boldsymbol{\beta}_{J_f}$ the regression coefficients associated respectively to the responses and to the free covariates, we can rewrite (1):

$$\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{S}; \boldsymbol{\beta}, \sigma_Y^2 = \boldsymbol{X}_f\boldsymbol{\beta}_f + \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}_Y. \tag{7}$$

Combining now (7) with (4), (5) and (6), and also independence between each $\boldsymbol{\varepsilon}_j$ and $\boldsymbol{\varepsilon}_Y$, we obtain the following closed-form for the distribution of $\{\boldsymbol{Y}|\boldsymbol{X}_f, \boldsymbol{S}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_Y^2, \boldsymbol{\sigma}^2\}$:

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{X}_f, \boldsymbol{S}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_Y^2, \boldsymbol{\sigma}^2 &= \boldsymbol{X}_f(\boldsymbol{\beta}_f + \sum_{j=1}^{d_r} \beta_{J_r^j}\boldsymbol{\alpha}_j^*) + \sum_{j=1}^{d_r} \beta_{J_r^j}\boldsymbol{\varepsilon}_j + \boldsymbol{\varepsilon}_Y & (8) \\ &= \boldsymbol{X}_f\boldsymbol{\beta}_f^* + \boldsymbol{\varepsilon}_Y^*, & (9) \end{aligned}$$

where $\boldsymbol{\alpha}_j^* \in \mathbb{R}^{d_f}$ with $(\boldsymbol{\alpha}_j^*)_{J_p^j} = \boldsymbol{\alpha}_j$ and $(\boldsymbol{\alpha}_j^*)_{J_f \setminus J_p^j} = \boldsymbol{0}$.
Consequently, we have obtained a new regression expression of $\boldsymbol{Y}$ but relying now *only* on uncorrelated

covariates $\boldsymbol{X}_f$. This decorrelation process has also acted like a specific variable selection process because $\boldsymbol{X}_f \subseteq \boldsymbol{X}$. These two statements are expected to decrease the variance of further estimates of $\boldsymbol{\beta}$. However, the counterpart is twofold. First, this regression has a higher residual variance than the initial one since it is now $\sigma_Y^{2*} = \sigma_Y^2 + \sum_{j=1}^{d_r} \beta_{J_r^j}^2 \sigma_j^2$ instead of σ_Y^2. Second, variable selection being equivalent to set $\hat{\boldsymbol{\beta}}_r = \mathbf{0}$, it implies possibly biased estimates of $\boldsymbol{\beta}_r$. As a conclusion, we are faced with a typical *bias-variance trade off*. We will illustrate it in the next section in the case of the OLS estimate.

In practice, the strategy we propose is to rely estimate of $\hat{\boldsymbol{\beta}}$ upon Equation (9). The practitioner can choose any estimate of its choice, like OLS or any variable selection procedure like LASSO. In other words, it is possible to see (9) as a kind of *pretreatment* for decorrelating covariates, while assuming nothing on the subsequent estimate process.

Remark

- As a consequence of Hypothesis 1 to 3 , "free" covariates $\boldsymbol{X}_f$ are *all* decorrelated (see the Lemma in the Appendix about identifiability).

In the following, we will denote by CorReg the new proposed strategy.

2.3 Illustration of the bias-variance trade off with OLS

Qualitative illustration Noting $\boldsymbol{\beta}^*$ the vector of coefficients of the same dimension as $\boldsymbol{\beta}$ with $\boldsymbol{\beta}^*_{J_r} = \boldsymbol{\beta}^*_r = \mathbf{0}$ and $\boldsymbol{\beta}^*_{J_f} = \boldsymbol{\beta}^*_f$, (9) can be rewritten

$$\boldsymbol{Y}|\boldsymbol{X}_f, \boldsymbol{S}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_Y^2, \boldsymbol{\sigma}^2 = \boldsymbol{X}_f \boldsymbol{\beta}^*_f + \boldsymbol{X}_r \boldsymbol{\beta}^*_r + \boldsymbol{\varepsilon}^*_Y. \tag{10}$$

The OLS estimate of $\boldsymbol{\beta}^*$ is then given by

$$\hat{\boldsymbol{\beta}}^*_f = \left(\boldsymbol{X}'_f \boldsymbol{X}_f\right)^{-1} \boldsymbol{X}'_f \boldsymbol{Y} \quad \text{and} \quad \hat{\boldsymbol{\beta}}^*_r = \mathbf{0}. \tag{11}$$

As usual, the OLS estimate $\hat{\boldsymbol{\beta}}^*$ of $\boldsymbol{\beta}^*$ is unbiased but, however, contrary to $\hat{\boldsymbol{\beta}}$, it could be a biased estimate of $\boldsymbol{\beta}$ since

$$\mathbb{E}(\hat{\boldsymbol{\beta}}^*_f) = \boldsymbol{\beta}_f + \sum_{j=1}^{d_r} \beta_{J_r^j} \boldsymbol{\alpha}^*_j \quad \text{and} \quad \mathbb{E}(\hat{\boldsymbol{\beta}}^*_r) = \mathbf{0}. \tag{12}$$

In return, its variance could be reduced compared to this one of $\hat{\boldsymbol{\beta}}$ given in (3) as soon as values of σ_j are small enough (it means strong correlations in sub-regressions) as we can see in the following expression

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}^*_f) = (\sigma_Y^2 + \sum_{j=1}^{d_r} \sigma_j^2 \beta_{J_r^j}^2)(\boldsymbol{X}'_f \boldsymbol{X}_f)^{-1} \quad \text{and} \quad \mathrm{Var}(\hat{\boldsymbol{\beta}}^*_r) = \mathbf{0}. \tag{13}$$

Indeed, no correlations between covariates $\boldsymbol{X}_f$ imply that the matrix $\boldsymbol{X}'_f \boldsymbol{X}_f$ could be sufficiently better conditioned than the matrix $\boldsymbol{X}'\boldsymbol{X}$ involved in (3) to balance the added variance $\sum_{j \in I_r} \sigma_j^2 \beta_j^2$ in (13). This bias-variance trade off can be resumed by the Mean Squared Error (MSE) associated to both estimates:

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}) = \sigma_Y^2 \,\mathrm{Tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}), \tag{14}$$

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}^*) = \| \sum_{j=1}^{d_r} \beta_{J_r^j} \boldsymbol{\alpha}^*_j \|_2^2 + \| \boldsymbol{\beta}_r \|_2^2 + (\sigma_Y^2 + \sum_{j=1}^{d_r} \sigma_j^2 \beta_{J_r^j}^2) \,\mathrm{Tr}((\boldsymbol{X}'_f \boldsymbol{X}_f)^{-1}). \tag{15}$$

Numerical illustration We now illustrate the bias-variance trade off, through a numerical example. Let a simple case with $d = 5$ covariates $\boldsymbol{X} = (\boldsymbol{X}^1, \boldsymbol{X}^2, \boldsymbol{X}^4, \boldsymbol{X}^5)$ independently drawn by four independent Gaussian $\mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I})$. Thus, $J_f = \{1, 2, 4, 5\}$, $d_f = 4$ and $\boldsymbol{X}_f = (\boldsymbol{X}^1, \boldsymbol{X}^2, \boldsymbol{X}^4, \boldsymbol{X}^5)$. Let also $d_r = 1$ response covariate $\boldsymbol{X}^3|\boldsymbol{X}^1, \boldsymbol{X}^2 = \boldsymbol{X}^1 + \boldsymbol{X}^2 + \boldsymbol{\varepsilon}_1$ where $\boldsymbol{\varepsilon}_1 \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_1^2 \boldsymbol{I})$. Thus, $\boldsymbol{\alpha}_1 = (1, 1)'$, $\boldsymbol{J}_r = (3)$, $J_r = \{3\}$, $\boldsymbol{X}_r = (\boldsymbol{X}^3)$, $\boldsymbol{d}_p = (2)$, $\boldsymbol{J}_p = (\{1, 2\})$, $\boldsymbol{X}^{\boldsymbol{J}_p^1} = (\boldsymbol{X}^1, \boldsymbol{X}^2)$ and $\boldsymbol{S} = ((3), (\{1, 2\}))$. Concerning now the regression with $\boldsymbol{Y}$, we define $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{X}^1 + \boldsymbol{X}^2 + \boldsymbol{X}^3 + \boldsymbol{X}^4 + \boldsymbol{X}^5 + \boldsymbol{\varepsilon}_Y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_Y$, where $\boldsymbol{\beta} = (1, 1, 1, 1, 1)'$ and $\sigma_Y \in \{10, 20\}$. Finally, we deduce that $\boldsymbol{Y}|\boldsymbol{X}_f, \boldsymbol{S} = 2\boldsymbol{X}^1 + 2\boldsymbol{X}^2 + \boldsymbol{X}^4 + \boldsymbol{X}^5 + \boldsymbol{\varepsilon}_1 + \boldsymbol{\varepsilon}_Y$ It is clear that $\boldsymbol{X}'\boldsymbol{X}$ will become more ill-conditioned as σ_1 gets smaller.

We compute now the theoretical MSE of the OLS estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^*$ for several values of σ_1 (strength of the sub-regression) and the sample size n. Figure 2.3 displays the MSE evolution with the strength of the sub-regression expressed by a function of the standard coefficient of determination

$$1 - R^2 = \frac{\mathrm{Var}(\boldsymbol{\varepsilon}_1)}{\mathrm{Var}(\boldsymbol{X}^3)} = \frac{\sigma_1^2}{\sigma_1^2 + 2}. \tag{16}$$

The lower is the value of $1 - R^2$, the larger is the strength of the sub-regression.

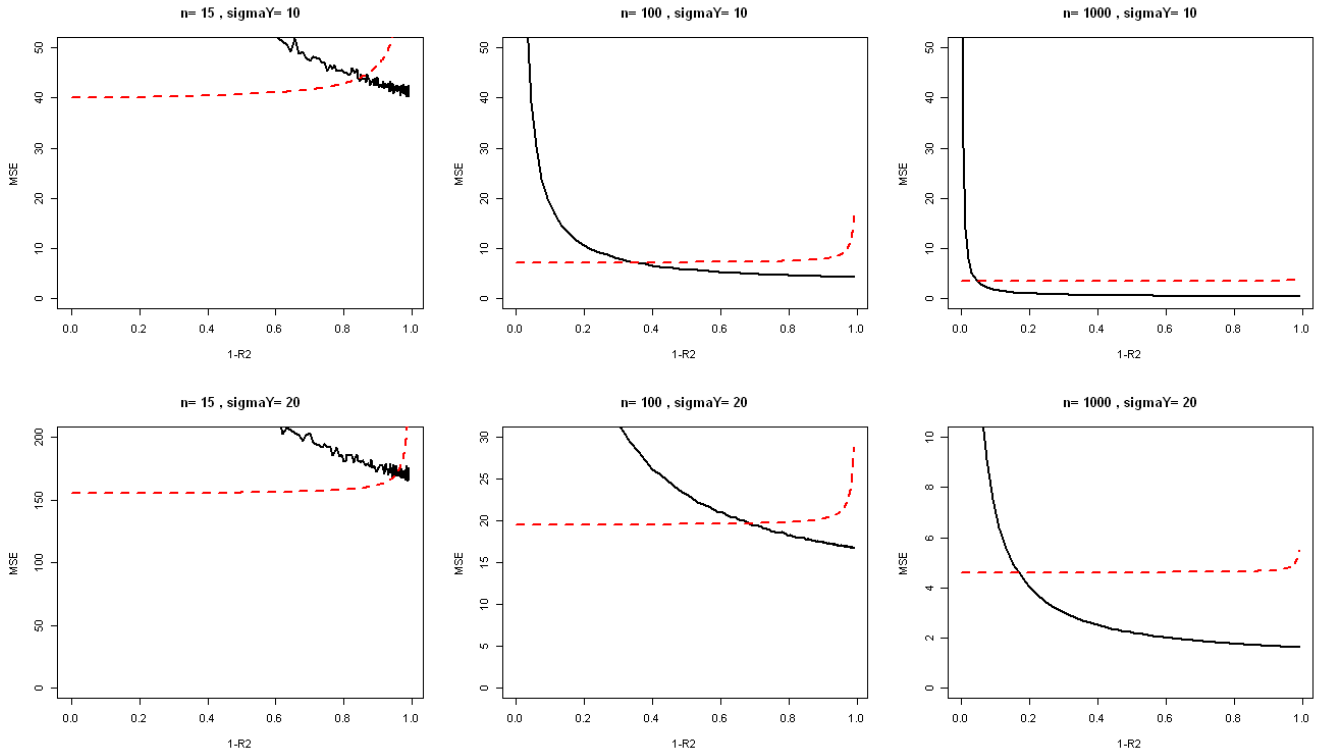


Figure 1: Values of MSE($\hat{\boldsymbol{\beta}}$) (plain) and of MSE($\hat{\boldsymbol{\beta}}^*$) (dotted) when varying the strength ($1 - R^2$) of the sub-regression, and also the values n and σ_Y.

It appears that, when the sub-regression is strong (low $1-R^2$), $\hat{\boldsymbol{\beta}}^*$ is a better estimate than $\hat{\boldsymbol{\beta}}$: the gain on MSE can even be very significant. This effect is amplified by the σ_Y increase but is reduced by the n increases. Thus, the estimate $\hat{\boldsymbol{\beta}}^*$ should be particularly useful when some covariates are highly correlated, when also the sample sizes is small and when the residual variance of $\boldsymbol{Y}$ is large. It corresponds to expected difficult practical situations.

Further results will be provided in Section 4 and 5.

3 Sub-regressions model selection

The question we address now is twofold: which criterion to retain for selecting a sub-regression structure $\boldsymbol{S}$ and which strategy to adopt for exploring the large space of models $\boldsymbol{S}$. Obviously, $\boldsymbol{S}$ being very simply understood even by non statisticians, practitioners could easily transform their expert

knowledge on the phenomenon at hand, if any, into a given structure $\boldsymbol{S}$. For instance, Structural Equations Models (SEM), which are related to our model as already mentioned in Section 1, are often used in social sciences and economy where a structure $\boldsymbol{S}$ is generally "hand-made". However, in the general situation, $\boldsymbol{S}$ has to be estimated. A standard method as graphical LASSO [6], which searches for a structure on the precision matrix (inverse of the variance-covariance matrix) by setting some coefficients of the precision matrix to zero, can not be applied since it is not designed to estimate oriented structures like $\boldsymbol{S}$.

It is important to note that selecting $\boldsymbol{S}$ only relies on $\boldsymbol{X}$, not on $\boldsymbol{Y}$.

3.1 Designing two specific BIC criteria

The Bayesian model selection paradigm consists of retaining the model $\boldsymbol{S}$ maximizing the posterior distribution [22, 1, 3]

$$\begin{aligned} \mathbb{P}(\boldsymbol{S}|\boldsymbol{X}) &\propto \mathbb{P}(\boldsymbol{X}|\boldsymbol{S})\mathbb{P}(\boldsymbol{S}) && (17)\\ &= \mathbb{P}(\boldsymbol{X}_r|\boldsymbol{X}_f,\boldsymbol{S})\mathbb{P}(\boldsymbol{X}_f|\boldsymbol{S})\mathbb{P}(\boldsymbol{S}). && (18)\end{aligned}$$

In order to implement this paradigm, we need first to define the three probabilities which are in the right hand of the previous equation.

Defining $\mathbb{P}(\boldsymbol{X}_r|\boldsymbol{X}_f,\boldsymbol{S})$ corresponds to the integrated likelihood based on $\mathbb{P}(\boldsymbol{X}_r|\boldsymbol{X}_f,\boldsymbol{S};\boldsymbol{\alpha},\boldsymbol{\sigma^2})$. It can be approximated by a BIC-like approach [23]

$$-2\ln\mathbb{P}(\boldsymbol{X}_r|\boldsymbol{X}_f,\boldsymbol{S}) \approx -2\ln\mathbb{P}(\boldsymbol{X}_r|\boldsymbol{X}_f,\boldsymbol{S};\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\sigma}}^2) + (|\boldsymbol{\alpha}|+|\boldsymbol{\sigma}^2|)\ln(n) = \text{BIC}_r(\boldsymbol{S}), \quad (19)$$

where $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\sigma}}^2$ designate respectively the Maximum Likelihood Estimates (MLE) of $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}^2$, and $|\boldsymbol{\psi}|$ designates the number of free continuous parameters associated to the space of any parameter $\boldsymbol{\psi}$.

Defining $\mathbb{P}(\boldsymbol{X}_f|\boldsymbol{S})$ It corresponds to the integrated likelihood based on a not yet defined distribution $\mathbb{P}(\boldsymbol{X}_f|\boldsymbol{S};\boldsymbol{\theta})$ on the uncorrelated covariates $\boldsymbol{X}_f$ and parameterized by $\boldsymbol{\theta}$. In this purpose, we need the following new hypothesis.

Hypothesis 4 *All covariates $\boldsymbol{X}^j$ with $j \in J_f$ are mutually independent and arise from the following Gaussian mixture of k_j components*

$$\mathbb{P}(\boldsymbol{X}_f^j|\boldsymbol{S};\boldsymbol{\pi}_j,\boldsymbol{\mu}_j,\boldsymbol{\lambda}_j^2) = \sum_{h=1}^{k_j}\pi_{hj}\mathcal{N}_n(\mu_{hj}.(1,\ldots,1)',\lambda_{hj}^2\boldsymbol{I}), \quad (20)$$

where $\boldsymbol{\pi}_j = (\pi_{1j},\ldots,\pi_{k_jj})$ is the vector of mixing proportions with $\forall 1 \leq h \leq k_j, \pi_{hj} > 0$ and $\sum_{h=1}^{k_j}\pi_{hj} = 1$, $\boldsymbol{\mu}_j = (\mu_{1j},\ldots,\mu_{k_jj})$ is the vector of centers and $\boldsymbol{\lambda}_j^2 = (\lambda_{1j}^2,\ldots,\lambda_{k_jj}^2)$ is the vector of variances. We stack together all the mixture parameters in $\boldsymbol{\theta} = (\boldsymbol{\pi}_j,\boldsymbol{\mu}_j,\boldsymbol{\lambda}_j^2; j \in J_f)$.

Noting $\hat{\boldsymbol{\theta}}$ the MLE of $\boldsymbol{\theta}$, the BIC approximation can then be used again:

$$-2\ln\mathbb{P}(\boldsymbol{X}_f|\boldsymbol{S}) \approx -2\ln\mathbb{P}(\boldsymbol{X}_f|\boldsymbol{S};\hat{\boldsymbol{\theta}}) + |\boldsymbol{\theta}|\ln(n) = \text{BIC}_f(\boldsymbol{S}). \quad (21)$$

Defining $\mathbb{P}(\boldsymbol{S})$ The most standard choice consists of putting a uniform distribution on the model space $\mathcal{S}$, this choice being noted $\mathbb{P}_U(\boldsymbol{S}) = |\mathcal{S}|^{-1}$, with $|\mathcal{S}|$ the space dimension of $\mathcal{S}$.

However, $\mathcal{S}$ being combinatorial, $|\mathcal{S}|$ is huge. It has two cumulated consequences: First, the *exact* probability $\mathbb{P}(\boldsymbol{S}|\boldsymbol{X})$ may be of the same order of magnitude for a large number of candidates $\boldsymbol{S}$, including the best one; Second, the BIC *approximations* of this quantity may introduce additional confusion to wisely distinguish between model probabilities. In order to limit this problem, we propose

to introduce some information in $\mathbb{P}(\boldsymbol{S})$ promoting simple models through the following *hierarchical* uniform distribution denoted by $\mathbb{P}_H(\boldsymbol{S})$:

$$\begin{aligned}
\mathbb{P}_H(\boldsymbol{S}) &= \mathbb{P}_H(\boldsymbol{J}_r, \boldsymbol{J}_p) && (22)\\
&= \mathbb{P}_H(\boldsymbol{J}_r, \boldsymbol{J}_p, d_r, \boldsymbol{d}_p) && (23)\\
&= \mathbb{P}_U(\boldsymbol{J}_p|\boldsymbol{d}_p, \boldsymbol{J}_r, d_r) \times \mathbb{P}_U(\boldsymbol{d}_p|\boldsymbol{J}_r, d_r) \times \mathbb{P}_U(\boldsymbol{J}_r|d_r) \times \mathbb{P}_U(d_r) && (24)\\
&= \left[\prod_{j=1}^{d_r} \left(\begin{array}{c} d-d_r \\ d_p^j \end{array} \right)\right]^{-1} \times [d-d_r]^{-d_r} \times \left[\left(\begin{array}{c} d \\ d_r \end{array} \right)\right]^{-1} \times [d+1]^{-1}, && (25)
\end{aligned}$$

where $\left(\begin{array}{c} a \\ b \end{array} \right)$ means the number of b-element subsets of an a-element set and where all probabilities $\mathbb{P}_U(\cdot)$ denote uniform distribution on the related space at hand. $\mathbb{P}_H(\boldsymbol{S})$ gives decreasing probabilities to more complex models, provided that the following new hypothesis is verified:

Hypothesis 5 *We set $d_r < d/2$ and also $d_p^j < d/2$ $(j = 1, \ldots, d_r)$.*

These two thresholds are sufficiently large to be wholly realistic.

Final approximation of $\mathbb{P}(\boldsymbol{S}|\boldsymbol{X})$ Merging the previous three expressions, it leads to the following two *global* BIC criteria, to be minimized, denoted by BIC_U or BIC_H, depending on the choice of $\mathbb{P}_U(\boldsymbol{S})$ or $\mathbb{P}_H(\boldsymbol{S})$ respectively:

$$\begin{aligned}
\text{BIC}_U(\boldsymbol{S}) &= \text{BIC}_r(\boldsymbol{S}) + \text{BIC}_f(\boldsymbol{S}) - 2\ln \mathbb{P}_U(\boldsymbol{S}) && (26)\\
\text{BIC}_H(\boldsymbol{S}) &= \text{BIC}_r(\boldsymbol{S}) + \text{BIC}_f(\boldsymbol{S}) - 2\ln \mathbb{P}_H(\boldsymbol{S}). && (27)
\end{aligned}$$

In the following, we will denote by BIC_* any of both BIC_U and BIC_H. Numerical results in Section 4.2 will allow to compare behaviour of both criteria.

Remarks

- Hypothesis 4 is the keystone to define a full generative model on the whole covariates $\boldsymbol{X}$. On the one hand, the BIC criterion can be applied in this context, avoiding to use a cross-validation criterion which can be much more time-consuming. On the other hand, the great flexibility of Gaussian mixture models [18], provided that the number of components k_j has to be estimated, implies that Hypothesis 4 is particularly weak in fact.
- In practice, Gaussian mixture models are estimated only once for each variable $\boldsymbol{X}^j$ $(j = 1, \ldots, d)$. Thus, there is no combinatorial difficulty associated with them. An EM algorithm [5] will be used for estimating the mixture parameters and a classical BIC criterion [23] will be used for selecting the different number of components k_j.
- A sufficient condition for identifiability of the structure of sub-regressions $\boldsymbol{S}$ is that all sub-regressions contain at least two predictor covariates ($d_p^j \geq 2$ for all $j = 1, \ldots, d_r$). In fact, if there exists some sub-regressions with only one regressor, identifiability holds for these sub-regressions only up to a permutation between the related response and predictor covariates. However, even in this case, full identifiability may occur thanks to constraints on response and predictor covariates given in Hypothesis 1 and 2.
- As any BIC criterion, the BIC_U and BIC_H criteria are consistent [13].
- Even if it favors more parsimonious models, $\mathbb{P}_H(\boldsymbol{S})$ can be also viewed as a poor informative prior on $\boldsymbol{S}$ since it is a combination of non informative priors.

3.2 Exploring the structure space with an MCMC algorithm

We present now an MCMC algorithm devoting to minimize the criterion BIC_* to find the optimal estimate of the structure $\boldsymbol{S}$. This Markov chain is regular and ergodic with a finite state space, thus it has a stationary distribution $\pi \propto \exp(-\text{BIC}_*)$ on the space $\mathcal{S}$ of $\boldsymbol{S}$ [8]. Consequently, the chain is expected to be more concentrated around the mode of π, where the optimal value of $\boldsymbol{S}$ stands.

This algorithm alternates two steps: the definition of a neighbourhood $\mathcal{V}(\boldsymbol{S})$ around the current structure $\boldsymbol{S}$ and then the generation of a new structure $\boldsymbol{S}^+$ belonging to this neighbourhood according to its posterior probability.

Note that the $\mathcal{S}$ has to be a *regular* space. It means that it has to verify Hypothesis 1, 2 and 5. In addition, we note below $\boldsymbol{S}$ and $\boldsymbol{S}^+$ the structures at the current and the next iteration of the algorithm, respectively.

3.2.1 Definition of a neighbourhood $\mathcal{V}(\boldsymbol{S})$

We define a *global* neighbourhood space $\mathcal{V}(\boldsymbol{S})$ of $\boldsymbol{S}$ composed by the following four *specific* neighbourhood spaces $\mathcal{V}(\boldsymbol{S}) = \mathcal{V}_{r+}(\boldsymbol{S}) \cup \mathcal{V}_{r-}(\boldsymbol{S}) \cup \mathcal{V}_{p+}(\boldsymbol{S}) \cup \mathcal{V}_{p-}(\boldsymbol{S})$ described below:

- **Adding a sub-regression**: a new sub-regression with only one predictor covariate is added to $\boldsymbol{S}$

$$\mathcal{V}_{r+}(\boldsymbol{S}) = \left\{ \tilde{\boldsymbol{S}} : \tilde{\boldsymbol{S}} \in \mathcal{S}, (\tilde{\boldsymbol{J}}_r, \tilde{\boldsymbol{J}}_p)^{1,\dots,d_r} = (\boldsymbol{J}_r, \boldsymbol{J}_p), \tilde{J}_r^{d_r+1} \in J_f, \tilde{J}_p^{d_r+1} = \{j\}, j \in J_f \right\}. \tag{28}$$

- **Removing a sub-regression**: a sub-regression is removed from $\boldsymbol{S}$

$$\mathcal{V}_{r-}(\boldsymbol{S}) = \left\{ \tilde{\boldsymbol{S}} : \tilde{\boldsymbol{S}} \in \mathcal{S}, (\tilde{\boldsymbol{J}}_r, \tilde{\boldsymbol{J}}_p) = (\boldsymbol{J}_r, \boldsymbol{J}_p)^{\{1,\dots,d_r\}\setminus j}, j \in \{1,\dots,d_r\} \right\}. \tag{29}$$

- **Adding a predictor covariate**: a predictor covariate is added to one sub-regression of $\boldsymbol{S}$

$$\mathcal{V}_{p+}(\boldsymbol{S}) = \left\{ \tilde{\boldsymbol{S}} : \tilde{\boldsymbol{S}} \in \mathcal{S}, \tilde{\boldsymbol{J}}_r = \boldsymbol{J}_r, \tilde{J}_p^{\{1,\dots,d_r\}\setminus\{j\}} = J_p^{\{1,\dots,d_r\}\setminus\{j\}}, \tilde{J}_p^j = J_p^j \cup \{h\}, j \in \{1,\dots,d_r\}, h \in J_f \setminus J_p^j \right\}. \tag{30}$$

- **Removing a predictor covariate**: a predictor covariate is removed from one sub-regression of $\boldsymbol{S}$

$$\mathcal{V}_{p-}(\boldsymbol{S}) = \left\{ \tilde{\boldsymbol{S}} : \tilde{\boldsymbol{S}} \in \mathcal{S}, \tilde{\boldsymbol{J}}_r = \boldsymbol{J}_r, \tilde{J}_p^{\{1,\dots,d_r\}\setminus\{j\}} = J_p^{\{1,\dots,d_r\}\setminus\{j\}}, \tilde{J}_p^j = J_p^j \setminus \{h\}, j \in \{1,\dots,d_r\}, h \in J_p^j \right\}. \tag{31}$$

3.2.2 Generation of a new structure $\boldsymbol{S}^+$

We then generate $\boldsymbol{S}^+$ from the following transition probability defined in $\mathcal{V}(\boldsymbol{S})$

$$\mathbb{P}(\boldsymbol{S}^+ | \mathcal{V}(\boldsymbol{S})) = \frac{\exp\left(-\text{BIC}_*(\cdot)\right)}{\sum_{\tilde{\boldsymbol{S}} \in \mathcal{V}(\boldsymbol{S})} \exp\left(-\text{BIC}_*(\tilde{\boldsymbol{S}})\right)}. \tag{32}$$

3.2.3 Detailed use of the algorithm

- **Estimate $\hat{\boldsymbol{S}}$**: we retain the structure $\hat{\boldsymbol{S}}$ having the lowest value of the criterion $\text{BIC}_*(\tilde{\boldsymbol{S}})$ during the walk.

- **Initialization**: the initial structure is randomly from a distribution taking into account the absolute value of the correlations.

- **Long versus short runs**: we prefer to run multiple short chains than to run a unique long chains [7].

4 Numerical results on simulated datasets

We now aim to access the numerical behaviour of the proposed strategy CorReg, and its related estimation and model selection processes, through some simulated datasets. We will also evaluate robustness of the method in some disadvantageous situations as non linear correlations between covariates.

4.1 Experimental design

We consider regressions on $\boldsymbol{Y}$ with $d = 40$ covariates and with a R^2 value equal to 0.4. Sub-regressions will have R^2 successively set to $(0.1, 0.3, 0.5, 0.7, 0.99)$. Variables in $\boldsymbol{X}_f$ arise from a Gaussian mixture model whose the number of components follows a Poisson's law of mean parameter equal to 5. The coefficients of $\boldsymbol{\beta}$ and of the $\boldsymbol{\alpha}_j$'s are independently generated according to the same Poisson distribution but with a uniform random sign. All sub-regressions are of length two ($\forall j = 1, \ldots, d_r, d_p^j = 2$ and we have $d_r = 16$ sub-regressions. The datasets are then scaled, so that covariates $\boldsymbol{X}_r$ avoid large distortions for variances or for means. Different sample sizes $n \in (30, 50, 100, 400)$ are chosen, thus considering experiments in both situations $n < d$ and $n > d$. In all figures, the thickness of the lines will represent various values of n, the thicker being the greater.

We used Rmixmod [14] to estimate the Gaussian mixture densities of each covariate $\boldsymbol{X}$, settings being used at default values both for parameter and the number of components estimation. For each configuration, the mcmc walk was launched on 10 initial structures with 1 000 iterations each time. All the results are provided by the CorReg package available on the cran[2].

In the following, Section 4.2 evaluates the quality of the procedures to estimate the structure $\hat{\boldsymbol{S}}$. Section 4.3 and 4.4 compare predictive performance of standard methods with/without CorReg in some "standard" and "robustness" cases, respectively.

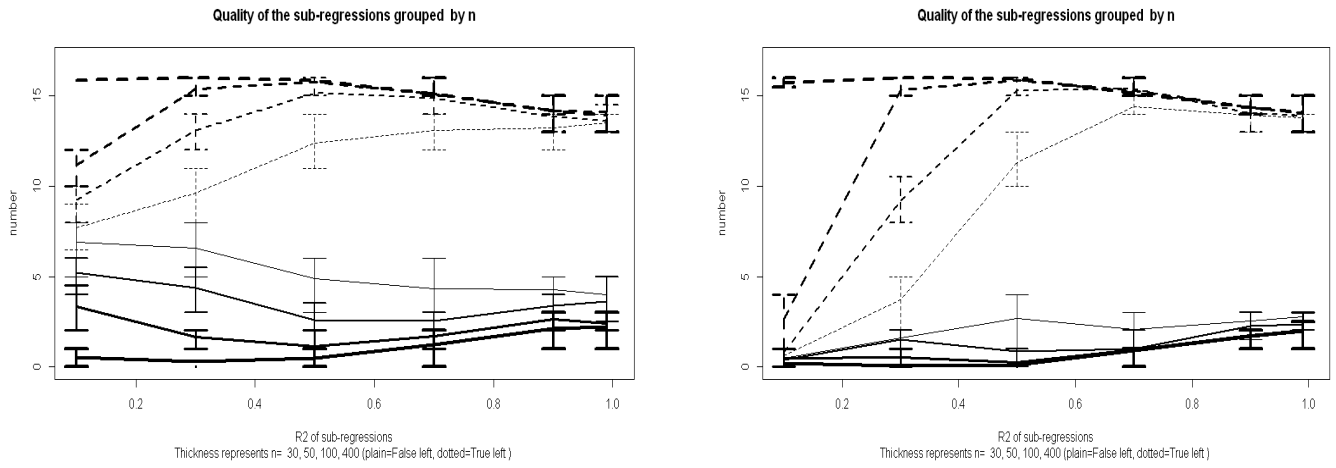
4.2 Evaluation of BIC_* to estimate $\boldsymbol{S}$

To evaluate the quality of the estimated structure $\hat{\boldsymbol{S}}$ structure retained both by BIC_U and BIC_H we define two complementary indicators to compare it to the true model $\boldsymbol{S}$.

- $T_r = |J_r \cap \hat{J}_r|$ ("True Response"): it corresponds to the number of estimate response covariates in $\hat{\boldsymbol{S}}$ which are *truly* response covariates in the true model $\boldsymbol{S}$.
- $W_r = |\hat{J}_r| - T_r$ ("Wrong Response"): it corresponds to the number of estimate response covariates in $\hat{\boldsymbol{S}}$ which are are *wrongfully* response covariates in the true model $\boldsymbol{S}$.

The T_r and W_r quality values are displayed in Figure 2(a) and (b) for the BIC_U and BIC_H respectively. We observe that BIC_H provides notably less wrong sub-regressions than BIC_U for any strength R^2 of the true sub-regressions. In addition, BIC_H need to have significantly strong sub-regressions ($R^2 > 0.4$) to detect them. However, this is not a problem since our CorReg strategy is expected to involve only quite strong correlated covariates to have potential interest. Finally, we keep now the BIC_H criterion as the best one for our purpose.

[2] http://cran.r-project.org/web/packages/CorReg/index.html



(a) BIC_U criterion (b) BIC_H criterion

Figure 2: Average quality of the estimate subregressions $\hat{\boldsymbol{S}}$ (dotted for T_r, plain for W_r) obtained by (a) BIC_U and (b) BIC_H. Thickness represents $n = 30, 50, 100, 400$. Inter-quartile intervals are also displayed.

4.3 Evaluation of the prediction quality

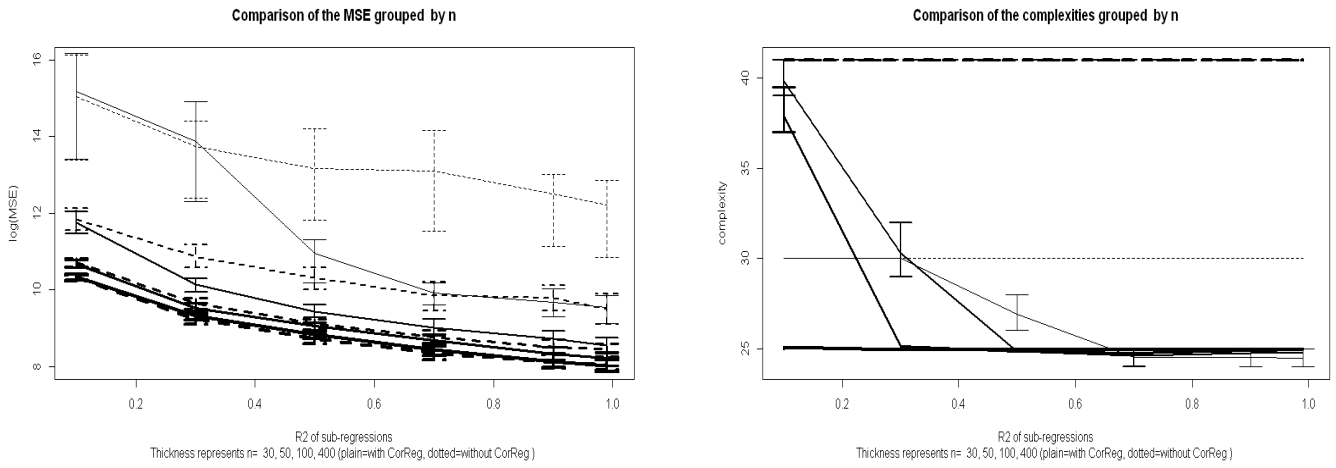
To evaluate the prediction quality of CorReg as a pretreatment method, we consider three *scenarii*: First, the response variable $\boldsymbol{Y}$ depends on all covariates $\boldsymbol{X}$; Second, $\boldsymbol{Y}$ depends on only covariates $\boldsymbol{X}_f$; Finally, $\boldsymbol{Y}$ depends on only covariates $\boldsymbol{X}_r$. It correspond respectively to a *neutral* situation for CorReg, a *favorable* one and an *unfavorable* one. The BIC_H criterion is always used for the model selection.

4.3.1 $\boldsymbol{Y}$ depends on all $\boldsymbol{X}$

We compare different standard estimation methods (OLS, LASSO, ridge, stepwise) with and without CorReg as a pre-treatment. When $n < p$, the OLS method is associated as usual with the Moore-Penrose generalized inverse [12]. In addition, when using penalized estimators for variable selection like LASSO, an OLS step is used for coefficient estimation after shrinkage for better estimation [28]. When CorReg is combined with a standard estimation method, OLS for instance, the whole method will be simply noted CorReg+OLS. Results will be evaluated though a predictive Mean Squared Error value (MSE) on a validation sample of 1 000 individuals and also by the complexity of the regression in $\boldsymbol{Y}$ (*i.e.* its number of variables). Associated figures will display both mean and inter-quartile intervals of this predictive MSE.

Comparison 1 We compare OLS with CorReg+OLS. Figure 3 (a) shows that CorReg improves significantly the prediction power of OLS for small values of n and/or heavy sub-regression structures. This advantage then shrinks when n increases because the matrix to invert becomes better-conditioned and since CorReg does not allow to retrieve that $\boldsymbol{Y}$ depends on all $\boldsymbol{X}$ because of the marginalization of some covariates implicated in the sub-regressions. Figure 3 (b) illustrates also the model regression in $\boldsymbol{Y}$ retained by CorReg is more parsimonious, provided that sub-regressions are strong enough.

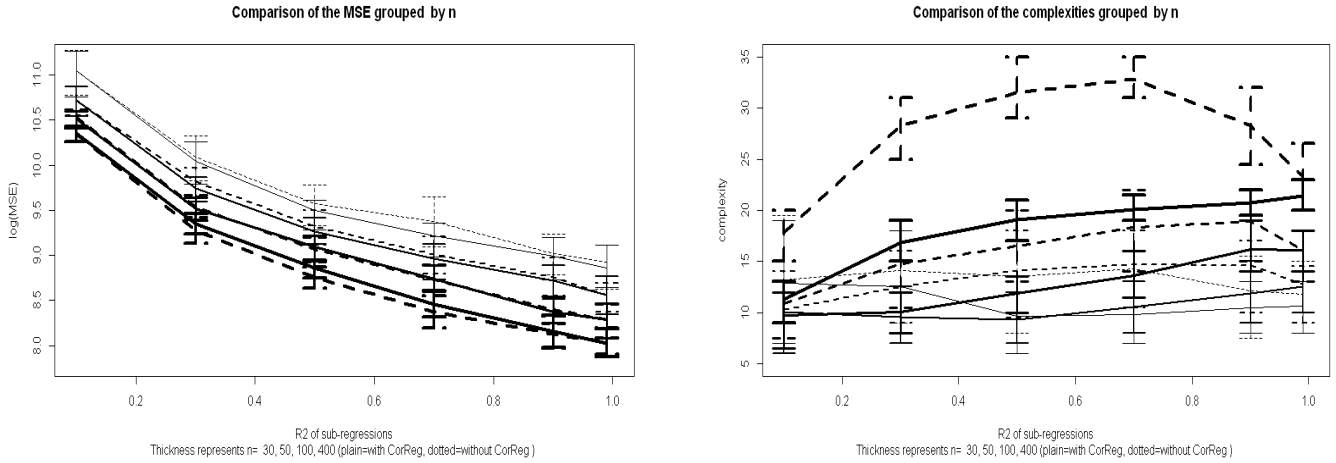
Comparisons 2 to 4 We compare now three variable selections methods: LASSO with CorReg+LASSO, elasticnet with CorReg+electicnet and also stepwise with CorReg+stepwise. Figures 4, 5 and 6 respectively display the results on the same manner as the previous OLS with CorReg+OLS. We see that CorReg, used as a pre-treatment, provides similar prediction accuracy as the three variable selection methods (this prediction is often better for small datasets) but with much more parsimonious regression models on $\boldsymbol{Y}$. This is remarkable because the true model depends on



(a) MSE on a validation subset (b) Complexity

Figure 3: Comparison of OLS (dotted) and CorReg + OLS (plain) when $\boldsymbol{Y}$ depends on all covariates $\boldsymbol{X}$. Thickness represents $n = 30, 50, 100, 400$. Inter-quartile intervals are also displayed.

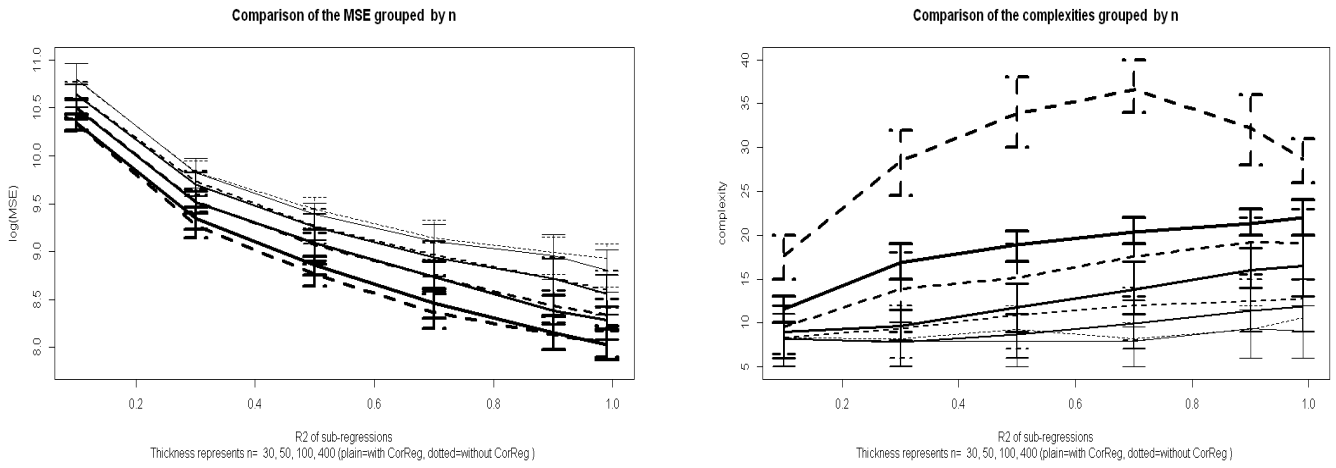
all covariates, so it highlights that other variable selection methods may be really penalized by the correlations between the covariates.



(a) MSE on a validation subset (b) Complexity

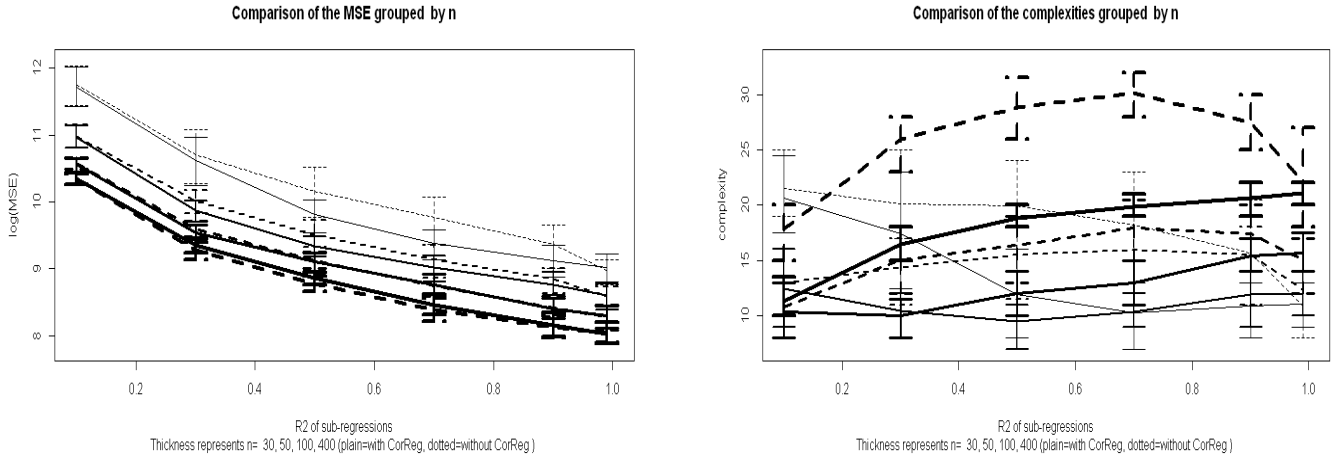
Figure 4: Comparison of LASSO (dotted) and CorReg + LASSO (plain) when $\boldsymbol{Y}$ depends on all covariates $\boldsymbol{X}$. Thickness represents $n = 30, 50, 100, 400$. Inter-quartile intervals are also displayed.

Comparison 5 We compare now ridge with CorReg+ridge. Figure 7 shows that the ridge regression is efficient in prediction when confronted to correlated covariates. It is directly made to improve the conditioning and keeps all the covariates (as the true model) so it logically gives better predictions than CorReg which removes some covariates. Nevertheless, we notice that the CorReg pre-treatment before the ridge regression gives MSE values that are really close to the ridge regression ones (inter-quartile intervals are very mingled) but with drastically more parsimonious models. Indeed, ridge regression alone gives a full model (40 covariates even with only 30 individuals) and will give models too complex to be easily interpreted. Thus, the combination CorReg+ridge provides high benefits for interpretation while preserving good prediction accuracy.



(a) MSE on a validation subset (b) Complexity

Figure 5: Comparison of elasticnet (dotted) and CorReg + elasticnet (plain) when $\boldsymbol{Y}$ depends on all covariates $\boldsymbol{X}$. Thickness represents $n = 30, 50, 100, 400$. Inter-quartile intervals are also displayed.



(a) MSE on a validation subset (b) Complexity

Figure 6: Comparison of stepwise (dotted) and CorReg + stepwise (plain) when $\boldsymbol{Y}$ depends on all covariates $\boldsymbol{X}$. Thickness represents $n = 30, 50, 100, 400$. Inter-quartile intervals are also displayed.

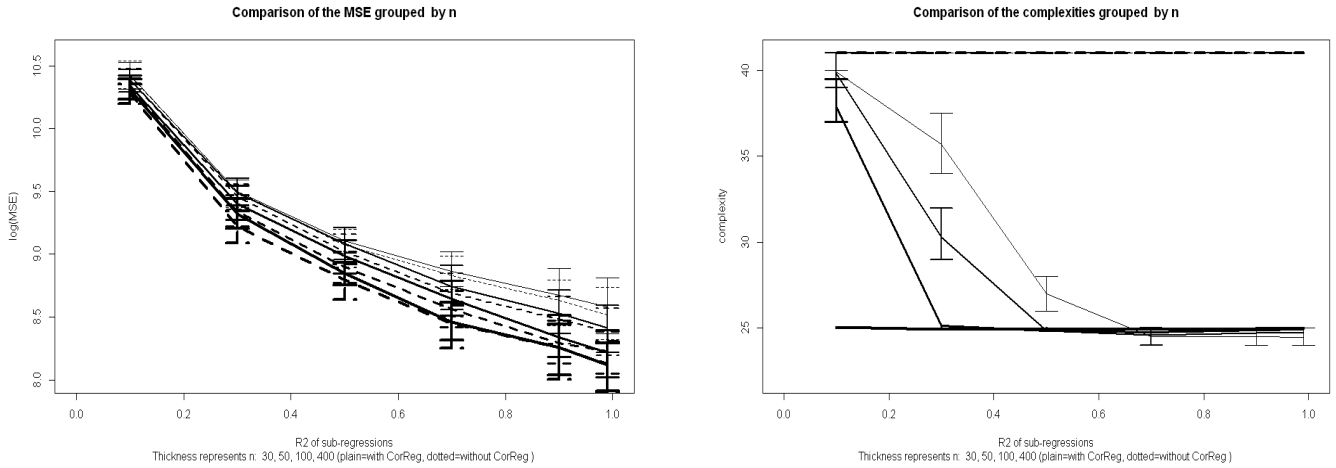
4.3.2 $\boldsymbol{Y}$ depends only on $\boldsymbol{X}_f$

Figures 8 and 9 display results for, respectively, OLS with CorReg+OLS and LASSO with CorReg+LASSO. Even if the true model involved correlated but irrelevant covariates then classical variable selection methods gain to be associated with a pre-treatment by CorReg. Figure 10 displays results for RIDGE with CorReg+RIDGE. It shows the accuracy prediction of ridge regression is heavily penalized since the true model is parsimonious. In that case, CorReg can significantly help to improve the results.

4.3.3 $\boldsymbol{Y}$ depends only on $\boldsymbol{X}_r$

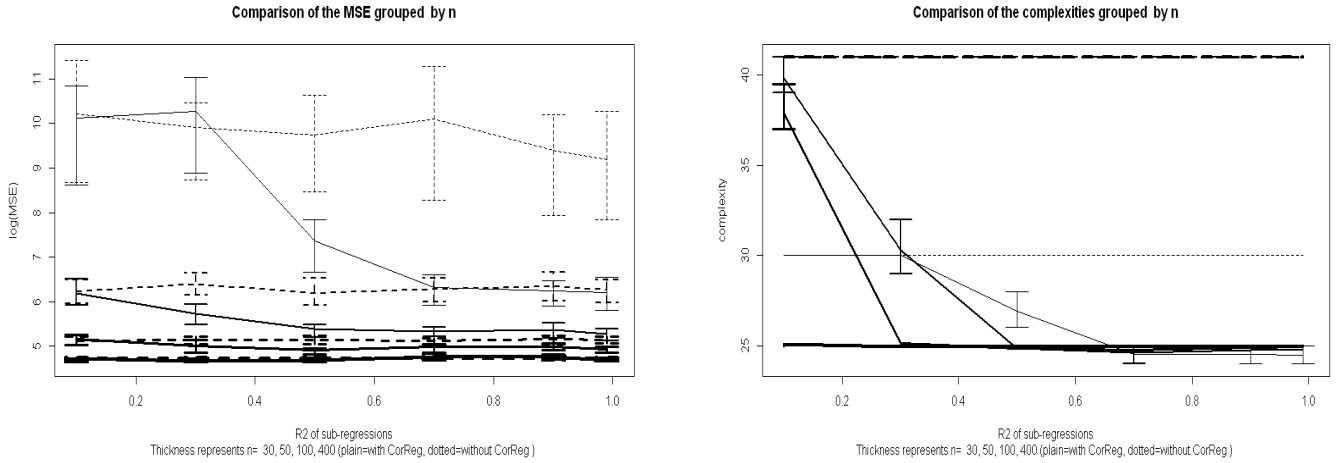
We now try the method with a response depending only on variables in $\boldsymbol{X}_r$. Depending only on $\boldsymbol{X}_r$ implies sparsity and impossibility to obtain the true model, even if using the true structure $\boldsymbol{S}$.

Figure 11 compares OLS with CorReg+OLS. It reveals that CorReg is still better than OLS for strong correlations and limited values of n. When n rises, the sub-regression are detected and relevant covariates are removed. As a consequence, CorReg can not improve the results and increases the MSE. However, for strong correlations, the error shrinks as the model tends to be less identifiable and



(a) MSE on a validation subset (b) Complexity

Figure 7: Comparison of ridge (dotted) and CorReg + ridge (plain) when $\boldsymbol{Y}$ depends on all covariates $\boldsymbol{X}$. Thickness represents $n = 30, 50, 100, 400$. Inter-quartile intervals are also displayed.



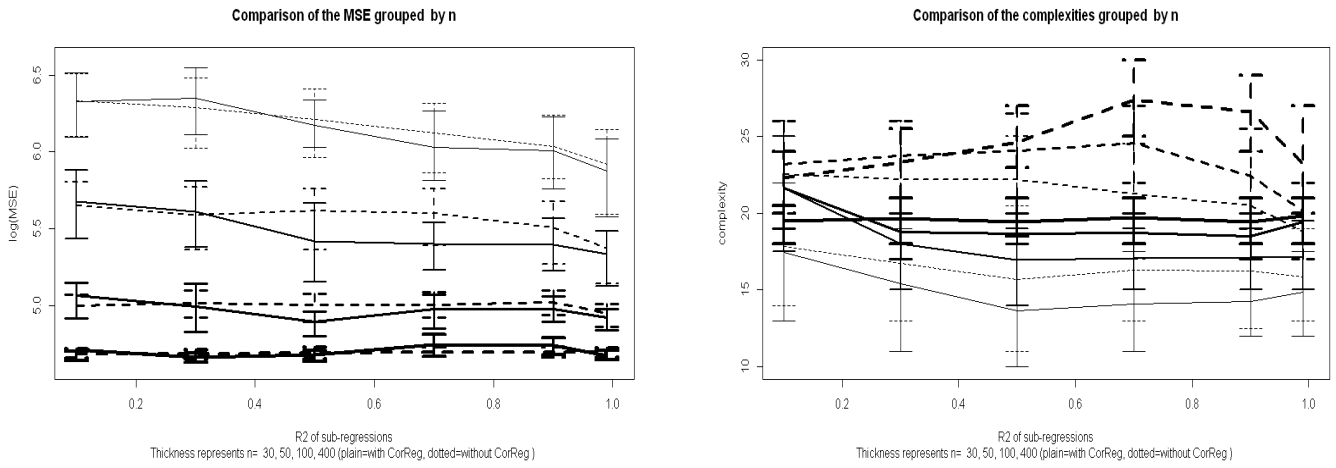
(a) MSE on a validation subset (b) Complexity

Figure 8: Comparison of OLS (dotted) and CorReg + OLS (plain) when $\boldsymbol{Y}$ depends only on $\boldsymbol{X}_f$. Thickness represents $n = 30, 50, 100, 400$. Inter-quartile intervals are also displayed.

switching variables is not a problem anymore even with large values of n.

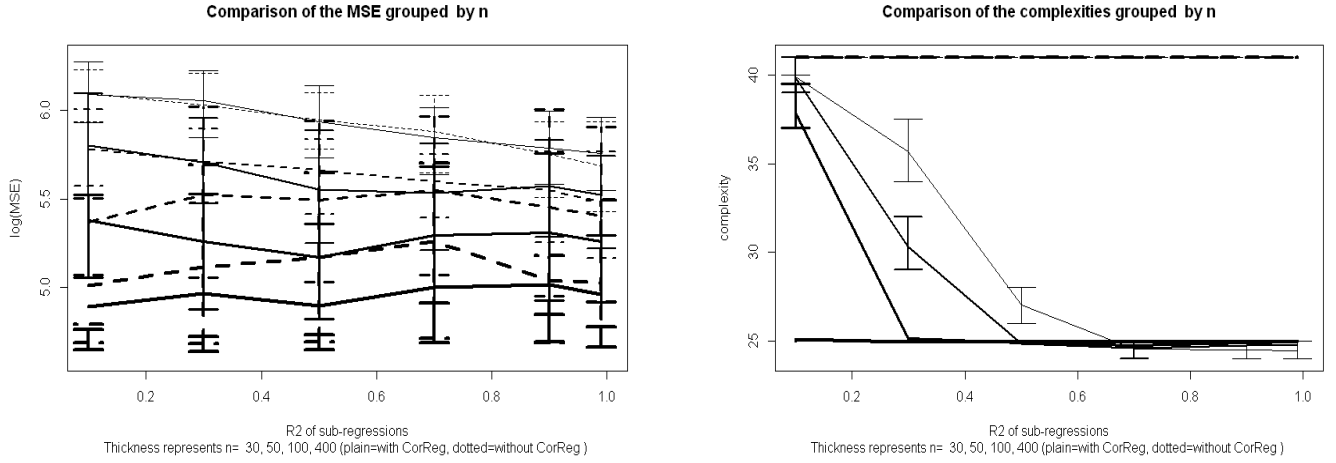
Figure 12 compares LASSO with CorReg+LASSO. It shows that LASSO naturally tends to keep $\boldsymbol{X_r}$ and thus is better because it corresponds to the true model. So CorReg is almost always the worst method. In such a case, it is recommended to compare both LASSO with CorReg+LASSO with a model choice criterion. Note that this model choice is only between two models so it avoids multiple tests issues and computational cost explosion. Thus, we suggest to always compute the "with CorReg" and the "without CorReg" solutions and then to compare them with the more pertinent criterion (AIC, cross-validation, validation sample, *etc.*) according to the context (size of the datasets for example).

Remark Since the structure $\boldsymbol{S}$ does not depend on $\boldsymbol{Y}$, it can be interesting for interpretation in cases when the "with CorReg" solutions are not kept in favor of "without CorReg" solutions. CorReg can then be seen both not only like a pre-treatment but also like a pre-study.



(a) MSE on a validation subset (b) Complexity

Figure 9: Comparison of LASSO (dotted) and CorReg + LASSO (plain) when $\boldsymbol{Y}$ depends only on $\boldsymbol{X}_f$. Thickness represents $n = 30, 50, 100, 400$. Inter-quartile intervals are also displayed.

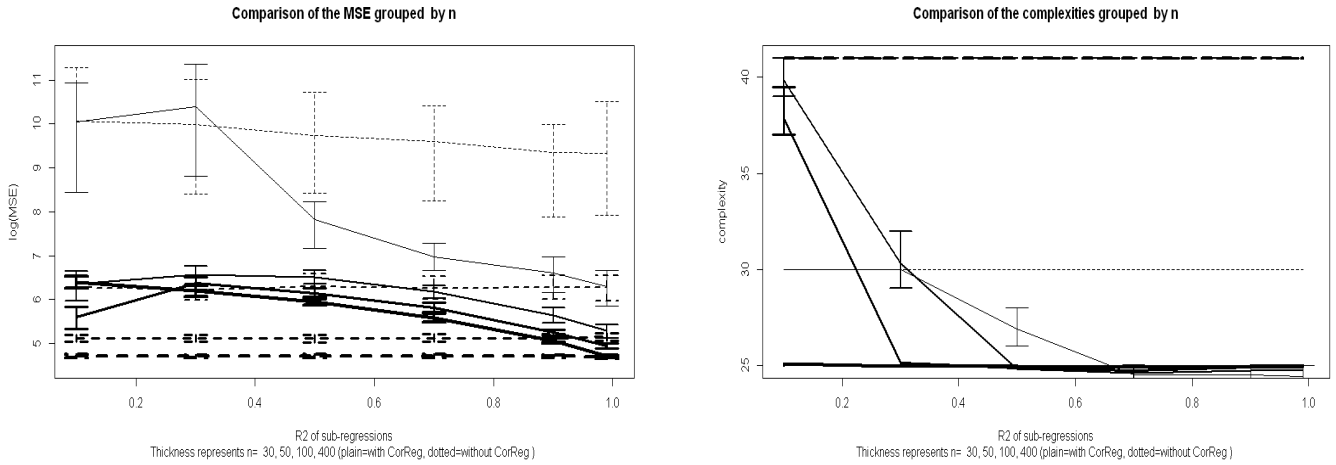


(a) MSE on a validation subset (b) Complexity

Figure 10: Comparison of ridge (dotted) and CorReg + ridge (plain) when $\boldsymbol{Y}$ depends only on $\boldsymbol{X}_f$y. Thickness represents $n = 30, 50, 100, 400$. Inter-quartile intervals are also displayed.

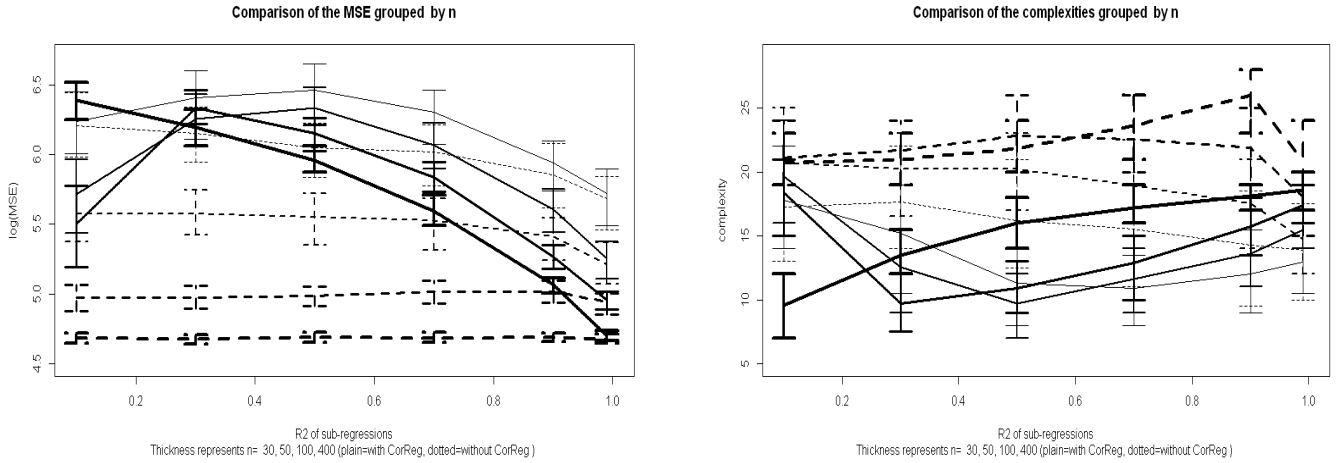
4.4 Robustess study through a non-linear case

We have generated a non-linear sub-regression to test the robustness of our model. $\boldsymbol{X}_f$ is a set of 6 independent Gaussian mixtures defined as before. We design a unique, possibly non linear, sub-regression $\boldsymbol{X}_7 = a\boldsymbol{X}_1^2 + \boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{\varepsilon}_1$. The matrix $\boldsymbol{X}$ is then scaled and we set $\boldsymbol{Y} = \sum_{i=1}^{7} \boldsymbol{X}_i + \boldsymbol{\varepsilon}_Y$. We let a vary between 0 and 10 to increase progressively the non-linear part of the sub-regression. Figure 13 shows that the MCMC algorithm has more difficulties to find a linear structure (by BIC_H as the non-linear part of the sub-regression increases with a. But the model is quite robust, preserving good efficient for small values of a. In addition, Figure 14 illustrates the advantage of using CorReg, even with non-linear sub-regressions, concerning the quality of the MSE.



(a) MSE on a validation subset (b) Complexity

Figure 11: Comparison of OLS (dotted) and CorReg + OLS (plain) when $\boldsymbol{Y}$ depends only on $\boldsymbol{X}_r$. Thickness represents $n = 30, 50, 100, 400$. Inter-quartile intervals are also displayed.



(a) MSE on a validation subset (b) Complexity

Figure 12: Comparison of LASSO (dotted) and CorReg + LASSO (plain) when $\boldsymbol{Y}$ depends only on $\boldsymbol{X}_r$. Thickness represents $n = 30, 50, 100, 400$. Inter-quartile intervals are also displayed.

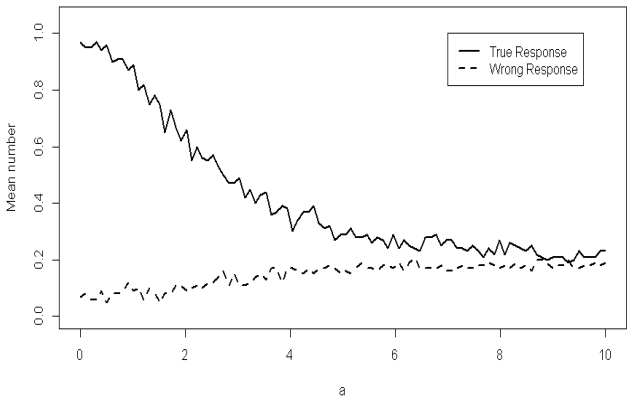


Figure 13: Evolution of the quality of $\hat{\boldsymbol{S}}$ when the paramater a increases.

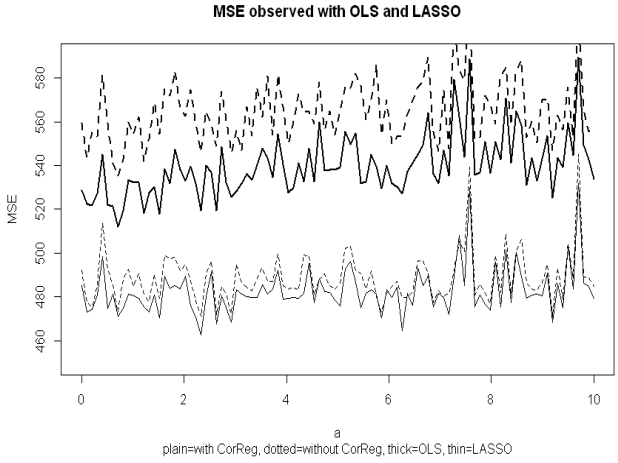


Figure 14: MSE on the main regression for OLS (thick) and LASSO (thin) used both with (plain) or without CorReg (dotted).

5 Numerical results on two real datasets

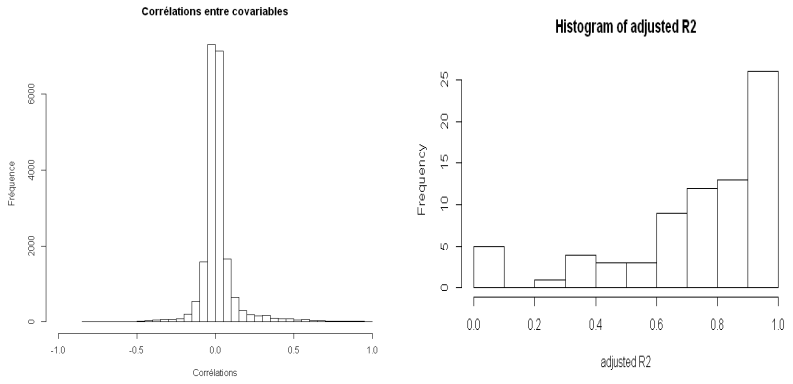
5.1 Quality case study in steel industry

This work takes place in steel industry context, with a quality oriented objective. Indeed, the purpose is to understand and to prevent quality problems on finished products, knowing the whole process. The correlations between involved features can be strong here because many parameters of the whole process are highly correlated (physical laws, process rules, *etc.*). We have a quality parameter (confidential) as response variable $\boldsymbol{Y}$ and $d = 205$ variables from the whole process to explain it. We get a training set of $n = 3\,000$ products described by these 205 variables from the industrial process and also a validation sample of 847 products.

The objective here is not only to predict non-quality but to understand and then to avoid it. CorReg provides an automatic method without any *a priori* and can be combined with any variable selection methods. So it allows to obtain, in a small amount of time (several hours for this dataset), some indications on the source of the problem, and to use human resources efficiently. When quality crises occur, time is extremely precious so automation is a real stake. The combinatorial aspect of the sub-regression models makes it impossible to do manually.

To illustrate that some industrial variables are naturally highly correlated, we can measure the correlation ρ between some couple of variables. For instance, the width and the weight of a steel slab gives $|\rho| = 0.905$, the temperature before and after some tool gives $|\rho| = 0.983$, the roughness of both faces of the product gives $|\rho| = 0.919$ and a particular mean and a particular max gives $|\rho| = 0.911$. For an overview of correlations, Figure 15(a) gives an histogram of ρ where we can see that, however, many other variables are not so highly correlated.

CorReg estimated a structure of $d_r = 76$ sub-regressions with a mean of $\bar{\boldsymbol{d}}_p = 5.17$ predictors. In the resulting uncorrelated covariate set $\boldsymbol{X}_f$ the number of values $|\rho| > 0.7$ is 79.33% smaller than in $\boldsymbol{X}$. Indeed, Figure 15(b) displays the histogram of adjusted R^2 value (R^2_{adj}) and we can see that essentially large values of R^2_{adj} are present. When we have a look at a more detailed level, we can see also that CorReg has been able non only to retrieve the above correlations (the width and the weight of a steel slab, *etc.*) but also to detect more complex structures describing physical models, like the width in function of the mean flow and the mean speed, even if the true physical model is not linear since "width = flow / (speed * thickness)" (here thickness is constant). Non-linear regulation models used to optimize the process were also found (but are confidential). These first results are easily understandable and meet metallurgists expertise. Sub-regressions with small values of R^2 are associated with non-linear model (chemical kinetics for example).

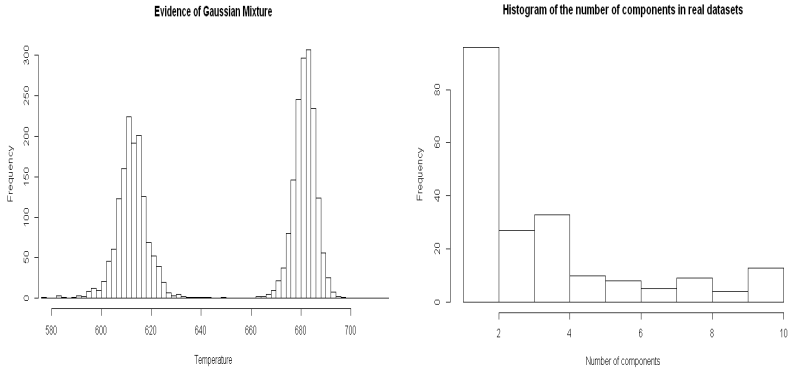


(a) (b)

Figure 15: Quality case study: (a) Histogram of correlations ρ in $\boldsymbol{X}$, (b) histogram of the adjusted R^2_{adj} for the $d_r = 76$ sub-regressions.

Note that the uncorrelated variables can be very well-modeled by parsimonious Gaussian mixtures as it is illustrated by Figure 16(a). In particular, the number of components is quite moderate as seen

in Figure 16(b).



(a) (b)

Figure 16: Quality case study: (a) Example of a non-Gaussian real variable easily modeled by a Gaussian mixture, (b) distribution of the number of components found for each covariate.

Table 1 displays predictive results associated to different estimation methods with and without CorReg. We can see that CorReg improves the results for each method tested in terms of prediction, with generally a more parsimonious regression on $\boldsymbol{Y}$. In terms of interpretation, this regression gives a better understanding of the consequences of corrective actions on the whole process. It typically permits to determine the *tuning parameters* whereas variable selection alone would point variables we can not directly act on. So it becomes easier to take corrective actions on the process to reach the goal. The stakes are so important that even a little improvement leads to consequent benefits, and we do not even talk about the impact on the market shares that is even more important.

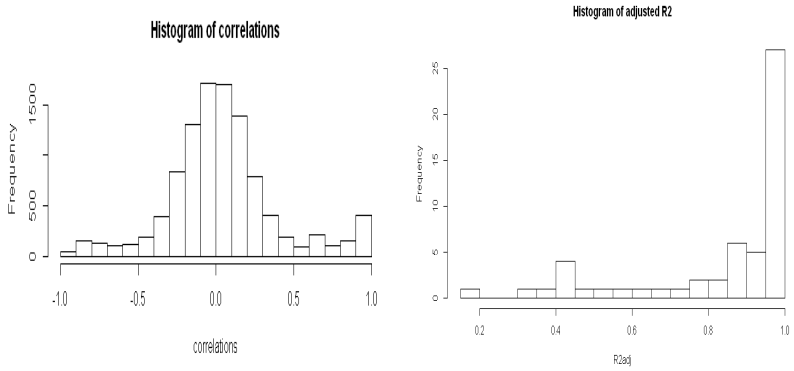
Method	Indicator	With CorReg	Without CorReg
OLS	MSE	13.30	14.03
	complexity	130	206
LASSO	MSE	12.77	12.96
	complexity	24	21
elasticnet	MSE	**12.15**	13.52
	complexity	40	78
ridge	MSE	12.69	13.09
	complexity	130	206

Table 1: Quality case study: Results obtained on a validation sample ($n = 847$ individuals). In bold, the best MSE value.

5.2 Production case study

This second example is about a phenomenon that impacts the productivity of a steel plant. It is described by a (confidential) response variable $\boldsymbol{Y}$ and $d = 145$ covariates from the whole process to explain it but only $n = 100$ individuals are present. The stake is to gain 20% of productivity on a specific product with high added value.

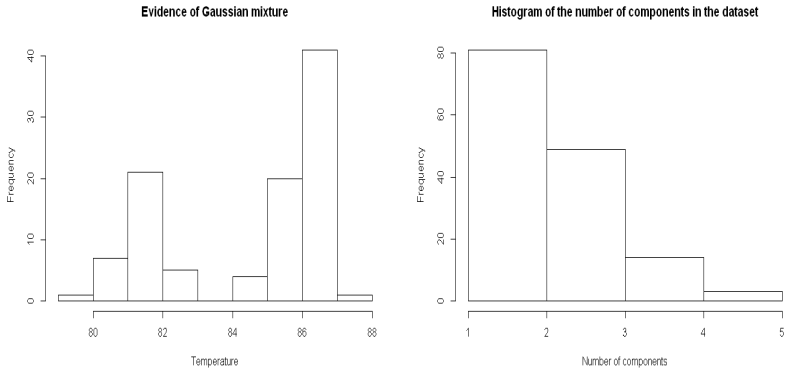
Figure 17(a) shows that many variables are highly correlated. CorReg found $d_r = 55$ sub-regressions and corresponding R^2_{adj} values are displayed in Figure 17(b). One of them seems to be weak ($R^2_{adj} = 0.17$) but it corresponds in fact to a non-linear regression: It points out a link between diameter of a coil and some shape indicator. In this precise case, CorReg found a structure that helped to decorrelate covariates and to find the relevant part of the process to optimize. This product is made by a long process that requires several steel plants so it was necessary to point out the steel plant where the problem occurred.



(a) (b)

Figure 17: Production case study: (a) Histogram of correlations ρ in $\boldsymbol{X}$, (b) histogram of the adjusted R^2_{adj} for the $d_r = 55$ sub-regressions.

As in the previous quality case study, we note that the uncorrelated variables can be very well-modeled by parsimonious Gaussian mixtures as it is illustrated by Figure 18(a). In particular, the number of components is really moderate as seen in Figure 18(b).



(a) (b)

Figure 18: Production case study: (a) Example of a non-Gaussian real variable easily modeled by a Gaussian mixture, (b) distribution of the number of components found for each covariate.

Table 2 displays predictive results associated to different estimation methods with and without CorReg. Note that mse is calculated though a leave-one-out method because of the small sample size. We can again see that CorReg globally improves the results for each method tested in terms of prediction, with always a more parsimonious regression on $\boldsymbol{Y}$.

Method	Indicator	With CorReg	Without CorReg
OLS	MSE	1.95	51 810
	complexity	91	100
LASSO	MSE	**0.106**	0.120
	complexity	27	34
elasticnet	MSE	0.140	0.148
	complexity	10	13
ridge	MSE	0.179	0.177
	complexity	91	146

Table 2: Production case study: Results obtained with leave-one out cross-validation ($n = 100, d = 145$). MSE is calculated though a leave-one-out method because of the small sample size. In bold, the best MSE value.

6 Conclusion and perspectives

We have seen that correlations can lead to serious estimation and variable selection problems in linear regression. In such a situation, it can be useful to explicitly model the structure between the covariates and to use this structure by simple probabilistic marginalization to avoid correlations issues. It has led to the so called CorReg method, which can be viewed as a variable pre-selection based on covariates correlations and which has to be then combined with any standard estimate and variable selection procedure. The CorReg strategy is able to give not only efficiently prediction but also a better understanding of the phenomenon at hand thanks to the sub-regressions description of correlated covariates. Its strength is then its great interpretability of the model, composed of several short linear regression easily managed by non-statisticians. In particular, we have illustrated both advantages of CorReg in two industrial contexts related to the steel industry.

Future works we plan is to allow CorReg to manage missing values. They are very common in industry for instance. Indeed, the structure can be used to estimate missing values in $\boldsymbol{X}$ thanks to the full generative model assumed in CorReg. Another perspective would be to take back lost information (the residual of each sub-regression) to improve predictive efficiency when needed. It would only consists in a second step of linear regression between the residuals and would thus still be able to use any selection method.

Package `CorReg` is accessible on CRAN: http://cran.r-project.org/web/packages/CorReg/index.html

Acknowledgements We want to thank ArcelorMittal Atlantique & Lorraine that has granted this work, given the chance to use CorReg on real datasets and authorized the package to be open-sourced licensed (CeCILL), especially Dunkerque's site where most of the work has been done.

References

[1] C. Andrieu and A. Doucet. Joint bayesian model selection and estimation of noisy sinusoids via reversible jump mcmc. *Signal Processing, IEEE Transactions on*, 47(10):2667–2676, 1999.

[2] H.D. Bondell and B.J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

[3] H. Chipman, E.I. George, R.E. McCulloch, M. Clyde, D.P. Foster, and R.A. Stine. The practical implementation of bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134, 2001.

[4] R. Davidson and J.G. MacKinnon. Estimation and inference in econometrics. *Oxford University Press Catalogue*, 1993.

[5] A.P. Dempster, N.M. Laird, D.B. Rubin, et al. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, 39(1):1–38, 1977.

[6] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[7] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov chain Monte Carlo in practice*, volume 2. CRC press, 1996.

[8] C. M. Grinstead and J.L. Snell. *Introduction to probability*. American Mathematical Society, 1997.

[9] A.E. Hoerl and R.W. Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, pages 69–82, 1970.

[10] H. Ishwaran and J.S. Rao. Spike and slab variable selection: frequentist and bayesian strategies. *Annals of Statistics*, pages 730–773, 2005.

[11] T. Isobe, E.D. Feigelson, M.G. Akritas, and G.J. Babu. Linear regression in astronomy. *The astrophysical journal*, 364:104–113, 1990.

[12] V.N. Katsikis and D. Pappas. Fast computing of the moore-penrose inverse matrix. *Electronic Journal of Linear Algebra*, 17(1):637–650, 2008.

[13] É. Lebarbier and T. Mary-Huard. Une introduction au critère bic: fondements théoriques et interprétation. *Journal de la SFdS*, 147(1):39–57, 2006.

[14] R. Lebret, S. Iovleff, F. Langrognet, C. Biernacki, and G. Celeux, Govaert. Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library. November 2013.

[15] N.T. Longford. A revision of school effectiveness analysis. *Journal of Educational and Behavioral Statistics*, 37(1):157–179, 2012.

[16] D.W. Marquardt and R.D. Snee. Ridge regression in practice. *American Statistician*, pages 3–20, 1975.

[17] Cathy Maugis, Gilles Celeux, and M-L Martin-Magniette. Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics & Data Analysis*, 53(11):3872–3882, 2009.

[18] G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[19] Alan Miller. *Subset selection in regression*. CRC Press, 2002.

[20] D.C. Montgomery, E.A. Peck, and G. Vining. *Introduction to linear regression analysis*, volume 821. John Wiley & Sons, 2012.

[21] Adrian E Raftery and Nema Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.

[22] A.E. Raftery. Bayesian model selection in social research. *Sociological methodology*, 25:111–164, 1995.

[23] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[24] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[25] N.H. Timm. *Applied multivariate analysis*. Springer Verlag, 2002.

[26] L. Yengo, J. Jacques, C. Biernacki, et al. Variable clustering in high dimensional linear regression models. 2012.

[27] A. Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, pages 348–368, 1962.

[28] Yongli Zhang and Xiaotong Shen. Model selection procedure for high-dimensional data. *Statistical Analysis and Data Mining*, 3(5):350–358, 2010.

[29] P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, December 2006.

[30] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Appendix

7 Identifiability

7.1 Definition

We call identifiability:

$$\nexists(\boldsymbol{S}, \tilde{\boldsymbol{S}}) \in \mathcal{S}_d \times \mathcal{S}_d \text{ with } \boldsymbol{S} \neq \tilde{\boldsymbol{S}} \text{ and } \mathbb{P}(\boldsymbol{X}; \boldsymbol{S}) = \mathbb{P}(\boldsymbol{X}; \tilde{\boldsymbol{S}}) \tag{33}$$

To avoid label-switching consideration, we suppose here (without loss of generality) that $\boldsymbol{J}_r$ is ordered by ascending order of the labels of the covariates. Hence identifiability is paired with the hypotheses we made on $\mathcal{S}_d$. It is not sufficient to find a structure of linear sub-regression, the structure also has to verify hypotheses 1 to 3 (uncrossing rule + dependencies exhaustively described by the structure and then independence between the conditional response covariates). As a consequence, the covariance between two covariates is not null if and only if these covariates are linked by some sub-regressions.

7.2 Sufficient condition for identifiability

Identifability criterion: The model $\boldsymbol{S}$ is identifiable if

$$\forall j \in \{1, \dots, d_r\}, d_p^j > 1. \tag{34}$$

That is to have at least two regressors in each sub-regression.

To prove the sufficiency of this condition for identifiability we rely on the following lemma.

Lemma: With $\boldsymbol{X}$ and $\boldsymbol{S}$ following hypotheses 1 to 3, covariance between two distinct covariates does differ from 0 in only two cases:

1. One of the two variables is a regressor of the other in a sub-regression

$$j \in \{1, \dots, d_r\}, i \in J_p^j \text{ then } \operatorname{cov}(\boldsymbol{X}^i, \boldsymbol{X}^{J_r^j}) \neq 0 \tag{35}$$

2. Both variables are regressed by a common covariate in their respective sub-regression:

$$\begin{aligned} \exists k \in J_f, \exists (i,j) \in \{1, \dots, d_r\} \times \{1, \dots, d_r\} \text{ with } i \neq j, \qquad & k \in J_p^i \text{ and } k \in J_p^j \\ \text{then } \operatorname{cov}(\boldsymbol{X}^{J_r^i}, \boldsymbol{X}^{J_r^j}) \quad & \neq \quad 0 \end{aligned} \tag{36}$$

proof of the lemma: The two cases lead immediately to non-zero covariance so we just look at other combinations of covariates.

- if $\exists(i,j) \in \{1,\dots,d_r\} \times \{1,\dots,d_r\}, \text{cov}(\boldsymbol{X}^{J_r^i}, \boldsymbol{X}^{J_r^j}) \neq 0$ then hypothesis 2 (uncrossing rule) guarantee that the two covariates are not in a same sub-regression so the covariance must come from the noises of the sub-regression but hypothesis 3 say that they are independent. The only remaining case is then the second case of the lemma: common covariate in the sub-regressions.
- if $(i,j) \in J_f \times J_f$ then $\text{cov}(\boldsymbol{X}^i, \boldsymbol{X}^j) = 0$ because covariates in $\boldsymbol{X}_f$ are orthogonal (by hypotheses 1 and 2).
- if $j \in \{1,\dots,d_r\}, i \in J_f$ and $\text{cov}(\boldsymbol{X}^{J_r^j}, \boldsymbol{X}^i) \neq 0$ then $i \in J_p^j$ by hypotheses 1 and 3 and equation 6.

□

proof of the identifiability criterion: We suppose that 34 is verified and the model is not identifiable:

$$\exists(\boldsymbol{S}, \tilde{\boldsymbol{S}}) \in \mathcal{S}_d \times \mathcal{S}_d \text{ with } \boldsymbol{S} \neq \tilde{\boldsymbol{S}} \text{ and } \mathbb{P}(\boldsymbol{X}; \boldsymbol{S}) = \mathbb{P}(\boldsymbol{X}; \tilde{\boldsymbol{S}}) \tag{37}$$

$\tilde{\boldsymbol{S}} = (\tilde{\boldsymbol{J}}_r, \tilde{\boldsymbol{J}}_p)$ contains $\tilde{d}_r$ sub-regressions and is characterized by $\tilde{\boldsymbol{J}}_r = (\tilde{J}_r^1, \dots, \tilde{J}_r^{\tilde{d}_r}), \tilde{\boldsymbol{J}}_p = (\tilde{J}_p^1, \dots, \tilde{J}_p^{\tilde{d}_r})$. Because $\boldsymbol{S} \neq \tilde{\boldsymbol{S}}$ we have $\boldsymbol{J}_r \neq \tilde{\boldsymbol{J}}_r$ or $\boldsymbol{J}_p \neq \tilde{\boldsymbol{J}}_p$.

- If $\boldsymbol{J}_r = \tilde{\boldsymbol{J}}_r$ and $\boldsymbol{S} \neq \tilde{\boldsymbol{S}}$ then one sub-regression as a predictor that stands only in one of the two structures. We suppose (without loss of generality) that $\exists j \in \{1,\dots,d_r\}$ for which $\exists i \in J_p^j$ with $i \notin \tilde{J}_p^j$ so $\text{cov}_S(\boldsymbol{X}^{J_r^j}, \boldsymbol{X}^i) \neq 0$ and $\text{cov}_{\tilde{S}}(\boldsymbol{X}^{J_r^j}, \boldsymbol{X}^i) = 0$ (from the lemma) so the two structure don't give the same joint distribution, leading to a contradiction.
- If $\boldsymbol{J}_r \neq \tilde{\boldsymbol{J}}_r$ then one of the two models has a sub-regression that is not in the other. We suppose (without loss of generality) that $\exists J_r^j \in J_r$ with $J_r^j \notin \tilde{J}_r$ then $J_r^j \in \tilde{J}_f$ (recall $J_f = \{1,\dots,d\} \setminus J_r$). We note that $J_r^j \in J_r$ means $\exists k_1 \neq k_2, \{k_1,k_2\} \subset J_p^j \subset J_f$. Then $\text{cov}_{\boldsymbol{S}}(\boldsymbol{X}^{J_r^j}, \boldsymbol{X}^{k_1}) \neq 0$ and $\text{cov}_{\boldsymbol{S}}(\boldsymbol{X}^{J_r^j}, \boldsymbol{X}^{k_2}) \neq 0$ so by the lemma k_1 and k_2 are responses variables in $\tilde{\boldsymbol{S}}$: $\exists(l_1,l_2) \in \{1,\dots,\tilde{d}_r\} \times \{1,\dots,\tilde{d}_r\}, \tilde{J}_r^{l_1} = k_1, \tilde{J}_r^{l_2} = k_2$ and J_r^j is a regressor of k_1 and k_2: $J_r^j \in J_p^{l_1}, J_r^j \in J_p^{l_2}$ thus $\text{cov}_{\tilde{\boldsymbol{S}}}(\boldsymbol{X}^{k_1}, \boldsymbol{X}^{k_2}) \neq 0$ that is not possible because $\{k_1,k_2\} \subset J_p^j \subset J_f$ and covariates in $\boldsymbol{X}^{J_f}$ are orthogonal by hypotheses.

Finally, condition 34 is sufficient for identifiability of $\boldsymbol{S}$. □

Remark: Because sub-regressions with at least two regressors are identifiable, the only non-identifiable sub-regressions could be those with only one regressor, leading only to pairwise correlations that can be seen directly in the correlation matrix. Such sub-regression can be permuted without any impact on interpretation so such trivial sub-regression are not a problem even if they may occur with real datasets. One more thing: exact sub-regression with at least two sub-regressors are identifiable with our hypotheses.