Model-based variable selection in regression with highly correlated variables.

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

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

Abstract. Linear regression outcomes are known to be damaged by highly correlated covariates. However many modern datasets are expected to convey more and more highly correlated covariates due to the global increase of the amount of variables in datasets. We propose to explicitly model the correlations by a family of linear regressions between the covariates, some covariates explaining others. It allows then to obtain by marginalisation on the explained covariates a parsimonious correlation-free regression model. It corresponds to a kind of variable selection preliminary step which has then to be followed by standard linear estimation methods including classical variables selection procedures for instance. The structure of correlations is found with an MCMC algorithm aiming at optimizing a specific BIC criterion. An R package (CorReg) available on the CRAN implements this new method which will be illustrated on both simulated datasets and real-life datasets from steel industry where correlated variables are frequent.

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

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 [18], industry (illustrated in the present paper), astronomy [10], sociology [13] ...

With the rise of informatics, datasets contain more and more covariates, increasing the chance to have correlated ones. Many estimators rely on the inversion of matrix that will be ill-conditioned in such a context, increasing the variance of these estimators and leading to misleading interpretations and reduced prediction efficiency.

We know the minimum-variance linear unbiased estimator for linear regression that is the Ordinary Least Squares (OLS) estimator but it suffers from ill-conditioned matrices inversion when the covariates are highly correlated (section 2.1). Many traditional try to reduce the variance introducing some bias to improve the bias-variance tradeoff and get better prediction by selection of only some covariates, grouping of some covariates, *etc.*

Ridge regression [14] proposes a biased estimator that can be written in terms of a parametric L_2 penalty but it is the same for each covariate no matter how much it is correlated or not to others.Moreover, coefficients tend to 0 but don't reach 0 so it gives difficult interpretations for large number of covariates. The absence of variable selection is not compatible with the need to find small set of relevant covariates to explain the response variable.

Real datasets often imply many irrelevant variables so variable selection should be favoured when possible to reduce the dimension and obtain models small enough to be understood. Moreover, selection may keep only uncorrelated covariates and thus resolve the correlations problem. Variable selection methods may add some bias by deleting some relevant covariates but reduce the variance by the dimension reduction. The Least Absolute Shrinkage and Selection Operator (LASSO [22]) 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. Indeed, LASSO is also known to face consistency problems [27] when confronted with correlated covariates. So the quality

of interpretation is compromised.

Elastic net [28] is a method developed to be a compromise between Ridge regression and the LASSO by a linear combination of L_1 and L_2 penalties. But it is based on the grouping effect so correlated covariates get similar coefficients and are selected together whereas LASSO will choose between one of them and will then obtain similar predictions with a more parsimonious model. Once again, nothing specifically aims to reduce the correlations. Interpretation will be misleading in both LASSO and Elastic net cases because nothing differentiates correlated and uncorrelated covariates.

Another way of improving the conditioning and the understandability is to consider clusters of variables with the same coefficients, like the Octogonal Shrinkage and Clustering Algorithm for Regression (OSCAR [2]) to reduce the dimension and also the correlations if correlated covariates are in the same clusters. The bias is added by the dimension reduction inherent to the coefficients clustering. The CLusterwise Effect REgression (CLERE [24]) 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 have a small number of groups of covariates and the mixture have few enough components the model will have a number of parameters to estimate 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.

But it requires to suppose having many covariates with the same level of effect on the response variable and seems to stay less efficient in prediction than elastic net. Spike and Slab variable selection [9] also relies on Gaussian mixture (the spike and the slab) hypothesis for the regression coefficients and gives a subset of covariates (not grouped) on which to compute OLS but has no specific protection against correlations issues. It is again a probable increase of the bias by variable selection to reduce the variance by dimension reduction. We see that none of the above methods takes explicitly the correlations into account, even if the clustering methods may group the correlated covariates together. These methods are not directly guided by the correlations whereas knowing the correlations between the covariates facilitates interpretation.

Outside of the linear regression field, model-based explicit correlations by linear sub-regressions between covariates claims to obtain good results in both interpretation and estimation quality [15]. But this is made in a clustering context only, with only irrelevant covariates being dependent from relevant ones (no sub-regressions between relevant or between irrelevant covariates) and the algorithm used to find the structure is a stepwise-like algorithm without protection against correlations [19] even if it is known to be often unstable [17]. We propose to transpose this method for linear regression with a specifically adapted algorithm to find the structure of sub-regression.

The idea is that if we know explicitly the correlations, we could use this knowledge to avoid the problem. We use the correlations as new information to reduce the variance without adding any bias. More precisely, we model the correlations with a system of linear sub-regressions for the joint distribution of the covariates to model the structure of the correlations. It helps to define the greatest set of orthogonal covariates to keep the maximum information but with an orthogonality constraint. Then we can define a marginal model with independent covariates. Thus we have an explained variable selection guided by the correlations and it improves interpretation, not only prediction. This marginal model still is the true model but in a different probability space, and OLS still gives an estimator without bias but with a reduced variance for the estimators. The model depends only on independent covariates because it is a marginal model on these covariates. This can be viewed as a pretreatment on the dataset that will be followed then by any other tools for estimation and dimension reduction without suffering from correlations. This pretreatment is specifically done to decorrelate the covariates, unlike methods described above. The linear structure is obtained by a MCMC algorithm optimizing the penalized likelihood of the joint distribution on the covariates, independently from the response variable. This algorithm is part of the R package CorReg accessible on CRAN.

This paper will first present the linear modelisation of the correlations and the by-product marginal

regression model before describing in Section 3 the random walk used to find the structure. We will then look at some numerical results on simulated (Section 4) and real industrial datasets (Section 5) before concluding and giving some perspectives in Section 6.

2 Model to select decorrelated covariates

2.1 A classical problem: correlations in regression

We note the linear regression model:

$$\boldsymbol{Y}_{|\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{X}$ is the $n \times p$ matrix of the explicative variables (that is a sub-matrix of $\tilde{\boldsymbol{X}}$ the $n \times \tilde{p}$ matrix of provided covariates), $\boldsymbol{Y}$ the $n \times 1$ response vector and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma_Y^2 \boldsymbol{I}_n)$ the noise of the regression, with $\boldsymbol{I}_n$ the n-sized identity matrix and $\sigma_Y > 0$. The $p \times 1$ vector $\boldsymbol{\beta}$ is the vector of the coefficients of the regression, that can be estimated by $\hat{\boldsymbol{\beta}}$ with Ordinary Least Squares (OLS):

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

with variance matrix

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

and without any bias. Estimation of $\boldsymbol{\beta}$ requires the inversion of $\boldsymbol{X}'\boldsymbol{X}$ which will be ill-conditioned or even singular if some covariates depend linearly from each other. Conditionning of $\boldsymbol{X}'\boldsymbol{X}$ get worse based on two aspects: the dimension p (number of covariates) of the model (the more covariates you have the greater variance you get) and the correlations within the covariates: strongly correlated covariates give bad-conditioning and increase variance of the estimators . When correlations between covariates are strong, the matrix to invert is ill-conditioned and the variance increases, giving unstable and unusable estimator [8]. Another problem is that matrix inversion requires $n \geq p$.

Running example: we look at a simple case with $p = 5$ variables defined by four independent scaled Gaussian $\mathcal{N}(0, 1)$ named $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{x}_3 = \boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{\varepsilon}_3$ where $\boldsymbol{\varepsilon}_3 \sim \mathcal{N}(\boldsymbol{0}, \sigma_3^2 \boldsymbol{I}_n)$. We also define another couple $\boldsymbol{x}_4, \boldsymbol{x}_5$ of covariates that are *i.i.d.* with $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and two *scenarii* for $\boldsymbol{Y}$ with $\boldsymbol{\beta} = (1, 1, 1, 1, 1)$ and $\sigma_Y \in \{10, 20\}$. It is clear that $\boldsymbol{X}'\boldsymbol{X}$ will become more ill-conditioned as σ_3 gets smaller.

2.2 Our proposal: modelisation of the correlations

We make the hypothesis that $\boldsymbol{X}$ can be described by a partition $\boldsymbol{X} = (\boldsymbol{X}_f, \boldsymbol{X}_r)$ given by an explicit structure S where variables in $\boldsymbol{X}_r$ are endogenous covariates resulting from linear sub-regressions based on $\boldsymbol{X}_f$, the submatrix of mutually independent exogenous covariates. So we model the correlations by $P(\boldsymbol{X}_r|\boldsymbol{X}_f)$ with $\boldsymbol{X}_f$ orthogonals. Then $\boldsymbol{X}_r$ is the $n \times p_r$ submatrix of $0 \leq p_r < p$ redundent covariates and $\boldsymbol{X}_f$ the $n \times (p - p_r)$ submatrix of the free (independent) covariates.

In the following, we note $\boldsymbol{X}^j$ the j^{th} column of $\boldsymbol{X}$. The structure S of p_r regressions within correlated covariates in $\boldsymbol{X}$ is described by:

$$\boldsymbol{X}_{r|\boldsymbol{X}_f,S} \text{ defined by } \forall \boldsymbol{X}^j \subset \boldsymbol{X}_r : \boldsymbol{X}^j_{|\boldsymbol{X}_f,S} = \boldsymbol{X}_f \boldsymbol{\alpha}_j + \boldsymbol{\varepsilon}_j \text{ with } \boldsymbol{\varepsilon}_j \sim \mathcal{N}(\boldsymbol{0}, \sigma_j^2 \boldsymbol{I}_n) \tag{4}$$

where $\boldsymbol{\alpha}_j \in \mathcal{R}^{(p-p_r)}$ are the sparse vectors of the regression coefficients between the covariates (each sub-regression freely implies different covariates).

The partition of $\boldsymbol{X}$ implies the uncrossing rule $\boldsymbol{X}_r \cap \boldsymbol{X}_f$ *i.e.* endogenous variables don't explain other covariates. This hypothesis ensures that S contains no cycle and is straightforward readable (no need to order the sub-regressions). It is not so restrictive because cyclic structures have no sense and any non-cyclic structure can be associated with a structure that verifies the uncrossing constraint by just successively replacing endogenous covariates by their sub-regression when they are also exogenous in some other sub-regressions.

We make the choice to distinguish the response variable from the other endogenous variables (that are on the left of a sub-regression). Thus we have one regression on the response variable ($P(\boldsymbol{Y}|\boldsymbol{X})$) and a system of sub-regressions (without the response variable: $P(\boldsymbol{X}_r|\boldsymbol{X}_f, S)$). Then we consider correlations between the explicative covariates of the main regression, not between the residuals. We see that the S does not depend on $\boldsymbol{Y}$ so it can be learnt independently, even with a larger dataset (if missing values in $\boldsymbol{Y}$).

The structure obtained gives a system of linear regression that can be viewed as a recursive Simultaneous Equation Model (SEM)[4] [23]. Here we suppose the $\boldsymbol{\varepsilon}_j$ independent but in other cases SUR (Seemingly Unrelated Regression [25]) takes into account correlations between residuals SUR (Seemingly Unrelated Regression [25]) and could be used to estimate the $\boldsymbol{\alpha}_j$.

In the running example: $\boldsymbol{X}_r = \boldsymbol{x}_3$, $\boldsymbol{X}_f = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_4, \boldsymbol{x}_5\}$, $p_r = 1$ and $\boldsymbol{\alpha}_3 = (1, 1, 0, 0)'$

2.3 A by-product model: marginal regression with decorrelated covariates

Now we know $P(\boldsymbol{X}_r|\boldsymbol{X}_f, S)$ by the structure of sub-regressions, we are able to define a marginal regression model $P(\boldsymbol{Y}|\boldsymbol{X}_f, S)$ based on the reduced set of independent covariates $\hat{\boldsymbol{\beta}}_f$ without significant information loss. We use the information of the correlations structure to rewrite the true model without bias in the marginal space defined by the independent covariates.
Using the partition $\boldsymbol{X} = [\boldsymbol{X}_f, \boldsymbol{X}_r]$ we can rewrite (1):

$$\boldsymbol{Y}_{|\boldsymbol{X}_f, \boldsymbol{X}_r, S} = \boldsymbol{X}_f\boldsymbol{\beta}_f + \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}_{\boldsymbol{Y}} \tag{5}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_f, \boldsymbol{\beta}_r) \in \mathcal{R}^p$ is the vector of the regression coefficients associated respectively to $\boldsymbol{X}_f$ and $\boldsymbol{I}_n$ the identity matrix. We note that (4) and (5) give also by simple integration on $\boldsymbol{X}_r$ a marginal regression model on $\boldsymbol{Y}$ *depending only on uncorrelated covariates* $\boldsymbol{X}_f$:

$$\begin{aligned} \boldsymbol{Y}_{|\boldsymbol{X}_f, S} &= \boldsymbol{X}_f(\boldsymbol{\beta}_f + \sum_{j \in I_r} \beta_j \boldsymbol{\alpha}_j) + \sum_{j \in I_r} \beta_j \boldsymbol{\varepsilon}_j + \boldsymbol{\varepsilon}_Y & (6) \\ &= \boldsymbol{X}_f\boldsymbol{\beta}_f^* + \boldsymbol{\varepsilon}_Y^* & (7) \end{aligned}$$

This model is still the true model and OLS estimator will still give an unbiased estimator, but its variance will be reduced by both dimension reduction and decorrelation (variables in $\boldsymbol{X}_f$ are independent so the matrix $\boldsymbol{X}_f'\boldsymbol{X}_f$ will be well-conditioned). So the information given by the structure S allows to reduce the variance without adding bias, by simple marginalization.
Nevertheless, to be able to compare the bias-variance tradeoff, we can see this model as a variable pre-selection independent of the response in $\boldsymbol{Y}_{|\boldsymbol{X}}$. We note that it is simply a linear regression on some of the original covariates so we only made a pretreatment on the dataset by selecting $\boldsymbol{X}_f$ because of the correlations given by S. So we also get the model

$$\boldsymbol{Y}_{|\boldsymbol{X}, S} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}_Y^* \text{ where } \boldsymbol{\beta}^* = (\boldsymbol{\beta}_f^*, \boldsymbol{\beta}_r^*) \text{ and } \boldsymbol{\beta}_r^* = \boldsymbol{0} \tag{8}$$

for which OLS estimator of the coefficients may be biased.

Running example: $\boldsymbol{Y}_{|\boldsymbol{X}_f} = 2\boldsymbol{x}_1 + 2\boldsymbol{x}_2 + \boldsymbol{x}_4 + \boldsymbol{x}_5 + \boldsymbol{\varepsilon}_3 + \boldsymbol{\varepsilon}_Y$

2.4 Strategy of use: pretreatment before classical estimation/selection methods

As a pretreatment, the model allows usage of any method in a second time to estimate $\boldsymbol{\beta}_f^*$, even with variable selection methods like LASSO or a best subset algorithm like stepwise [21]. However, we always have $\boldsymbol{X}_r = \boldsymbol{0}$

After selection and estimation we will obtain a model with *two steps of variable selection*: the decorrelation step by marginalization(coerced selection associated to redundant information defined in S) and the classical selection step, with different meanings for obtained zeros in $\hat{\boldsymbol{\beta}}_f^*$ (irrelevant covariates) and for $\hat{\boldsymbol{\beta}}_r^* = 0$ (redundant information). Thus we are able to distinguish the reasons of

selection and consistency issues don't mean interpretation issues any more. So we dodge the drawbacks of both grouping effect and variable selection.

The explicit structure is parsimonious and simply consists in linear regressions and thus is easily understood by non statistician, allowing them to have a better knowledge of the phenomenon inside the dataset and to take better actions. Expert knowledge can even be added to the structure, physical models for example.

Moreover, the uncrossing constraint (partition of $\boldsymbol{X}$) guarantee to keep a simple structure easily interpretable (no cycles and no chain-effect) and straightforward readable.

There is no theoretical guarantee that our model is better. It's just a compromise between numerical issues caused by correlations for estimation and selection versus increased variability due to structural hypothesis. We just play on the traditional bias-variance tradeoff.

2.5 Illustration of the tradeoff conveyed by the pretreatment

We compare the OLS estimator on $\boldsymbol{X}$ defined in section 2.1 with the estimator obtained by the pretreatment that is $\boldsymbol{X}_f$ selection.

For the marginal regression model defined in (7) we have the OLS unbiased estimator of $\boldsymbol{\beta}^*$:

$$\hat{\boldsymbol{\beta}}^*_f = (\boldsymbol{X}'_f\boldsymbol{X}_f)^{-1}\boldsymbol{X}'_f\boldsymbol{Y} \text{ and } \hat{\boldsymbol{\beta}}^*_r = \mathbf{0} \tag{9}$$

We see in (6) that it gives an unbiased estimation of $\boldsymbol{Y}$ and $\boldsymbol{\beta}^*$ but in terms of $\boldsymbol{\beta}$ this estimator is biased:

$$\mathrm{E}[\hat{\boldsymbol{\beta}}^*_f|\boldsymbol{X}_f] = \boldsymbol{\beta}_f + \sum_{j\in I_r}\beta_j\boldsymbol{\alpha}_j \text{ and } \mathrm{E}[\hat{\boldsymbol{\beta}}^*_r|\boldsymbol{X}_f] = \mathbf{0} \tag{10}$$

with variance:

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}^*_f|\boldsymbol{X}_f] = (\sigma^2_Y + \sum_{j\in I_r}\sigma^2_j\beta^2_j)(\boldsymbol{X}'_f\boldsymbol{X}_f)^{-1} \text{ and } \mathrm{Var}[\hat{\boldsymbol{\beta}}^*_r|\boldsymbol{X}_f] = \mathbf{0} \tag{11}$$

We see that the variance is reduced compared to OLS described in equation (3)(no correlations and smaller matrix give better conditioning) for small values of σ_j *i.e.* strong correlations. So we play on the bias-variance tradeoff, reducing the variance by adding a bias.

The Mean Squared Error (MSE) on $\hat{\boldsymbol{\beta}}$ is:

$$\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= \| \text{ Bias } \|^2_2 + \mathrm{Tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}})) && (12)\\
\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}) &= 0 + \sigma^2_Y\,\mathrm{Tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}) && (13)\\
\mathrm{MSE}(\hat{\boldsymbol{\beta}}^*_{OLS}|\boldsymbol{X}) &= \| \sum_{j\in I_r}\beta_j\boldsymbol{\alpha}_j \|^2_2 + \| \boldsymbol{\beta}_r \|^2_2 + (\sigma^2_Y + \sum_{j\in I_2}\sigma^2_j\beta^2_j)\,\mathrm{Tr}((\boldsymbol{X}'_f\boldsymbol{X}_f)^{-1}) && (14)
\end{aligned}$$

To better illustrate the bias-variance tradeoff, we look at the running example. We observe the theoretical Mean Squared Error (MSE) of the estimator of both OLS and CorReg's marginal model for several values of σ_3 (strength of the sub-regression) and n. Figure 2.5 shows the theoretical MSE evolution with the strength of the sub-regression:

$$1 - \mathcal{R}^2 = \frac{\mathrm{Var}(\boldsymbol{\varepsilon})_{\mathbf{3}}}{\mathrm{Var}(\boldsymbol{x}_3)} = \frac{\sigma^2_3}{\sigma^2_3 + 2} \tag{15}$$

It is clear in Figure 2.5 that the marginal model is more robust than OLS on $\boldsymbol{X}$. And when sub-regression get weaker ($1-\mathcal{R}^2$ tends to 1) it remains stable until extreme values (sub-regression nearly fully explained by the noise). We also see that the error implied by strong correlations shrinks with the rise of n. We see that σ_Y multiplies $\mathrm{Tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}})) = \mathrm{Tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}}_f)) + \mathrm{Tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}}_r))$ for both models but for the marginal model $\mathrm{Tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}}_r)) = 0$. Thus, when σ^2_Y rises it increases the advantage of CorReg versus OLS. It illustrates the importance of dimension reduction when the model has a strong noise (very usual case on real datasets where true model is not even exactly linear). Further results are provided in sections 4 and 5.

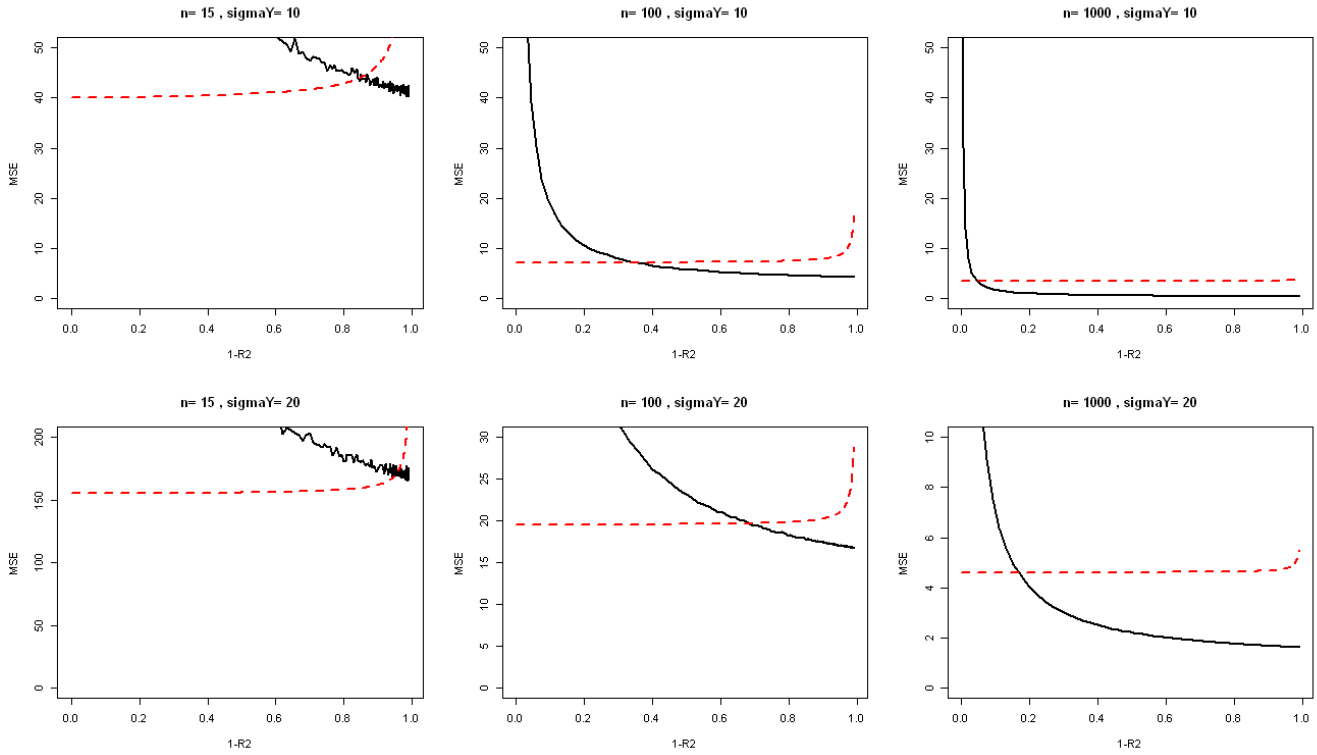

Figure 1: MSE of OLS and CorReg (dotted) estimators for varying $(1 - R^2)$ of the sub-regression, n and σ_Y.

3 Sub-regressions model selection

Structural equations models like SEM are often used in social sciences and economy where a structure is supposed "by hand" but here we want to find it automatically. Graphical LASSO [5] offers a method to obtain a structure based on the precision matrix (inverse of the variance-covariance matrix), setting some coefficients of the precision matrix to zero. But the resulting matrix is symmetric and we need an oriented structure for S to avoid cycles.

Cross-validation is very time-consuming and thus not friendly with combinatory problematics. Moreover, we need a criterion compatible with structures of different sizes (varying p_r) and not related with $\boldsymbol{Y}$ because the structure is inherent to $\boldsymbol{X}$ only. Thus it must be a global criterion. Because it is about model selection and we are able to provide a full generative model (section 3.1), we decide to follow a Bayesian approach ([20], [1],[3]).

We want to find the most probable structure S knowing the dataset, so we search for the structure that maximizes $P(S|\boldsymbol{X})$ and we have:

$$P(S|\boldsymbol{X}) \quad \propto \quad P(\boldsymbol{X}|S)P(S) = P(\boldsymbol{X}_r|\boldsymbol{X}_f, S)P(\boldsymbol{X}_f|S)P(S) \tag{16}$$

So we will try to maximize $\psi(\boldsymbol{X}, S) = P(\boldsymbol{X}|S)P(S)$.

3.1 Modeling the uncorrelated covariates: a full generative approach on $P(\boldsymbol{X})$

To be able to compare structures with $P(S|\boldsymbol{X})$, we need a full generative model on $\boldsymbol{X}$. Sub-regressions give $P(\boldsymbol{X}_r|\boldsymbol{X}_f, S)$ but $P(\boldsymbol{X}_f|S)$ is still undefined. We suppose that variables in $\boldsymbol{X}_f$ follow Gaussian mixtures of $k_j \in \mathbf{N}^*$ components:

$$\forall \boldsymbol{X}^j \notin \boldsymbol{X}_r : \boldsymbol{X}^j_{|S} \sim f(\boldsymbol{\theta}_j) = \mathcal{GM}(\boldsymbol{\pi}_j; \boldsymbol{\mu}_j; \boldsymbol{\sigma}^2_j) \text{ with } \boldsymbol{\pi}_j, \boldsymbol{\mu}_j, \boldsymbol{\sigma}^2_j \text{ vectors of size } K_j. \tag{17}$$

The great flexibility [16] of such models makes our model more robust. Gaussian case is just a special case ($K_j = 1$) of Gaussian mixture so it is included in our hypothesis but identifiability of S requires

to have Gaussian mixtures with at least two distinct components in each sub-regression (derived from the identifiability of the SR model in [15], more details in the Appendices 8.1).

Remark: Identifiability of S is not necessary to use a given structure but helps to find it.

Variables in $\boldsymbol{X}$ are in the followings supposed to be independent Gaussian mixtures with at least two distinct components each. We now have a full generative model.

3.2 Penalization of the integrated likelihood by $P(S)$

Our full generative generative model allows us to compare structures with criterions like the Bayesian Information Criterion (BIC) which penalize the log-likelihood of the joint law on $\boldsymbol{X}$ according to the complexity of the structure [12].

Uniform law on $P(S)$ gives $\psi(\boldsymbol{X}, S) \propto P(\boldsymbol{X}|S)$ so it is equivalent to a minimization of the BIC. We note $\boldsymbol{\Theta}$ the set of the parameters of the generative model

$$-2\log P(\boldsymbol{X}|S) \approx BIC = -2\mathcal{L}(\boldsymbol{X}, S, \boldsymbol{\Theta}) + |\boldsymbol{\Theta}|\log(n) \tag{18}$$

But BIC tends to give too complex structures because we test a great range of models. Thus we choose to penalise the complexity a bit more.

We note I_r the set of indices of endogenous variables in $\boldsymbol{X}$ (explained ones). We also define $I_f = \{I_f^1, \ldots, I_f^p\}$ the set of the sets of indices of exogenous covariates (explaining ones $= \boldsymbol{X}_f$) with $\forall j \notin I_r, I_f^j = \emptyset$. We see that I_f defines the non-null coefficients in $\boldsymbol{\alpha}_j$ (each sub-regression can be very parsimonious). Then we have the explicit structure characterized by $S = (I_f, I_r, p_f, p_r)$ where $p_r = |I_r|$, $\boldsymbol{p}_f = (p_f^1, \ldots, p_f^{p_r})$ is the vector of the number of covariates in each sub-regression and $p_f^j = |I_f^j|$, with $|.|$ the cardinal of an ensemble. Our running example is then described by $S = (\{\{1,2\}\}, \{3\}, (2), (1))$ We suppose a hierarchical uniform *a priori* distribution $P(S) = P(I_f|\boldsymbol{p}_f, I_r, p_r)P(\boldsymbol{p}_f|I_r, p_r)P(I_r|p_r)P(p_r)$ instead of the simple uniform law on S that is generally used and provides no penalty. Thus we have :

$$BIC_+(X|S) = BIC(X|S) - \ln(P(S)) \tag{19}$$

It increases penalty on complexity for $p_r \leq \frac{p}{2}$ and $p_f^j \leq \frac{p}{2}$. Hence this constraint on $\hat{p}_r$ and $\hat{p}_f^j$ is given in the research algorithm when the Hierarchical Uniform hypothesis is made instead of Uniform one in numerical experiments (section 4 and 5). BIC_+ does not change BIC but only $P(S)$ so the properties of BIC_+ are the same as classical BIC but we obtain better results when the constraints on the complexity are verified.

3.3 MCMC algorithm

Now we have a comparison criterion $\psi(\boldsymbol{X}, S)$, we define an MCMC algorithm to find the structure (R package CorReg on CRAN).

3.3.1 The neighbourhood

Let's define $\mathcal{S}$ the ensemble of feasible structures (those with $I_f \cap I_r = \emptyset$).
For each step, starting from $S \in \mathcal{S}$ we define a neighbourhood:

$$\mathcal{V}_{S,j} = \{S\} \cup \{S^{(i,j)} | 1 \leq i \leq p, i \neq j\} \tag{20}$$

$$\text{where} \quad j \sim \mathcal{U}(\{1, \ldots, p\}) \tag{21}$$

With $S^{(i,j)}$ defined by the following algorithm :

- if $i \notin I_f^j$ (add):
  - $I_f^j = I_f^j \cup \{i\}$
  - $I_f^i = \emptyset$ (explicative variables can't depend on others : column-wise relaxation)
  - $I_f = I_f \setminus \{j\}$ (dependent variables can't explain others : row-wise relaxation)

- else (remove): $I_f^j = I_f^j \setminus \{i\}$

At every moment, coherence between I_f and others parts of S can be done by $\forall 1 \leq j \leq p : p_f^j = |I_f^j|$, $I_r = \{j | p_f^j > 0\}$, $p_r = |I_r|$.

3.3.2 Transition probabilities

The walk follows a time-homogeneous Markov Chain whose transition matrix $\mathcal{P}$ has $|\mathcal{S}|$ rows and columns (combinatory so we just compute the probabilities when we need them). At each step the markov chain moves with probabiliy:

$$\forall (S, \tilde{S}) \in \mathcal{S}^2 : \mathcal{P}(S, \tilde{S}) \quad = \quad \sum_{j=1}^{p} \mathbf{1}_{\{\tilde{S} \in \mathcal{V}_{S,j}\}} \frac{\exp(-\frac{1}{2}\psi(\boldsymbol{X}, \tilde{S}))}{\sum_{S_l \in \mathcal{V}_{S,j}} \exp(-\frac{1}{2}\psi(\boldsymbol{X}, S_l))} \tag{22}$$

And $\mathcal{S}$ is a finite state space.

Because the walk follows a regular and thus ergodic markov chain with a finite state space, it has exactly one stationary distribution [7] and the output will be the best structure in terms of $P(S|\boldsymbol{X})$ which weights each candidate. Practically speaking, CorReg returns the best structure seen during the walk. Numerical results (Section 4) illustrates the efficiency of the walk when the true model really contains a linear structure or no structure at all (Table (1)) and when the structure is not linear (Table **??**)).

3.3.3 Initialisation(s)

If we have some knowledge about some sub-regressions (physical models for example) we can add them in the found and/or initial structure. So the model is really expert-friendly. The initial structure can be based on a first warming algorithm taking the correlations into account. Coefficients are randomly placed into I_f, weighted by the absolute value of the correlations. We do so in the followings. Then this structure could be for example reduced by the Hadamard product with the binary matrix obtained by Graphical Lasso[5] that makes selection in the precision matrix but it is time consuming.

One would rather test multiple short chains than lose time in initialisation or long chains [6]. It also helps to face local extrema. In the followings, the chain wzs launched with twenty initialisations each time.

4 Numerical results on simulated datasets

4.1 The datasets

Now we have defined the model and the way to obtain it, we can have a look on some numerical results to see if CorReg keeps its promises. The CorReg package has been tested on simulated datasets. Section 4.2.2 shows the results obtained in terms of $\hat{S}$. Sections **??** and 4.3.2 show the results obtained using only CorReg, or CorReg combined with other methods. Tables give both mean and standard deviation of the observed Mean Squared Errors (MSE) on a validation sample of 1000 individuals. For each simulation, $p = 40$, the R^2 of the main regression is 0.4, variables in $\boldsymbol{X}_f$ follow Gaussian mixture models of $\lambda = 5$ classes which means follow Poisson's law of parameter $\lambda = 5$ and which standard deviation is λ. The β_j and the coefficients of the $\boldsymbol{\alpha}_j$ are generated according to the same Poisson law but with a random sign. $\forall j \in I_r, p_1^j = 2$ (sub-regressions of length 2) and we have $p_r = 16$ sub-regressions. The datasets were then scaled so that covariates X_r don't have a greater variance or mean. We used Rmixmod to estimate the densities of each covariate. For each configuration, the MCMC walk was launched on 10 initial structures with a maximum of 1 000 steps each time. When $n < p$, a frequently used method is the Moore-Penrose generalized inverse [11], thus OLS can obtain some results even with $n < p$. When using penalized estimators for selection, a last Ordinary Least Square step is added to improve estimation because penalisation is made to select variables but also shrinks remaining coefficients. This last step allows to keep the benefits of

shrinkage (variable selection) without any impact on remaining coefficients (see [26]) and is applied for both classical and marginal model. We compare different methods with and without CorReg as a pretreatment. All the results are provided by the CorReg package.

4.2 Finding the structure

4.2.1 How to evaluate found structure?

The first criterion is $\psi(\boldsymbol{X}, S)$ which is maximized in the MCMC. But in our case, it is estimated by the likelihood (see (16))whose value don't have any intrinsic meaning. To show how far the found structure is from the true one in terms of S we define some indicators to compare the true model S and the found one $\hat{S}$. Global indicators :

- TL (True left) : the number of found dependent variables that really are dependent $TL = |I_r \cap \hat{I}_r|$
- WL (Wrong left) : the number of found dependent variables that are not dependent $WL = |\hat{I}_r| - TL$
- ML (Missing left) : the number of really dependent variables not found $ML = |I_r| - TL$
- Δp_r : the gap between the number of sub-regression in both model : $\Delta p_r = |I_r| - |\hat{I}_r|$. The sign defines if $\hat{S}$ is too complex or too simple
- $\Delta compl$: the difference in complexity between both model : $\Delta compl = \sum_{j \in p_r} p_f^j - \sum_{j \in \hat{p}_r} \hat{p}_f^j$

4.2.2 Results on S

We compare found structures in different contexts with both Uniform and Hierarchical Uniform *a priori* law on $P(S)$, noted respectively BIC and BIC_+. We see that usage of BIC_+ gives sparser models even with $\max(\hat{p}_f^j) = 5$. Moreover, this constraint is not active because observed complexities are smaller (this constraint only serves to accelerate the walk by reducing the dimension of $\mathcal{S}$ because each configuration was computed a hundred times), meaning that the stronger penalty implied by BIC_+ really is efficient. The datasets used for BIC and BIC_+ are the same to keep the comparison meaningful.

It is also notable that BIC_+ has a greater computational cost than BIC. We notice that the MCMC is faster when there are numerous correlations (rejecting more candidates).

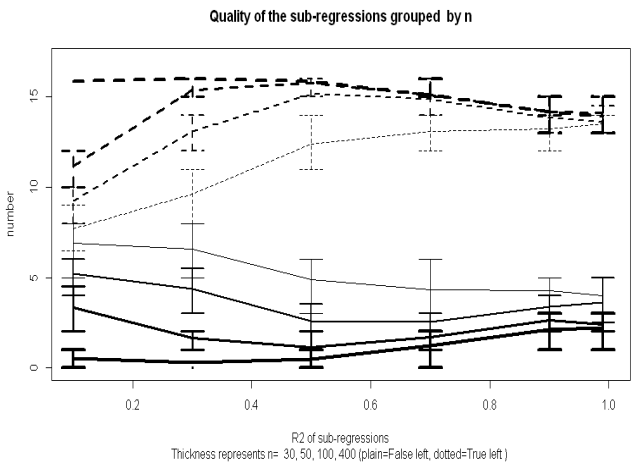

Figure 2: Quality of the subregressions found with classical BIC criterion

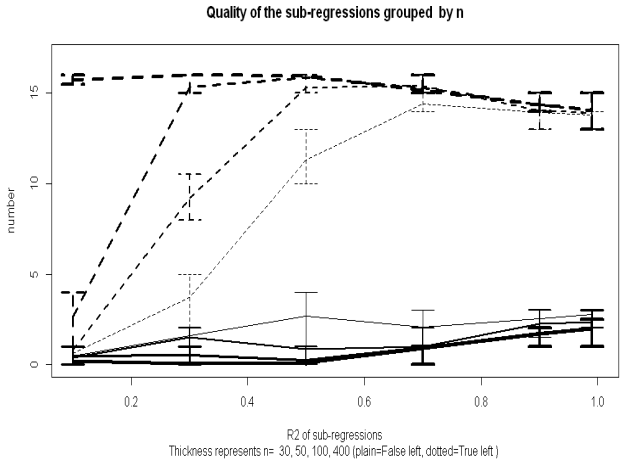

Figure 3: Quality of the subregressions found with our BIC_+ criterion

Configuration			Computing Time		Quality			Complexity	
n	p_r	$P(S)$	Time Mixmod	Time MCMC	TL	WL	ML	Δp_r	$\Delta compl$
30	0	BIC	0.4104	3.2428	0	5.43	0	-5.43	22.55
			(0.0275)	(0.3711)	(0)	(1.9346)	(0)	(1.9346)	(8.0884)
		BIC_+	0.4104	8.2338	0	0.53	0	-0.53	2.27
			(0.0275)	(0.9045)	(0)	(0.7844)	(0)	(0.7844)	(3.3024)
30	16	BIC	0.4182	2.7735	10.96	5.93	4.98	-0.95	38.16
			(0.0329)	(0.1498)	(1.9844)	(2.0313)	(1.9948)	(0.9987)	(6.499)
		BIC_+	0.4182	4.1876	11.61	4.57	4.33	-0.24	16.48
			(0.0329)	(0.2178)	(1.8743)	(1.9502)	(1.8752)	(0.4948)	(5.4892)
30	32	BIC	0.4456	2.9154	25.23	1.92	6.5	4.58	28
			(0.0429)	(0.1331)	(1.4761)	(1.0888)	(1.4668)	(0.9866)	(5.0831)
		BIC_+	0.4456	4.0233	16.96	3.04	14.77	11.73	4.35
			(0.0429)	(0.1091)	(1.3993)	(1.3993)	(1.4692)	(0.5478)	(5.8833)
50	0	BIC	0.5229	4.7068	0	4.2	0	-4.2	13.35
			(0.0519)	(0.5865)	(0)	(1.7233)	(0)	(1.7233)	(5.6468)
		BIC_+	0.5229	10.1198	0	0.13	0	-0.13	0.32
			(0.0519)	(0.5541)	(0)	(0.3667)	(0)	(0.3667)	(0.9732)
50	16	BIC	0.5205	3.3681	11.15	5.42	4.72	-0.7	22.85
			(0.0451)	(0.3123)	(1.93)	(1.9132)	(1.886)	(0.7317)	(5.7742)
		BIC_+	0.5205	4.909	11.42	4.59	4.45	-0.14	7.55
			(0.0451)	(0.4556)	(1.9079)	(1.8968)	(1.8333)	(0.3487)	(4.0611)
50	32	BIC	0.5833	3.2683	28.17	1.38	3.7	2.32	12.74
			(0.0628)	(0.316)	(1.3711)	(0.9077)	(1.3143)	(0.8394)	(4.3359)
		BIC_+	0.5833	4.3599	17.27	2.73	14.6	11.87	-2.61
			(0.0628)	(0.3729)	(1.1708)	(1.1708)	(1.2792)	(0.338)	(4.4854)
100	0	BIC	0.9623	12.9373	0	2.83	0	-2.83	6.23
			(0.077)	(1.7778)	(0)	(1.2953)	(0)	(1.2953)	(3.1999)
		BIC_+	0.9623	20.9817	0	0.01	0	-0.01	0.02
			(0.077)	(1.9421)	(0)	(0.1)	(0)	(0.1)	(0.2)
100	16	BIC	1.1223	6.9647	11.67	4.8	4.25	-0.55	12.58
			(0.1122)	(0.5473)	(2.0003)	(2.0646)	(1.956)	(0.7833)	(3.9471)
		BIC_+	1.1223	8.8486	12.04	3.95	3.88	-0.07	3.75
			(0.1122)	(0.7174)	(1.9223)	(1.9404)	(1.9137)	(0.2564)	(2.2625)
100	32	BIC	1.4343	5.9626	30.14	0.84	1.61	0.77	6.96
			(0.2528)	(0.3136)	(1.3928)	(0.8495)	(1.2941)	(0.7086)	(3.0975)
		BIC_+	1.4343	7.3741	17.49	2.51	14.26	11.75	-3.76
			(0.2528)	(0.2748)	(1.1849)	(1.1849)	(1.2441)	(0.4794)	(4.4859)

Table 1: Results of the Markov chain: Mean observed and standard deviation (sd).

4.2.3 Y depends only on covariates in X_f (best case for us)

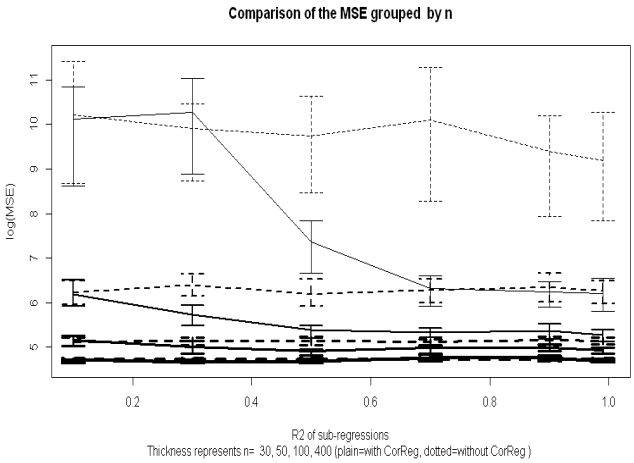

Figure 4: Comparison of the MSE between OLS and CorReg+OLS

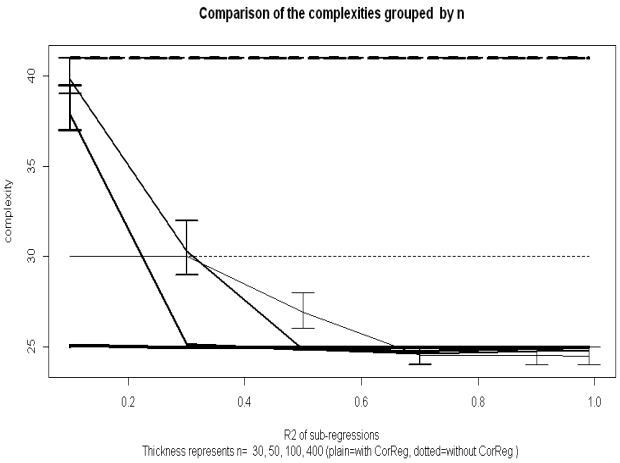

Figure 5: Comparison of the complexities between OLS and CorReg+OLS

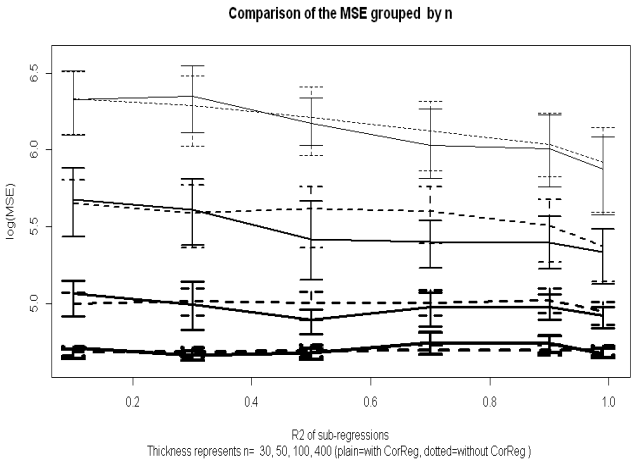

Figure 6: Comparison of the MSE between LASSO and CorReg+LASSO

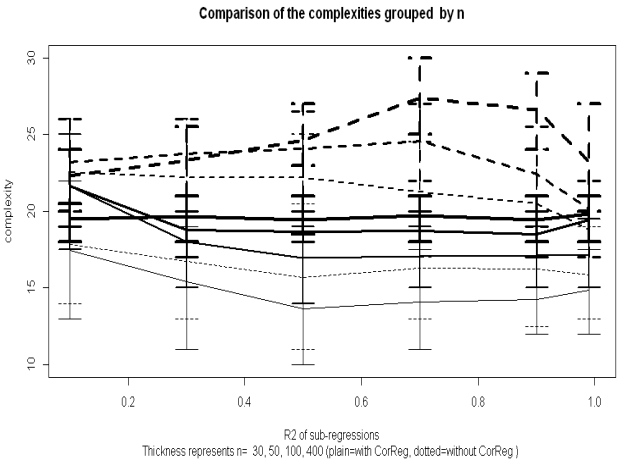

Figure 7: Comparison of the compexities between LASSO and CorReg+LASSO

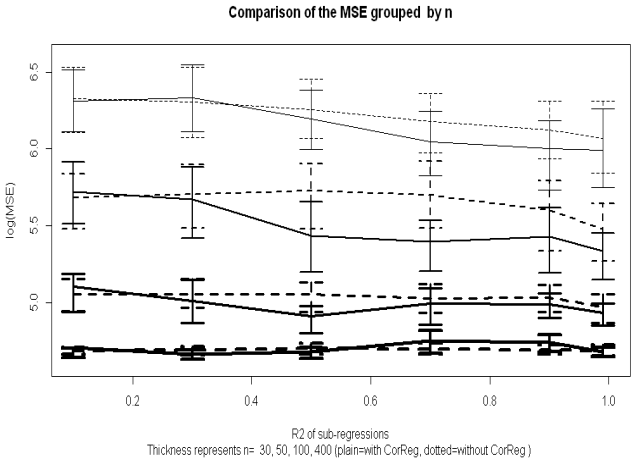

Figure 8: Comparison of the MSE between elasticnet and CorReg+elasticnet

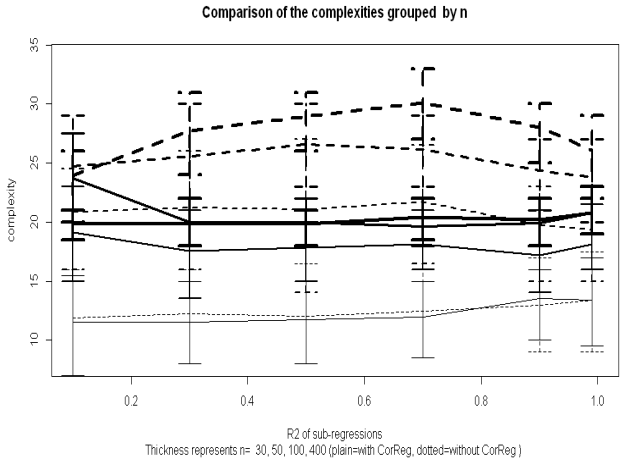

Figure 9: Comparison of the compexities between elasticnet and CorReg+elasticnet

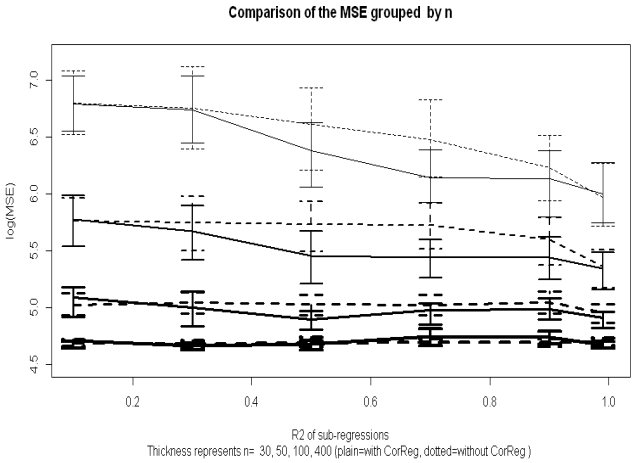

Figure 10: Comparison of the MSE between stepwise and CorReg+stepwise

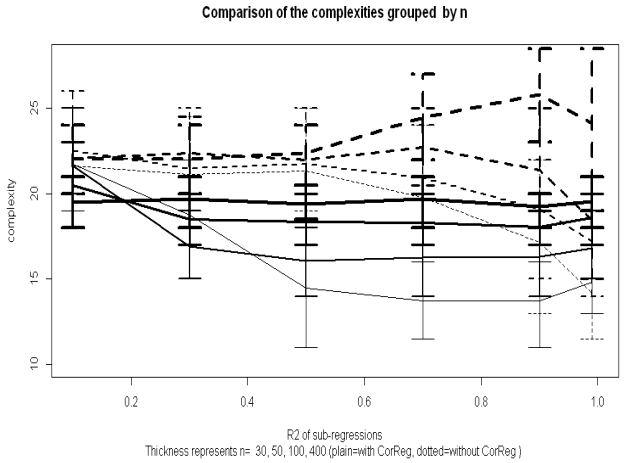

Figure 11: Comparison of the compexities between stepwise and CorReg+stepwise

4.3 Results on prediction

4.3.1 Y depends on all variables in X

We then try the method with a response depending on all covariates (CorReg reduces the dimension and can't give the true model if there is a structure). The datasets used here were those from table 1.

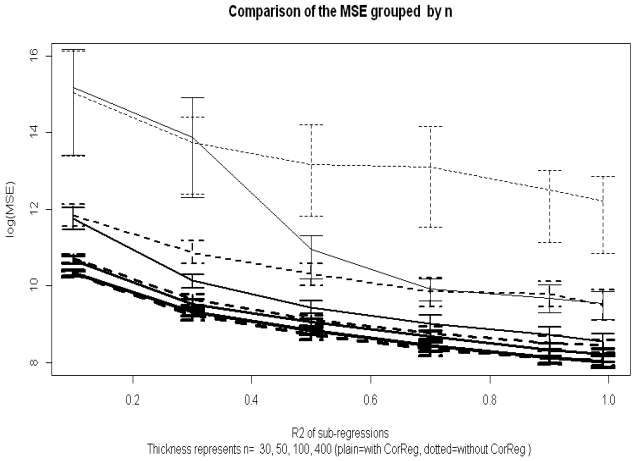

Figure 12: Comparison of the MSE between OLS and CorReg+OLS

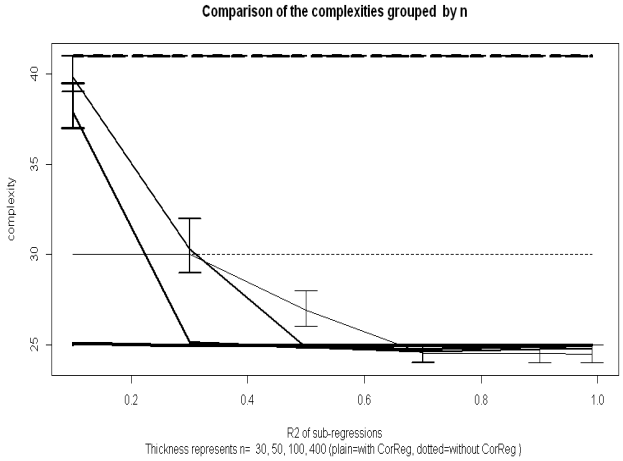

Figure 13: Comparison of the compexities between OLS and CorReg+OLS

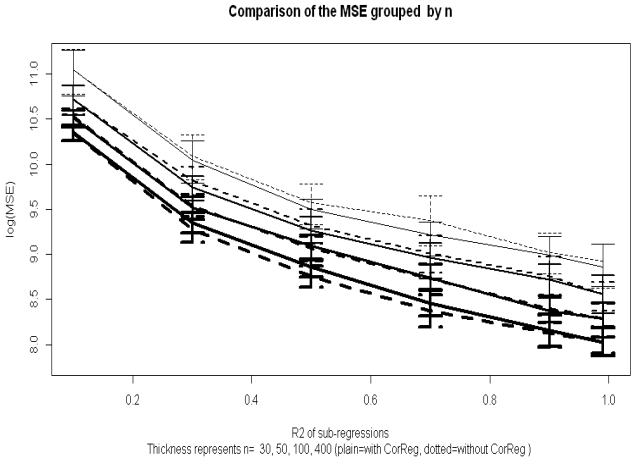

Figure 14: Comparison of the MSE between LASSO and CorReg+LASSO

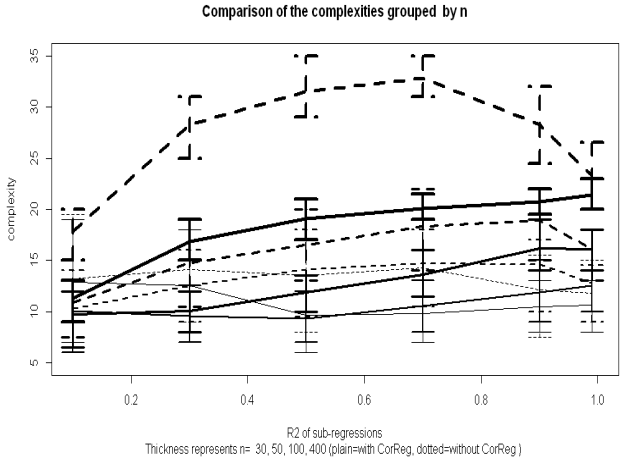

Figure 15: Comparison of the compexities between LASSO and CorReg+LASSO

We see that CorReg tends to give more parsimonious models and better predictions, even if the true model is not parsomious. We logically observe that when n rises, all the models get better and the correlations cease to be a problem so the complete model starts to be better (CorReg does not allow the true model to be choosen).

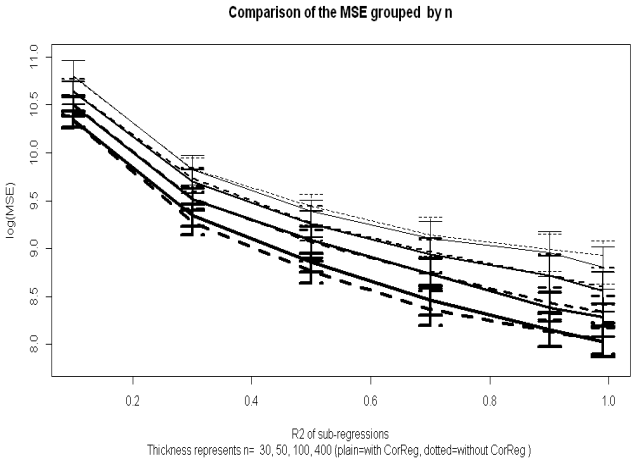

Figure 16: Comparison of the MSE between elasticnet and CorReg+elasticnet

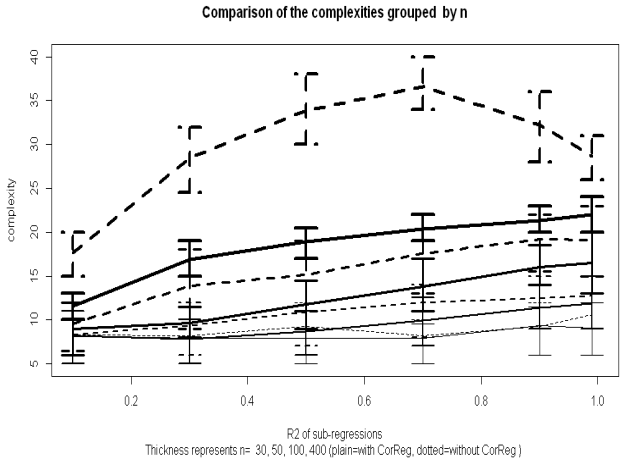

Figure 17: Comparison of the compexities between elasticnet and CorReg+elasticnet

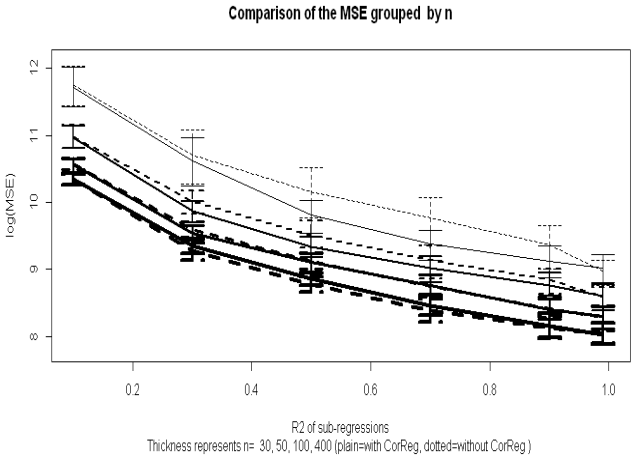

Figure 18: Comparison of the MSE between stepwise and CorReg+stepwise

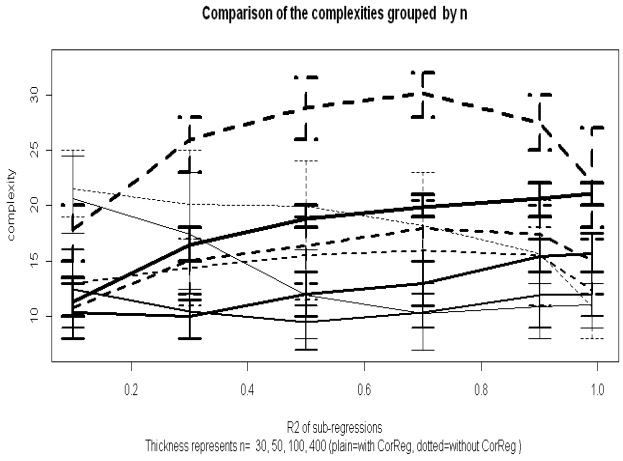

Figure 19: Comparison of the compexities between stepwise and CorReg+stepwise

4.3.2 Y depends only on covariates in X_r (worst case for us)

We now try the method with a response depending only on variables in $\boldsymbol{X}_r$. The datasets used here were still those from 1. Depending only on $\boldsymbol{X}_r$ implies sparsity and impossibility to obtain the true model when using the true structure.

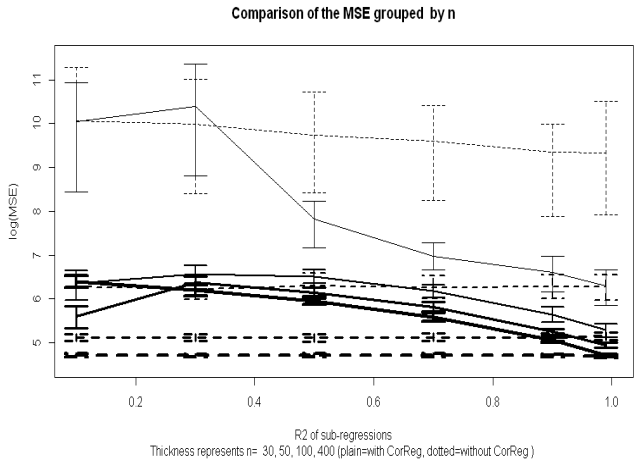

Figure 20: Comparison of the MSE between OLS and CorReg+OLS

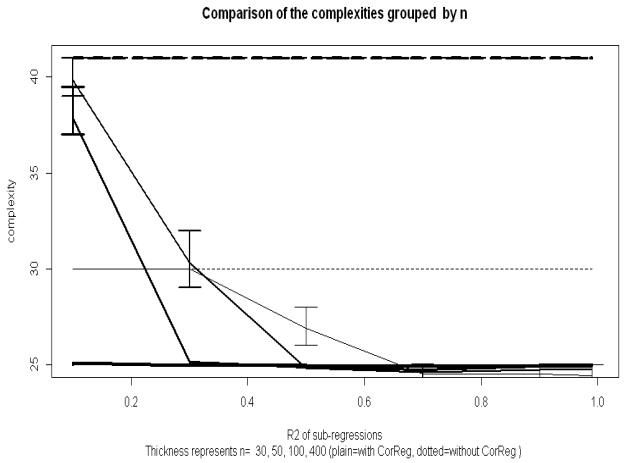

Figure 21: Comparison of the compexities between OLS and CorReg+OLS

CorReg is still better than OLS for strong correlations and limited values of n.

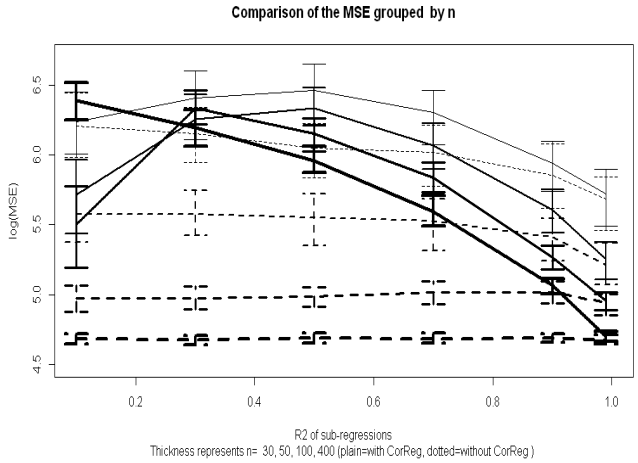

Figure 22: Comparison of the MSE between LASSO and CorReg+LASSO

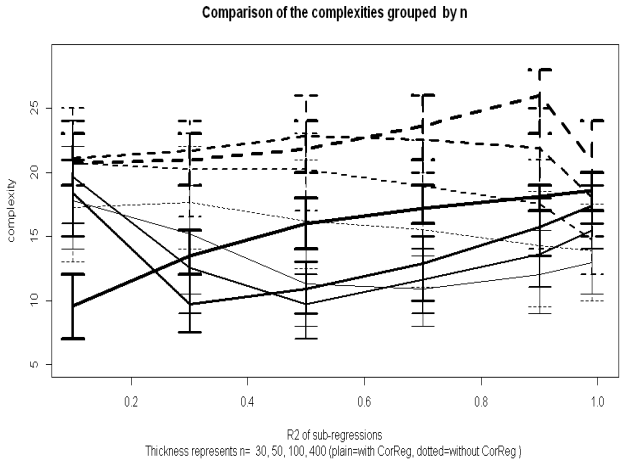

Figure 23: Comparison of the compexities between LASSO and CorReg+LASSO

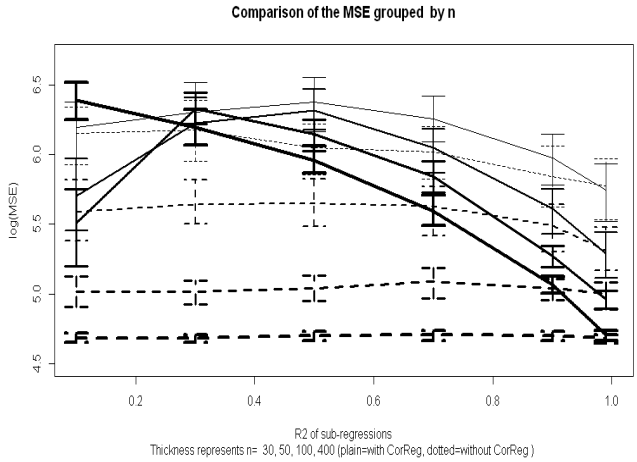

Figure 24: Comparison of the MSE between elasticnet and CorReg+elasticnet

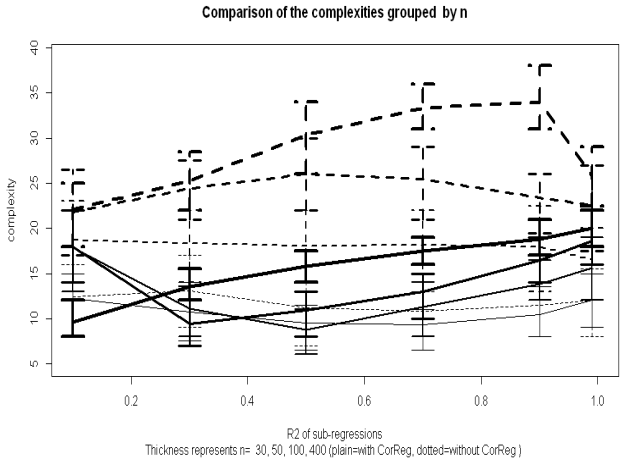

Figure 25: Comparison of the compexities between elasticnet and CorReg+elasticnet

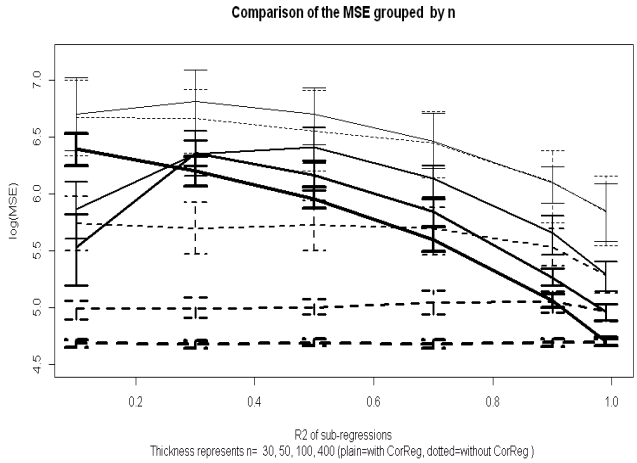

Figure 26: Comparison of the MSE between stepwise and CorReg+stepwise

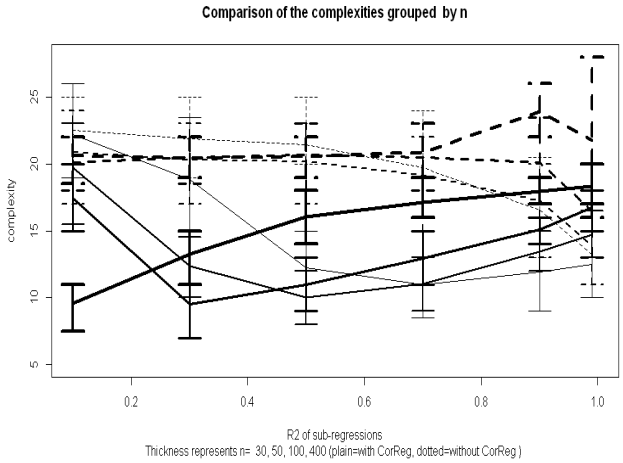

Figure 27: Comparison of the compexities between stepwise and CorReg+stepwise

4.4 Robustess of the model

We have generated a non-linear structure to test the robustess of the model. $\boldsymbol{X}_f$ is a set of 6 independent Gaussian mixtures defined as previously but with random signs for the components means. $\boldsymbol{X}_r = \boldsymbol{X}_7 = a\boldsymbol{X}_1^2 + \boldsymbol{X}_2 - 2\boldsymbol{X}_3 + \varepsilon$. The matrix $\boldsymbol{X}$ is then scaled and we get $\boldsymbol{Y} = \sum_{i=1}^{7} \boldsymbol{X}_i + \varepsilon_Y$. We let a vary between 0 and 10 to increase progressively the non-linear part of the sub-regression.

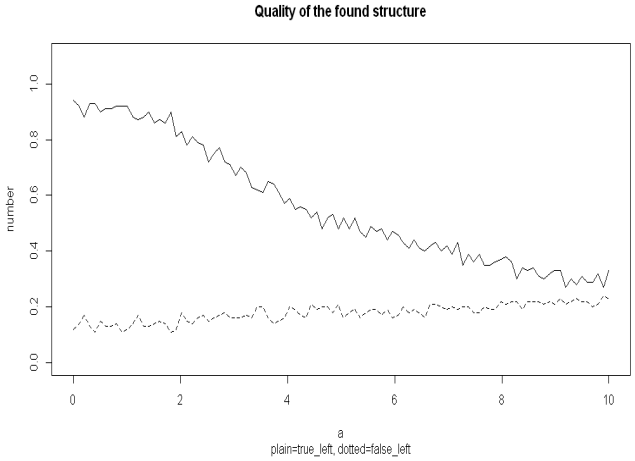

Figure 28: Quality of the structure found when the paramater a increases

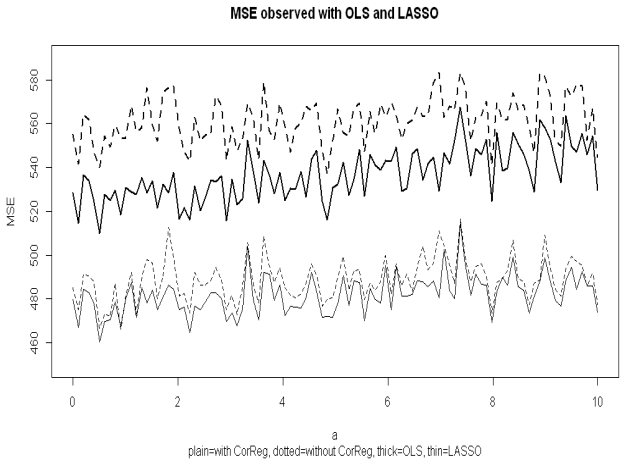

Figure 29: MSE on the main regression.

5 Numerical results on real datasets

5.1 Quality case study

This work takes place in steel industry context, with quality oriented objective : to understand and prevent quality problems on finished product, knowing the whole process. The correlations are strong here (many parameters of the whole process without any a priori and highly correlated because of physical laws, process rules, *etc.*).

We have :

- a quality parameter (confidential) as response variable,
- 205 variables from the whole process to explain it.
- The stakes : a hundred euros per ton (for information: Dunkerque's site aims to produce up to 7.5 millions tons a year)

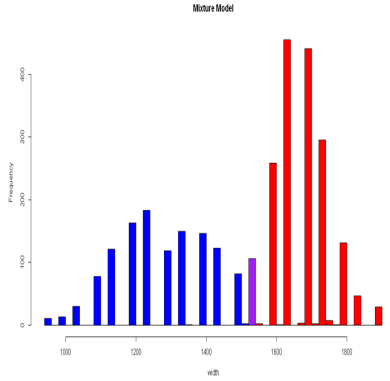

Figure 30: Example of non-Gaussian real variable easily modeled by a Gaussian mixture

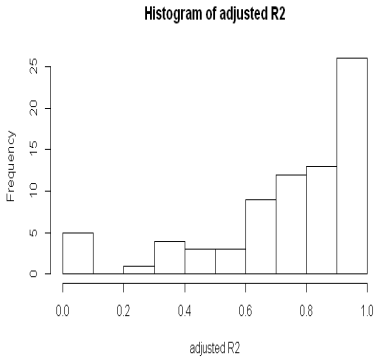

Figure 31: R^2_{adj} of the 76 sub-regressions.

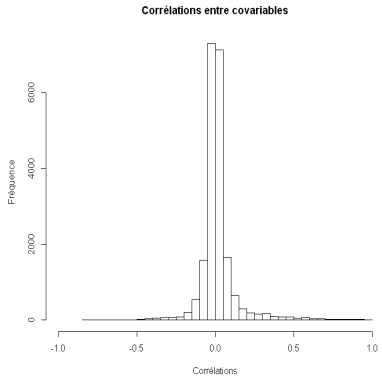

Figure 32: Histogram of correlations in $\boldsymbol{X}$.

We get a training set of $n = 3000$ products described by $p = 205$ variables from the industrial process and a validation sample of 847 products. Let's note ρ the absolute value of correlations between two covariates. Industrial variables are naturally highly correlated as the width and the weight of a steel slab ($\rho = 0.905$), the temperature before and after some tool ($\rho = 0.983$), the roughness of both faces of the product ($\rho = 0.919$), a mean and a max ($\rho = 0.911$). CorReg also found more complex structures describing physical models, like Width = f (Mean.flow , Mean.speed.CC) even if the true Physcial model is not linear : 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. The algorithm gives a structure of $p_r = 76$ subregressions with a mean of $\bar{\boldsymbol{p}}_f = 5.17$ regressors. In $\boldsymbol{X}_f$ the number of $\rho > 0.7$ is **79.33%** smaller than in $\boldsymbol{X}$.

It is now time to look at the predictive results (Figure 32). The best model found when not using CorReg is given by the LASSO. But when using CorReg elasticnet produces a better model in terms of prediction. LASSO gives a model with 21 non-zero coefficients and elasticnet with CorReg gives a model with 40 non-zero parameters but 6.40% better in prediction on the validation sample (847 products). 14 non-zero coefficients are common between the two models. Elasticnet alone get a model with 78 parameters that is improved by 9.75% in prediction when used with CorReg. When using LASSO with CorReg we obtain a model with 24 non-zero coefficients that is 4.11% better than LASSO alone. We also computed the OLS model (without selection) and the naive one (estimating the response by the mean of the learning set). All the MSE were modified here to obtain a value of 100 for the best (to preserve confidentiality). Elasticnet with CorReg is 13.51% better than OLS.

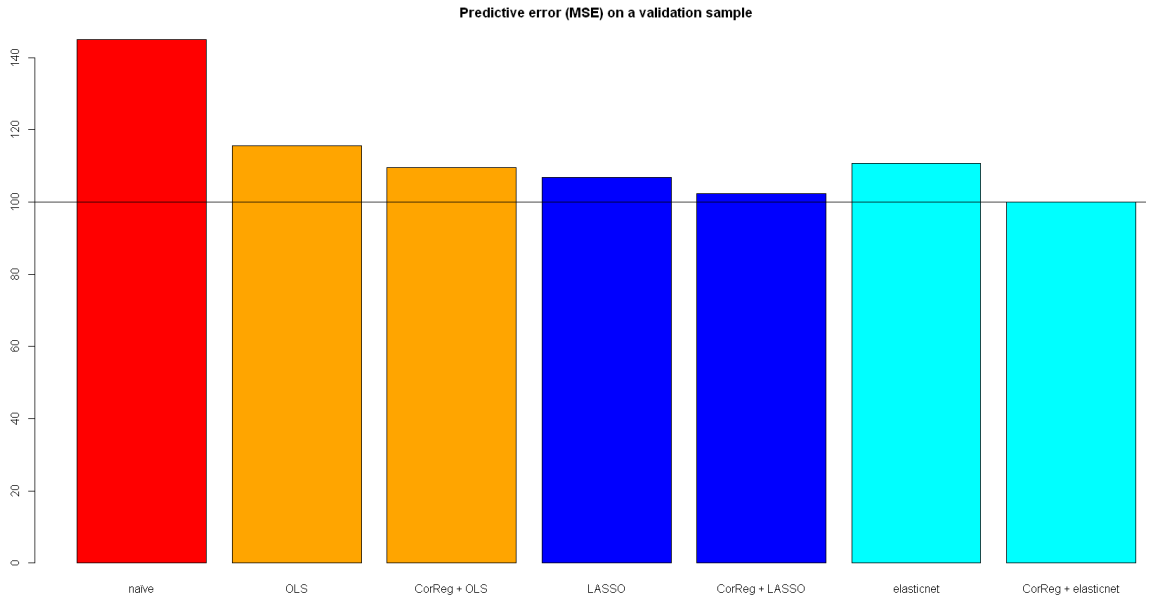

Figure 33: MSE comparison on industrial dataset. Learning set : 3 000 products, validation set : 847 products

Model	MSE	Complexity (with intercept)
OLS	115.63	206
CorReg + OLS	109.59	130
LASSO	106.84	21
CorReg + LASSO	102.45	24
elasticnet	110.81	78
CorReg + elasticnet	100	40

Table 2: Results obtained on a validation sample.

In terms of interpretation, the main regression comes with the family of regression so it gives a better understanding of the consequences of corrective actions on the whole process. It typically permits to determine the *tuning parameters* whereas LASSO would point variables we can't 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 don't even talk of the impact on the market shares that is even more important.

5.2 Production case study

This second example is about a phenomenon that impacts the productivity of a steel plan. We have :

- a (confidential) response variable,
- $p = 145$ variables from the whole process to explain it but only $n = 100$ individuals.
- The stakes : 20% of productivity to gain

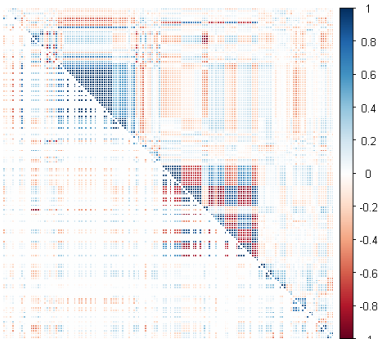

Figure 34: Correlations between the covariates in $\boldsymbol{X}$ (upper) and $\hat{\boldsymbol{X}}_f$ (lower).

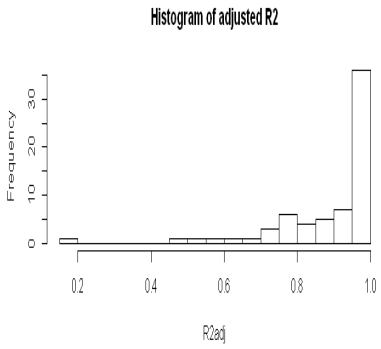

Figure 35: R^2_{adj} of the 67 sub-regressions.

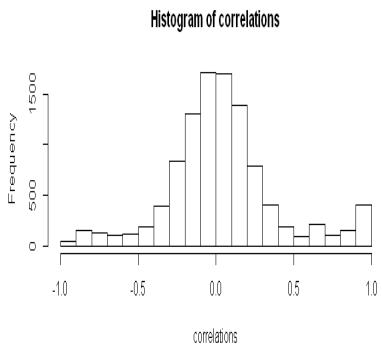

Figure 36: Histogram of correlations in $\boldsymbol{X}$.

Here $n < p$ so we only compare the leave-one-out cross-validation MSE. CorReg improves LASSO by 5.24% and elasticnet by 8.60%. CorReg combined with LASSO gives the best result but it is only a leave-on-out MSE. In this precise case, CorReg found a structure that helped to decorrelate

Model	MSE	Complexity (with intercept)
LASSO	105.54	34
CorReg + LASSO	100	18
elasticnet	129.94	13
CorReg + elasticnet	118.76	21

Table 3: Results obtained with leave-one out cross-validation. $n = 100, p = 145$.

covariates in interpretation and to find the relevant part of the process to optimize.

6 Conclusion and perspectives

We have seen that correlations can lead to serious estimation and variable selection problems in linear regression and that in such a context, it can be useful to explicitly model the structure between the covariates and to use this structure (even sequentially) to avoid correlations issues. We also show that real industrial context faces this kind of situations so our model can help to understand and predict physical phenomenon efficiently. But for now we still need a full dataset to learn the structure between the covariates and even if correlations are strong, some information is lost. Further work is needed to face these two challenges.

CorReg is accessible on CRAN and has already proved its efficiency on real regression problematics in industry. CorReg's strength is its great interpretability of the model, composed of several short linear regression easily managed by non-statisticians while strongly reducing correlations issues that are everywhere in industry. Nevertheless, we need to enlarge its application field to missing values, also very commons in industry. The structure can be used to estimate missing values in $\boldsymbol{X}_r$ but the actual generative model allows to go further (to manage missing values even in the MCMC algorithm) without supplementary hypothesis and this also is a strength of 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.

This paper only treats linear regression but such a pretreatment could be used for logistic regression, *etc.* So the subject is still wide opened.

7 Acknowledgements

We want to thanks ArcelorMittal Atlantique & Lorraine that has granted this work, given the chance to use CorReg on real dataset 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] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

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

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

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

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

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

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

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

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

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

[15] 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.

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

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

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

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

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

[21] G.A.F. Seber and A. J Lee. *Linear regression analysis*, volume 936. John Wiley & Sons, 2012.

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

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

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

[25] 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.

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

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

[28] 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.

8 Appendices

8.1 Identifiability of the structure

The model presented above relies on a discrete structure S between the covariates. But to find it we need identifiability property to insure the MCMC will asymptotically find the true model. Identifiability of the structure is asked in following terms: Is it possible to find another structure $\tilde{S}$ of linear regression between the covariates leading to the same joint distribution and marginal distributions?

If there are exact sub-regressions ($\sigma_j^2 = 0$), the structure won't be identifiable. But it only means that several structure will have the same likelihood and they will have the same interpretation. So it's not really a problem. Moreover, when an exact sub-regression is found, we can delete one of the implied variables without any loss of information and the structure will define a list of variable from which to delete. CorReg (Our R package) prints a warning to point out exact regressions when found. In the followings we suppose $\sigma_j^2 \neq 0$, then $\boldsymbol{X}'_f\boldsymbol{X}_f$ and $\boldsymbol{X}'\boldsymbol{X}$ are of full rank (but the later is ill-conditioned for small values of σ_j^2).

Our full generative model is a p-sized Gaussian mixture model of K distinct components and can be seen as a **SR** model defined by Maugis [15]. In this section, S will denote the set of variable as in the paper from Maugis and we call Gaussian mixtures the Gaussian mixtures with at least two distinct components. The equivalence with Maugis's model is defined by: $\boldsymbol{X}_r = \boldsymbol{y}^{S^c}$ and $\boldsymbol{X}_f = \boldsymbol{y}^R$. We have supposed independence between variables in $\boldsymbol{X}_f$ so the identifiability theorem from Maugis tells that

our model is identifiable if variables in $\boldsymbol{X}_f$ are Gaussian mixtures (what we supposed in section 3.1).

We define $\boldsymbol{X}^G \subsetneq \boldsymbol{X}_f$ containing Gaussian variables and we note the Gaussian mixtures $\boldsymbol{X}^{G^c} \neq \emptyset$ its complement in $\boldsymbol{X}_f$. We suppose that variables in $\boldsymbol{X}_r$ are all Gaussian mixtures. It implies that $\forall j \in I_r, \exists i \in I_f^j$ so that $\boldsymbol{X}^i \subset \boldsymbol{X}^{G^c}$ since any linear combination of Gaussian variable would only give a Gaussian (so each sub-regression contain at least one Gaussian mixture as a regressor).
We introduce the matricial notation $\boldsymbol{X}_r = \boldsymbol{X}_f\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\alpha}$ is the $(p-p_r)\times p_r$ matrix whose columns are the $\boldsymbol{\alpha}_j$ and $\boldsymbol{\varepsilon}$ is the $n \times p_r$ matrix whose columns are the $\boldsymbol{\varepsilon}_j$

The theorem from Maugis guarantee that a sub-regression between Gaussian mixtures is identifiable in terms of which one is regressed by others.

$$\boldsymbol{X}_{r|\boldsymbol{X}^G,\boldsymbol{X}^{G^c}} = \boldsymbol{X}^G\boldsymbol{\alpha}_G + \boldsymbol{X}^{G^c}\boldsymbol{\alpha}_{G^c} + \boldsymbol{\varepsilon} \tag{23}$$

$$\boldsymbol{X}_{r|\boldsymbol{X}^{G^c}} = \boldsymbol{X}^{G^c}\boldsymbol{\alpha}_{G^c} + \tilde{\boldsymbol{\varepsilon}} \text{ is identifiable where} \tag{24}$$

$$\tilde{\boldsymbol{\varepsilon}}_j = \boldsymbol{X}^G\boldsymbol{\alpha}_j^G + \boldsymbol{\varepsilon}_j \text{ is Gaussian.} \tag{25}$$

Here, $\otimes$ and $\oplus$ denote respectively the kronecker product and sum.

$$\forall j \in I_r : \ \boldsymbol{X}^j_{|\boldsymbol{X}_f} \sim \mathcal{N}(\boldsymbol{X}_f\boldsymbol{\alpha}_j, \sigma_j^2\boldsymbol{I}_n) \tag{26}$$

$$\forall j \in I_f : \boldsymbol{X}^j_{|S} \sim \mathcal{GM}(\boldsymbol{\pi}_j; \boldsymbol{\mu}_j; \boldsymbol{\sigma}_j^2) \text{ with } \boldsymbol{\pi}_j, \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2 \text{ vectors of size } K_j \tag{27}$$

$$\text{And we obtain, } \forall j \in I_r : \ \boldsymbol{X}^j \sim \mathcal{GM}(\bigotimes_{i\in I_f^j} \boldsymbol{\pi}_i \ ; \bigoplus_{i\in I_f^j} \alpha_{i,j}\boldsymbol{\mu}_i \ ; \boldsymbol{\sigma}_j^2 + \bigoplus_{i\in I_f^j} \alpha_{i,j}^2\boldsymbol{\sigma}_i^2) \tag{28}$$

So when we compare (27) and (28) we see that the number of components in I_r variables differs when subregressions are of length > 1 (almost 2 predictors) with multiple-class predictors (so the kronecker product is effective).

The presence of two Gaussian mixtures on the right of the sub-regression gives the identifiability of it (even with also Gaussian variables) by the strict increase of the component number when combining Gaussian mixtures. So a sub-regression is identifiable if it has only Gaussian mixtures (even only one) on the right or if it has at least two Gaussian mixtures (even with additional Gaussian variables)

8.2 The CorReg package

8.2.1 Alternative neighbourhoods for the MCMC

We have here at each step $|\mathcal{V}_{S,j}| = p$ candidates but some other constraints can be added on the definition of $\mathcal{S}$ and will consequently modify the size of the neighbourhood (for example a maximum complexity for the internal regressions or the whole structure, a maximum number of internal regressions, *etc.*). CorReg allows to modify this neighbourhood to better fit users constraints. Relaxation (column-wise and row-wise) is optional but gives more stability to the number of feasible candidates at each step and allows to modify several parts of I_f in only one step when needed. Hence it improves efficiency by a significant reinforcement of the irreductibility of the Markov chain. Rejecting candidates instead of doing the relaxation steps will however reduce the number of evaluated candidates and thus accelerate the walk. So it can be used for a warming phase when n is great and time is missing.

The hierchical uniform hypothesis made above for $P(S)$ implies $p_r < \frac{p}{2}$ and $p_f^j < \frac{p}{2}$ so candidates may be rejected to satisfy this hypothesis. Stronger constraints on p_r and/or p_f^j can be given in CorReg if relevant.

If the algorithm did not have time to converge (stationnarity), it can be continued with a few step for which the neighboorhood would only contain smaller candidates (in terms of complexity). It is equivalent to ask for each element in I_f if the criterion $P(S|\boldsymbol{X})$ would be better without it. Thus it can be seen as a final cleaning step. But in fact, it's just continuing the MCMC with a reduced neighbourhood.