Model-based pretreatment for regression with 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 correlated covariates. However many modern datasets are expected to convey more and more highly correlated covariates. We propose to explicitly model the correlations by a family of linear regressions between the covariates. Marginalization allows then to obtain a parsimonious correlation-free regression model, easily understandable and from which it is then possible to perform standard linear estimation methods including variables selection procedures for instance. The structure of correlations is found with an MCMC algorithm. 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

When one wants to explain a phenomenon based on some covariates, the first statistical method tried frequently is the linear regression. It provides a predictive model with a good interpretability even for non-statistician and is simple to learn. Therefore, linear regression is used in nearly all the fields where statistics are made, from industry (ballistic models to calibrate the process) to sociology (predicting some numerical properties of a population). Linear regression is a very classic situation but sometimes it has to face an also classical problem: the variance of the estimators, even without bias from modelisation. In this paper, vectors and matrices are in bold letters to distinguish them from scalars. 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, $\boldsymbol{Y}$ 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 $\sigma_Y \in \mathcal{R}$. The $p \times 1$ vector $\boldsymbol{\beta}$ is the vector of the coefficients of the regression, that can be estimated with Ordinary Least Squares (OLS). Estimation of the parameters requires the inversion of a matrix which will be ill-conditioned or even singular if some covariates depend linearly from each other. Variance of $\hat{\boldsymbol{\beta}}$ increases based on two aspects :

- The dimension p (number of covariates) of the model: the more covariates you have the greater variance you get.
- The correlations within the covariates: strongly correlated covariates give bad-conditioning and increase variance of the estimators .

With the rise of informatic, datasets contains more and more covariates and thus more and more useless covariates. So dimension reduction becomes a necessity. Moreover, when you use more covariates, you increase the chance to have correlated ones. It will be illustrated with examples from steel industry in the last section of this paper (many parameters of the whole process without any a priori) highly correlated (physical laws, process rules, etc). In such a context, variance of the estimators can lead to arbitrary results or even no results at all. Prediction and interpretation are both strongly needed, depending on the context. For instance, in an industrial context interpretation could be favored to improve the process instead of only predict defects.

Because OLS is the minimum-variance unbiased estimator, penalized methods try to reduce the variance introducing some bias to improve the bias-variance trade-off and get better prediction. Moreover, real datasets imply many irrelevant variables so we have to use variable selection methods.

In the following we note classical norms: $\| \boldsymbol{\beta} \|_2^2 = \sum_{i=1}^{p} (\beta_i)^2$ and $\| \boldsymbol{\beta} \|_1 = \sum_{i=1}^{p} |\beta_i|$.

Ridge regression[13] proposes a biased estimator that can be written in terms of a parametric L_2 penalty:

$$\hat{\boldsymbol{\beta}} = \text{argmin} \left\{ \| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \|_2^2 \right\} \text{ subject to } \| \boldsymbol{\beta} \|_2^2 \leq k \tag{2}$$

But Ridge regression do not aim at selecting covariates because coefficients tends to 0 but don't reach 0. So it gives difficult interpretations for large values of p.

The Least Absolute Shrinkage and Selection Operator (LASSO [16]) consists in a shrinkage of the regression coefficients based on a λ parametric L_1 penalty to obtain zeros in $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = \text{argmin} \left\{ \| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \|_2^2 \right\} \text{ subject to } \| \boldsymbol{\beta} \|_1 \leq \lambda \tag{3}$$

The Least Angle Regression (LAR [4]) Algorithm offers a very efficient way to obtain the whole LASSO path and is very attractive. It requires only the same order of magnitude of computational effort as OLS applied to the full set of covariates. And it really selects covariates with coefficients set exactly to 0. But LASSO also faces consistency problems [22] when confronted with correlated covariates. Some recent variants of the LASSO does exist for the choice of the penalization coefficient like the adaptative LASSO [23] or the random LASSO [18].

Elastic net[24] is a method developed to be a compromise between Ridge regression and the LASSO. Elastic net can be written:

$$\hat{\boldsymbol{\beta}} = (1 + \lambda_2) \, \text{argmin} \left\{ \| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \|_2^2 \right\}, \text{ subject to } (1 - \alpha) \| \boldsymbol{\beta} \|_1 + \alpha \| \boldsymbol{\beta} \|_2^2 \leq t \text{ for some } t \tag{4}$$

where $\alpha = \frac{\lambda_2}{(\lambda_1 + \lambda_2)}$. 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 same predictions with a more parsimonious model.

Another way of reducing the dimension is to consider clusters of variables with the same coefficients, like the Octogonal Shrinkage and Clustering Algorithm for Regression (OSCAR [1]). The CLusterwise Effect REgression (CLERE [19]) describes the β_j no longer as fixed effect parameters but as unobserved independant random variables with grouped β_j following a Gaussian Mixture distribution. The idea is to hope that the model have a small number of groups of covariates and that the mixture will have few enough components to have a number of parameters to estimate significantly lower than p. In such a case, it improves interpretability and ability to yeld reliable prediction with a smaller variance on $\hat{\boldsymbol{\beta}}$.

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 [10] also relies on Gaussian mixture (the spike and the slab) hypothesis for the β_j and gives a subset of covariates (not grouped) on which to compute OLS but has no specific protection against correlations issues.

Where some reduce the dimension without looking at correlations and then have to suppose having no correlations between the remaining covariates, we propose to focus on the correlations, giving a marginal model with orthogonal covariates and an explicit structure between covariates. Our work is based on the assumption that if we know explicitly the source of correlations, we could use this knowledge to avoid the problem. In fact we search the greatest set of orthogonal covariates to keep the maximum information but with an orthogonality constraint. This can be viewed as a pretreatment on the dataset allowing to use then other tools for dimension reduction and estimation without suffering

from correlations. Practically speaking we use a system of linear regressions between the covariates to model the structure of the correlations. This structure is obtained by a MCMC algorithm optimizing the penalized likelihood of the joint law on the dataset $\boldsymbol{X}$.

This paper will first present the marginal regression model and its properties 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 the last part.

2 Model to select decorrelated covariates

2.1 A classical problem: correlations in regression

For a model defined by

$$\boldsymbol{Y}_{|\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5}$$

where $\boldsymbol{X}$ is the $n \times p$ matrix of the explicative variables, $\boldsymbol{Y}$ the response vector and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma_Y^2 \boldsymbol{I_n})$ with $\boldsymbol{I_n}$ the n-sized identity matrix, we have the following Ordinary Least Squares (OLS) estimator:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1} \boldsymbol{X'Y} \tag{6}$$

With variance matrix

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma_Y^2 \, (\boldsymbol{X'X})^{-1} \tag{7}$$

and without any bias. When correlations between covariates are strong, the matrix to invert is ill-conditioned and the variance explodes. Another problem is that matricial inversion requires $n \geq p$. When it is not the case, a frequently used alternative is the Moore-Penrose generalized inverse [11]. In the followings, matrices inverses when $n < p$ will refer to this generalized inverse. Thus OLS can obtain some results even with $n < p$ (see section 4).

2.2 Our proposal: modelisation of the correlations

Let now $\boldsymbol{X} \in \mathcal{R}^{n \times p}$ be a set of p correlated covariates. We propose to explicitly define a family of p_2 internal regressions between covariates with I_2 the set of indices of endogenous variables in $\boldsymbol{X}$ (explained ones) and $I_1 = \{I_1^1, \ldots, I_1^p\}$ the set of the sets of indices of exogenous covariates (explaining ones) with $\forall j \notin I_2, I_1^j = \emptyset$. Then we have an explicit structure $S = (I_1, I_2, p_1, p_2)$ where $\boldsymbol{p_1} = (p_1^1, \ldots, p_1^{p_2})$ is the vector of the number of covariates in each internal regression. Thus we have $p_2 = |I_2|$ and $p_1^j = |I_1^j|$ where $|.|$ represents the cardinal of an ensemble.

In the following, we note $\boldsymbol{X^j}$ the j^{th} column of a matrix $\boldsymbol{X}$. For lighter notation we define $\boldsymbol{X_2} = \boldsymbol{X^{I_2}}$ the matrix of the endogenous covariates and $\boldsymbol{X_1} = \boldsymbol{X} \setminus \boldsymbol{X_2}$ the matrix of the remaining exogenous covariates.

The family of p_2 regressions within correlated covariates in $\boldsymbol{X}$ is noted:

$$\boldsymbol{X_2}_{|\boldsymbol{X_1},S} \text{ defined by } \forall j \in I_2 : \boldsymbol{X^j}_{|\boldsymbol{X_1},S} = \boldsymbol{X_1}\boldsymbol{\alpha_j} + \boldsymbol{\varepsilon_j} \text{ with } \boldsymbol{\varepsilon_j} \sim (\boldsymbol{0}, \sigma_j^2 \boldsymbol{I_n}) \tag{8}$$

where $\boldsymbol{\delta_j} \in \mathcal{R}^{(p-p_2)}$ are the vectors of the regression coefficients between the covariates (containing some zeros according to I_1^j).

We make the hypothesis of the uncrossing rule $I_1 \cap I_2 = \emptyset$ *i.e.* endogenous variables don't explain other covariates, thus we have a partition $\boldsymbol{X} = [\boldsymbol{X_1}, \boldsymbol{X_2}]$.

The discrete structure S is identifiable because we can't permute some regressions in (8) and obtain the same joint distribution $P(\boldsymbol{X}, \boldsymbol{Y})$, the residuals $(\boldsymbol{\varepsilon_j})$ would not stay Gaussian (see the appendices).

The structure obtained gives a system of linear regression that can be viewed as a recursive Simultaneous Equation Model (SEM)[3]. Such a system is easy to interpret but estimation don't take advantage of the explicit structure [17] when the structure is straight forward (recursive SEM). Some methods take into account correlations but they only consider covariances between residuals SUR (Seemingly Unrelated Regression [20]) or endogenous variables like SPRING (Structured selection of Primordial Relationships IN the General linear model [2]).

Here we suppose independency between the residuals but in other cases it remains the possibility to use such methods to estimate the $\boldsymbol{\delta_j}$ and σ_j.

We make the choice to distinguish the response variable from the other endogenous variables (that are on the left of a regression). Thus we have one regression on our response variable and a system of sub-regressions (without the response variable). Then we consider correlations between the explicative covariates of the main regression, not between the residuals. The structure is supposed to be the source of the correlations and allows us to define a marginal regression model based on a reduced set of independent covariates without significant information loss. So we may obtain two kinds of zeros in the marginal model : coerced zeros due to correlations (redundant information) and estimated ones with classical variable selection methods applied on remaining variables. This two kinds of zero won't be interpreted in the same way and thus consistency issues don't mean interpretation issues any more. So we dodge the drawbacks of both grouping effect and variable selection.

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

Let $\boldsymbol{Y} \in \mathcal{R}^n$ be a response variable we want to explain with $\boldsymbol{X}$:

$$\boldsymbol{Y}_{|\boldsymbol{X_1},\boldsymbol{X_2},S} = \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{X_2}\boldsymbol{\beta_2} + \boldsymbol{\varepsilon_Y} \text{ with } \boldsymbol{\varepsilon_Y} \sim \mathcal{N}(\boldsymbol{0}, \sigma_Y^2 \boldsymbol{I_n}); \tag{9}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta_1}, \boldsymbol{\beta_2}) \in \mathcal{R}^p$ is the vector of the regression coefficients and $\boldsymbol{I_n}$ the identity matrix. We note that (9) and (8) also give by simple integration on $\boldsymbol{X_2}$ a marginal regression model on $\boldsymbol{Y}$ *depending only on uncorrelated covariates* $\boldsymbol{X_1}$:

$$\begin{aligned} \boldsymbol{Y}_{|\boldsymbol{X_1},S} &= \boldsymbol{X_1}(\boldsymbol{\beta_1} + \sum_{j \in I_2} \beta_j \boldsymbol{\alpha_j}) + \sum_{j \in I_2} \beta_j \boldsymbol{\varepsilon_j} + \boldsymbol{\varepsilon_Y} & (10) \\ &= \boldsymbol{X_1}\boldsymbol{\beta_1^*} + \boldsymbol{\varepsilon_Y^*} = \boldsymbol{X}\boldsymbol{\beta^*} + \boldsymbol{\varepsilon_Y^*} \text{ where } \boldsymbol{\beta^*} = (\boldsymbol{\beta_1^*}, \boldsymbol{\beta_2^*}) \text{ and } \boldsymbol{\beta_2^*} = \boldsymbol{0} & (11) \end{aligned}$$

We note that it simply is a linear regression on some of the original covariates so we only made a pretreatment on the dataset by setting some coefficients to 0 because of correlations. These 0 won't be interpreted as independence with the response variable but as redundant information due to correlations between the covariates. It is a variable pre-selection independent of the response $\boldsymbol{Y}$.

2.4 Strategy of use: pretreatment before classical selection methods

As a pretreatment, it allows usage of any method in a second time to estimate αlike LASSO or even stepwise [15].

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. Expert knowledge can even be added to the structure.

After selection and estimation we will obtain a model with *two kinds of zeros*: those from decorrelating step and those from selection step, with different meanings. Thus consistency issues don't mean interpretation issues any more. So we dodge the drawbacks of both grouping effect and variable selection.

Each equation here is very simple (only linear regressions). Moreover, the uncrossing constraint ($I_1 \cap I_2 = \emptyset$) guarantee to keep a simple structure easily interpretable (no cycles and no chain-effect) and straightforward readable.

The marginal regression model simply is

$$Y_{|X_1} = X_1\alpha_1 + \varepsilon_\alpha \tag{12}$$

So we have the OLS unbiased estimator of α:

$$\hat{\alpha}_1 = (X_1'X_1)^{-1}X_1'Y \text{ and } \hat{\alpha}_2 = 0 \tag{13}$$

We see in (10) that it gives an unbiased estimation of Y and α but in terms of β this estimator is biased:

$$\mathrm{E}[\hat{\alpha_1}|X_1] = \beta_1 + \sum_{j \in I_2} \beta_j \delta_j \text{ and } \mathrm{E}[\hat{\alpha_2}|X_1] = 0 \tag{14}$$

with variance:

$$\mathrm{Var}[\hat{\alpha}_1|X_1] = (\sigma_Y^2 + \sum_{j \in I_2} \sigma_j^2 \beta_j^2)(X_1'X_1)^{-1} \text{ and } \mathrm{Var}[\hat{\alpha}_2|X_1] = 0 \tag{15}$$

We see that the variance is reduced compared to OLS described in equation (7)(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.

2.5 Illustration of the tradeoff conveyed by the pretreatment

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

$$\begin{aligned} \mathrm{MSE}(\hat{\beta}|\beta, X) &= \mathrm{Bias}^2 + \mathrm{Tr}(\mathrm{Var}(\hat{\beta})) & (16) \\ &= 0 + \sigma_Y^2 \, \mathrm{Tr}((X'X)^{-1}) & (17) \end{aligned}$$

To better illustrate the bias-variance tradeoff, we look at a simple example with $p = 3$ variables. X_1 is composed by two independent scaled Gaussian $\mathcal{N}(0, 1)$, $X_3 = x_1 + x_2 + \varepsilon_3$ where $\varepsilon_3 \sim \mathcal{N}(0, \sigma_3^2)$. We also have $\beta = (1, 1, 1)$ and $\sigma_Y = 10$. Then we observe the theoretical Mean Squared Error (MSE) of the estimator of both OLS and CorReg model: $MSE(\hat{\beta}|\beta, X) = Bias^2 + \mathrm{Tr}(\mathrm{Var}(\hat{\beta}))$ for several values of σ_3 (strength of the sub-regression) and n. Figures 1 to 3 show the theoretical MSE evolution with the strength of the sub-regression:

$$1 - \mathcal{R}^2 = \frac{\mathrm{Var}(\varepsilon_3)}{\mathrm{Var}(x_3)} = \frac{\sigma_3^2}{\sigma_3^2 + 2} \tag{18}$$

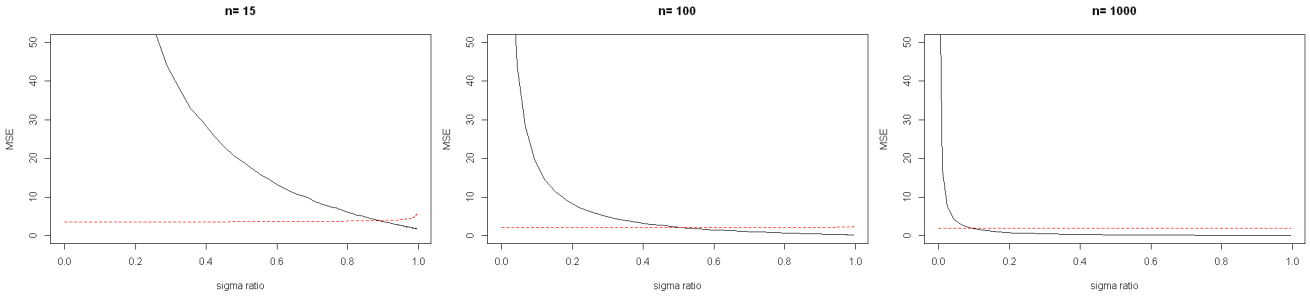

Figure 1: For $n = 15$. Dotted: CorReg, plain: OLS

Figure 2: For $n = 100$. Dotted: CorReg, plain: OLS

Figure 3: For $n = 1000$. Dotted: CorReg, plain: OLS.

It is clear in Figures 1 to 3 that the marginal model is more robust than OLS on X. And when sub-regression get weaker ($1 - \mathcal{R}^2$ tends to 1) it does not explode 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. $MSE_{OLS} = 0 + \mathrm{Tr}(\mathrm{Var}(\hat{\beta})) = \mathrm{Tr}(\mathrm{Var}(\hat{\beta}_1)) + \mathrm{Tr}(\mathrm{Var}(\hat{\beta}_2))$ where both terms are multiplied by σ_Y^2 and so is $\mathrm{Tr}(\mathrm{Var}(\hat{\alpha}_1))$. Thus, as $\mathrm{Var}(\hat{\alpha}_2) = 0$ when σ_Y^2 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).

3 Sub-regressions model selection

3.1 Modeling the uncorrelated covariates: a full generative approach

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 [6] offers a method to obtain a structure based

on the precision matrix (inverse of the variance-covariance matrix). It consists in a selection in the precision matrix, setting some covariances to zero. But the resulting matrix is symmetric and we need an oriented structure for our SEM. So we developped an MCMC algorithm to find the structure (R package CorReg on CRAN). The structure is based on a full generative model on $\boldsymbol{X}$ to be able to compare different size structures with a likelihood-based criterion in the MCMC.

To be able to compare structures with probabilistic criterions, we need a full generative model. We suppose that variables in X_1 follow Gaussian mixtures.

$$\forall j \notin I_2 : X^j_{|S} \sim f(\theta_j) = \mathcal{GM}(\pi_j; \mu_j; \sigma_j^2) \text{ with } \pi_j, \mu_j, \sigma_j^2 \text{ vectors of size } k_j; \tag{19}$$

The great flexibility [14] of such models makes our model more robust but one can use other laws if needed. Gaussian case is just a special case ($k_j = 1$) of Gaussian mixture so it is included in our hypothesis.

Variables in X_1 are supposed to be independent. Thus if one have some hypothesis on the distribution of some variables (exponentially distributed for example) it is possible to use it without impacting the model in other ways. We now have a full generative model.

3.2 Penalization of the integrated likelihood

To obtain the marginal regression model we need S, the linear structure between the covariates. Our full generative generative model allows us to compare structures with criterions like the Bayesian Information Criterion (BIC) which penalize the log-likelihood according to the complexity of the structure [12]. We will prefer this kind of comparison criterion instead of cross-validation that is very time-consuming and thus not friendly with combinatory problematics. We note Θ the set of the parameters of the generative model

$$\begin{aligned} P(S|X) &\propto P(X|S)P(S) && (20)\\ -2\log P(X|S) &\approx BIC = -2\mathcal{L}(X, S, \Theta) + |\Theta|\log(n) && (21)\\ & && (22) \end{aligned}$$

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 with a hierarchical uniform *a priori* law $P(S) = P(I_1|p_1, I_2, p_2)P(p_1|I_2, p_2)P(I_2|p_2)P(p_2)$ instead of a simple uniform law on S. It increases penalty on complexity for $p_2 < \frac{p}{2}$ and $p_1^j < \frac{p}{2}$. Hence this constraint on $\hat{p}_2$ and $\hat{p}_1^j$ is made in the MCMC in the followings. But we can imagine to use other criterions, like the RIC (Risk Inflation Criterion [5]) that choose a penalty in $\log p$ instead of $\log n$ and thus gives more parsimonious models when p is larger than n (high dimension) or any other criterion [7] thought to be better in a given context.
We will now use the following notation: $\psi(S) = \psi(X|S)$ where $\psi(X|S)$ is the chosen criterion, that will be BIC with hierarchical uniform hypothesis in our numerical results. We do not change BIC but only $P(S)$ so the properties are the same as classical BIC but we will obtain better results when our hypothesis is verified.

3.3 MCMC algorithm

3.3.1 The neighbourhood

Let's define $\mathcal{S}$ the ensemble of feasible structures (those with $I_1 \cap I_2 = \emptyset$).
For each step, starting from $S \in \mathcal{S}$ we define a neighbourhood:

$$\begin{aligned} \mathcal{V}_{S,j} &= \{S\} \cup \{S^{(i,j)} | 1 \leq i \leq p, i \neq j\} && (23)\\ \text{where} \quad j &\sim \mathcal{U}(\{1, \ldots, p\}) && (24) \end{aligned}$$

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

- if $i \notin I_i^j$ (add):
  - $I_1^j = I_1^j \cup \{i\}$

- $I_1^i = \emptyset$ (explicative variables can't depend on others : column-wise relaxation)
- $I_1 = I_1 \setminus \{j\}$ (dependent variables can't explain others : row-wise relaxation)

- else (remove): $I_1^j = I_1^j \setminus \{i\}$

At every moment, coherence between I_1 and others parts of S can be done by $\forall 1 \leq j \leq p : p_1^j = |I_1^j|$, $I_2 = \{j | p_1^j > 0\}$, $p_2 = |I_2|$, .

3.3.2 Transition probabilities

We first make the approximation

$$P(S|X) \approx exp(\psi(S)). \tag{25}$$

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

$$\forall (S, \tilde{S}) \in \mathcal{S}^2 : \mathcal{P}(S, \tilde{S}) \quad = \quad \sum_{j=1}^{p} \mathbf{1}_{\{\tilde{S} \in \mathcal{V}_{S,j}\}} \frac{\exp(-\frac{1}{2}\psi(\tilde{S}))}{\sum_{S_l \in \mathcal{V}_{S,j}} \exp(-\frac{1}{2}\psi(S_l))} \tag{26}$$

$$\tag{27}$$

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 [9] and the output will be the best structure in terms of ψ 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 10)).

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_1, 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[6] 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 [8]. It also helps to face local extrema. In the followings, the chain was launched with twenty initialisations.

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 4.3.1 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, $p = 40$, $\sigma_Y = 10$, $\sigma = 0.001$, variables X_1 follow Gaussian mixture models of $\lambda = 5$ classes which means follow Poisson's law of parameter λ and which standard deviation also is λ. The δ_j and β_j are generated according to the same Poisson law but with a random sign. S only contains binary relationships but CorReg was only constrained to $\max(\hat{p}_1^j) = 5$. We used Rmixmod to estimate the densities of each covariate. For each configuration, the walk was launched on 20 initial structures with a maximum of 9 000 steps each time.

4.2 Finding the structure

4.2.1 How to evaluate found structure?

The first criterion is ψ which is minimised in the MCMC. But in our case, ψ is based on the likelihood 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 $TL = |I_2 \cap \hat{I}_2|$
- WL (Wrong left) : the number of found dependent variables that are not dependent $WL = |\hat{I}_2| - TL$
- ML (Missing left) : the number of really dependent variables not found $ML = |I_2| - TL$
- Δp_2 : the gap between the number of sub-regression in both model : $\Delta p_2 = |I_2| - |\hat{I}_2|$. The sign defines if $\hat{S}$ is too complex or too simple
- $\Delta compl$: the difference in complexity between both model : $\Delta compl = \sum_{j \in p_2} p_1^j - \sum_{j \in \hat{p}_2} \hat{p}_1^j$

4.2.2 Results on S

In table 1 we compare found structures in different contexts with both Uniform (U) and Hierarchical Uniform (HU) a priori law on $P(S)$. We see that (HU) hypothesis gives sparser models even if $\max(\hat{p}_1^j) = 5$. $p = 40$ so the maximum value for $\hat{p_2}$ is 20 in (HU) case and we see that this max is not reached. The stronger penalty implied by (HU) really is efficient. The datasets used for (U) and (HU) are the same to keep the comparison meaningful. These datasets and $\hat{S}$ are those used for tables 2 to 9.

It is also notable that (HU) has a greater computational cost than (U). But next versions of the package may optimize a bit more the penalization step and reduce this gap. We also notice that the MCMC is faster when there are numerous correlations (rejecting more candidates).

n	p_2	ψ	Time Mixmod	Time MCMC	TL	WL	ML	Δp_2	$\Delta compl$
30	0	U	0.4104	3.2428	0	5.43	0	-5.43	22.55
			(0.0275)	(0.3711)	(0)	(1.9346)	(0)	(1.9346)	(8.0884)
		HU	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	U	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)
		HU	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	U	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)
		HU	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	U	0.5229	4.7068	0	4.2	0	-4.2	13.35
			(0.0519)	(0.5865)	(0)	(1.7233)	(0)	(1.7233)	(5.6468)
		HU	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	U	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)
		HU	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	U	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)
		HU	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	U	0.9623	12.9373	0	2.83	0	-2.83	6.23
			(0.077)	(1.7778)	(0)	(1.2953)	(0)	(1.2953)	(3.1999)
		HU	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	U	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)
		HU	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	U	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)
		HU	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 with constraint $\hat{p}_1 \leq 5$. Mean observed and standard deviation (sd).

4.3 Results on prediction

4.3.1 Y depends on all variables in X

We first 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 1.

We observe that CorReg is better than classical methods especially with the Hierarchical Uniform law. When the complexity of the true model is higher than $\frac{p}{2}$ Uniform hypothesis logically is better but (HU) still beats classical methods. We also observe in table 5 that simple methods like stepwise (here from the package lars) can give good results in prediction.

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 [21]).

n	p_2	ψ	indicator	OLS	CorReg $\hat{S}$	CorReg S
30	0	U	MSE (sd)	262627.57 (732019)	5928332.4 (49005690.2)	262627.57 (732019)
			cpl (sd)	30 (0)	29.99 (0.1)	30 (0)
		HU	MSE (sd)	262627.57 (732019)	10381962.64 (90962496.9)	262627.57 (732019)
			cpl (sd)	30 (0)	30 (0)	30 (0)
30	16	U	MSE (sd)	510747.53 (2287539.8)	635.42 (335.2)	610.32 (424.9)
			cpl (sd)	30 (0)	24.11 (1)	25 (0)
		HU	MSE (sd)	510747.53 (2287539.8)	603.24 (415.6)	610.32 (424.9)
			cpl (sd)	30 (0)	24.82 (0.6)	25 (0)
30	32	U	MSE (sd)	178323.95 (1426610.2)	180.02 (44.9)	141.03 (27.8)
			cpl (sd)	30 (0)	13.85 (0.9)	9 (0)
		HU	MSE (sd)	178323.95 (1426610.2)	330.31 (134.8)	141.03 (27.8)
			cpl (sd)	30 (0)	21 (0)	9 (0)
50	0	U	MSE (sd)	528.08 (228.6)	886.54 (364.3)	528.08 (228.6)
			cpl (sd)	41 (0)	36.8 (1.7)	41 (0)
		HU	MSE (sd)	528.08 (228.6)	542.29 (235)	528.08 (228.6)
			cpl (sd)	41 (0)	40.87 (0.4)	41 (0)
50	16	U	MSE (sd)	612.72 (291.5)	239.17 (89.4)	200.32 (42.6)
			cpl (sd)	41 (0)	24.43 (0.7)	25 (0)
		HU	MSE (sd)	612.72 (291.5)	207.6 (51.3)	200.32 (42.6)
			cpl (sd)	41 (0)	24.99 (0.5)	25 (0)
50	32	U	MSE (sd)	555.44 (262.5)	128.98 (18.1)	121.08 (11.9)
			cpl (sd)	41 (0)	11.45 (0.9)	9 (0)
		HU	MSE (sd)	555.44 (262.5)	171.9 (31.1)	121.08 (11.9)
			cpl (sd)	41 (0)	21 (0)	9 (0)
100	0	U	MSE (sd)	167.71 (20.6)	323.44 (124.1)	167.71 (20.6)
			cpl (sd)	41 (0)	38.17 (1.3)	41 (0)
		HU	MSE (sd)	167.71 (20.6)	167.98 (20.8)	167.71 (20.6)
			cpl (sd)	41 (0)	40.99 (0.1)	41 (0)
100	16	U	MSE (sd)	168.68 (22.4)	158.51 (51.6)	133.49 (12.2)
			cpl (sd)	41 (0)	24.53 (0.7)	25 (0)
		HU	MSE (sd)	168.68 (22.4)	137.03 (20.3)	133.49 (12.2)
			cpl (sd)	41 (0)	25.01 (0.4)	25 (0)
100	32	U	MSE (sd)	173.25 (22.9)	112.8 (10.7)	110.63 (6.9)
			cpl (sd)	41 (0)	10.02 (0.8)	9 (0)
		HU	MSE (sd)	173.25 (22.9)	127.4 (11.9)	110.63 (6.9)
			cpl (sd)	41 (0)	21 (0)	9 (0)

Table 2: OLS and OLS combined with constrained CorReg.Y depends on all variables in X. CorReg logically wins. When $n < p$ the dataset was reduced to $p = n$ automatically by a QR decomposition as the lm function of R does. Without selection, all models have $\min(n, p)$ non-zero coefficients.

n	p_2	ψ	indicator	LASSO	CorReg $\hat{S}$	CorReg S
30	0	U	MSE (sd)	1246.35 (350.5)	1433.98 (526.7)	1246.35 (350.5)
			cpl (sd)	17.84 (5.5)	15.79 (5.6)	17.84 (5.5)
		HU	MSE (sd)	1246.35 (350.5)	1248.84 (341.4)	1246.35 (350.5)
			cpl (sd)	17.84 (5.5)	17.63 (5.6)	17.84 (5.5)
30	16	U	MSE (sd)	712.79 (405.2)	566.32 (226.4)	554.01 (257.1)
			cpl (sd)	16.68 (4.4)	15.35 (4.3)	16.14 (4.5)
		HU	MSE (sd)	712.79 (405.2)	551.92 (242.4)	554.01 (257.1)
			cpl (sd)	16.68 (4.4)	15.96 (4.4)	16.14 (4.5)
30	32	U	MSE (sd)	216.12 (128.3)	157.78 (39.4)	147.04 (28.4)
			cpl (sd)	11.23 (4.4)	8.56 (1.8)	7.91 (1.3)
		HU	MSE (sd)	216.12 (128.3)	178.88 (68.9)	147.04 (28.4)
			cpl (sd)	11.23 (4.4)	9.86 (3)	7.91 (1.3)
50	0	U	MSE (sd)	658.38 (221.4)	872.32 (239.4)	658.38 (221.4)
			cpl (sd)	28.2 (6)	22.95 (5.8)	28.2 (6)
		HU	MSE (sd)	658.38 (221.4)	652.72 (211.9)	658.38 (221.4)
			cpl (sd)	28.2 (6)	28.36 (5.9)	28.2 (6)
50	16	U	MSE (sd)	274.34 (101.7)	268.76 (106.3)	225.64 (67.5)
			cpl (sd)	22.14 (4.1)	19.52 (2.9)	20.47 (2.8)
		HU	MSE (sd)	274.34 (101.7)	233.8 (78.9)	225.64 (67.5)
			cpl (sd)	22.14 (4.1)	20.19 (2.8)	20.47 (2.8)
50	32	U	MSE (sd)	165.6 (73.7)	128.34 (18.8)	123.73 (12.8)
			cpl (sd)	12.93 (6.4)	8.51 (1.5)	8.09 (1.1)
		HU	MSE (sd)	165.6 (73.7)	135.84 (24.1)	123.73 (12.8)
			cpl (sd)	12.93 (6.4)	10.17 (3.1)	8.09 (1.1)
100	0	U	MSE (sd)	183.33 (31.1)	357.15 (133.4)	183.33 (31.1)
			cpl (sd)	37.78 (2.4)	32.93 (3.7)	37.78 (2.4)
		HU	MSE (sd)	183.33 (31.1)	183.78 (31.6)	183.33 (31.1)
			cpl (sd)	37.78 (2.4)	37.75 (2.4)	37.78 (2.4)
100	16	U	MSE (sd)	148.83 (18.6)	164.44 (54.3)	139.57 (16.1)
			cpl (sd)	25.22 (4)	21.72 (2)	22.3 (1.7)
		HU	MSE (sd)	148.83 (18.6)	142.97 (22.1)	139.57 (16.1)
			cpl (sd)	25.22 (4)	22.24 (1.8)	22.3 (1.7)
100	32	U	MSE (sd)	124.29 (19.9)	113.36 (11)	111.54 (7.2)
			cpl (sd)	13.18 (5.9)	8.68 (1.2)	8.53 (1)
		HU	MSE (sd)	124.29 (19.9)	118.33 (12)	111.54 (7.2)
			cpl (sd)	13.18 (5.9)	11.02 (2.9)	8.53 (1)

Table 3: LASSO (with LAR) combined with constrained CorReg.Y depends on all variables in X. CorReg logically wins. LASSO is better than OLS but is improved by CorReg even for large values of n.

n	p_2	ψ	indicator	Elasticnet	CorReg $\hat{S}$	CorReg S
30	0	U	MSE (sd)	1 326.54 (388.4)	1 356.98 (331.4)	1 326.54 (388.4)
			cpl (sd)	12.14 (5.1)	12.5 (5.2)	12.14 (5.1)
		HU	MSE (sd)	1326.54 (388.4)	1307.96 (356.6)	1326.54 (388.4)
			cpl (sd)	12.14 (5.1)	12.27 (5.1)	12.14 (5.1)
30	16	U	MSE (sd)	1 400.56 (1598.2)	668.57 (274.6)	653.59 (283.7)
			cpl (sd)	13.86 (6.8)	14.31 (5.7)	14.83 (5.3)
		HU	MSE (sd)	1 400.56 (1598.2)	643.25 (277.9)	653.59 (283.7)
			cpl (sd)	13.86 (6.8)	14.83 (5.6)	14.83 (5.3)
30	32	U	MSE (sd)	855.74 (582.6)	181.08 (51.9)	146.79 (32.7)
			cpl (sd)	15.57 (6.3)	11.19 (2.4)	8.37 (1.3)
		HU	MSE (sd)	855.74 (582.6)	311.57 (163.9)	146.79 (32.7)
			cpl (sd)	15.57 (6.3)	14.87 (3.9)	8.37 (1.3)
50	0	U	MSE (sd)	738.94 (254.5)	914.56 (283.7)	738.94 (254.5)
			cpl (sd)	25.85 (8.7)	20.97 (7.8)	25.85 (8.7)
		HU	MSE (sd)	738.94 (254.5)	751.06 (262.2)	738.94 (254.5)
			cpl (sd)	25.85 (8.7)	25.43 (8.8)	25.85 (8.7)
50	16	U	MSE (sd)	516.36 (226.6)	276.6 (133.6)	228.85 (89.3)
			cpl (sd)	26.72 (7.4)	20.49 (3.8)	21.8 (3.3)
		HU	MSE (sd)	516.36 (226.6)	239.77 (103.7)	228.85 (89.3)
			cpl (sd)	26.72 (7.4)	21.57 (3.4)	21.8 (3.3)
50	32	U	MSE (sd)	328.7 (146.3)	130.61 (20.6)	124.01 (15.4)
			cpl (sd)	24.7 (7.1)	9.95 (1.5)	8.22 (1.2)
		HU	MSE (sd)	328.7 (146.3)	169.13 (40.2)	124.01 (15.4)
			cpl (sd)	24.7 (7.1)	16.04 (3.8)	8.22 (1.2)
100	0	U	MSE (sd)	175.93 (25.9)	355.79 (143.8)	175.93 (25.9)
			cpl (sd)	39.71 (1.7)	34.11 (4.6)	39.71 (1.7)
		HU	MSE (sd)	175.93 (25.9)	176.2 (26.3)	175.93 (25.9)
			cpl (sd)	39.71 (1.7)	39.69 (1.7)	39.71 (1.7)
100	16	U	MSE (sd)	173.65 (25.1)	164.96 (59)	139.51 (17.6)
			cpl (sd)	35.54 (4.2)	22.58 (2.2)	23.26 (1.9)
		HU	MSE (sd)	173.65 (25.1)	143.04 (23.2)	139.51 (17.6)
			cpl (sd)	35.54 (4.2)	23.16 (1.9)	23.26 (1.9)
100	32	U	MSE (sd)	161.92 (25.2)	113.61 (10.9)	111.46 (7.2)
			cpl (sd)	30.78 (6.4)	9.25 (1.2)	8.67 (1)
		HU	MSE (sd)	161.92 (25.2)	127.44 (13.3)	111.46 (7.2)
			cpl (sd)	30.78 (6.4)	17.85 (2.7)	8.67 (1)

Table 4: Elasticnet combined with constrained CorReg.Y depends on all variables in X. CorReg logically wins. LASSO was better.

n	p_2	ψ	indicator	Stepwise	CorReg $\hat{S}$	CorReg S
30	0	U	MSE (sd)	1 919.43 (861.8)	2 014.13 (683.6)	1 919.43 (861.8)
			cpl (sd)	22.8 (3.3)	19.34 (4.9)	22.8 (3.3)
		HU	MSE (sd)	1 919.43 (861.8)	1 885.39 (814)	1 919.43 (861.8)
			cpl (sd)	22.8 (3.3)	22.73 (3.1)	22.8 (3.3)
30	16	U	MSE (sd)	696.46 (492.9)	662.07 (285.9)	660.93 (354)
			cpl (sd)	15.9 (3.8)	15.1 (3.7)	15.76 (3.7)
		HU	MSE (sd)	696.46 (492.9)	655.94 (325.7)	660.93 (354)
			cpl (sd)	15.9 (3.8)	15.72 (3.6)	15.76 (3.7)
30	32	U	MSE (sd)	394.93 (356)	155.18 (36.4)	147.81 (30.3)
			cpl (sd)	16.1 (6.3)	8.26 (1.7)	7.77 (1.3)
		HU	MSE (sd)	394.93 (356)	189.49 (79)	147.81 (30.3)
			cpl (sd)	16.1 (6.3)	10.23 (3)	7.77 (1.3)
50	0	U	MSE (sd)	745.88 (241.7)	962.12 (319.3)	745.88 (241.7)
			cpl (sd)	27.19 (4.6)	22.9 (4.5)	27.19 (4.6)
		HU	MSE (sd)	745.88 (241.7)	749.08 (244.2)	745.88 (241.7)
			cpl (sd)	27.19 (4.6)	27.21 (4.5)	27.19 (4.6)
50	16	U	MSE (sd)	279.82 (151.1)	290.13 (117.3)	239.61 (69.5)
			cpl (sd)	21.24 (5.2)	18.13 (2.8)	19.23 (2.7)
		HU	MSE (sd)	279.82 (151.1)	252.35 (108.5)	239.61 (69.5)
			cpl (sd)	21.24 (5.2)	18.97 (2.7)	19.23 (2.7)
50	32	U	MSE (sd)	199.3 (158.7)	126.81 (17.4)	124.19 (13)
			cpl (sd)	14.31 (6.2)	8.16 (1.2)	8.04 (1)
		HU	MSE (sd)	199.3 (158.7)	136.32 (25.2)	124.19 (13)
			cpl (sd)	14.31 (6.2)	9.59 (2.6)	8.04 (1)
100	0	U	MSE (sd)	193.26 (34.9)	387.18 (151.6)	193.26 (34.9)
			cpl (sd)	36.3 (2.5)	30.94 (4.2)	36.3 (2.5)
		HU	MSE (sd)	193.26 (34.9)	194.43 (37.3)	193.26 (34.9)
			cpl (sd)	36.3 (2.5)	36.22 (2.7)	36.3 (2.5)
100	16	U	MSE (sd)	145.46 (20.1)	165.74 (56.1)	140.27 (17.5)
			cpl (sd)	23.44 (3.8)	21.23 (2.1)	21.93 (1.7)
		HU	MSE (sd)	145.46 (20.1)	143.74 (23.2)	140.27 (17.5)
			cpl (sd)	23.44 (3.8)	21.86 (1.7)	21.93 (1.7)
100	32	U	MSE (sd)	128.44 (18.4)	113.27 (10.7)	111.77 (7.2)
			cpl (sd)	13.58 (5.2)	8.59 (1.1)	8.48 (1)
		HU	MSE (sd)	128.44 (18.4)	118.61 (11.5)	111.77 (7.2)
			cpl (sd)	13.58 (5.2)	10.75 (2.8)	8.48 (1)

Table 5: Stepwise combined with constrained CorReg. Y depends on all variables in X. CorReg logically wins but stepwise is quite good compared to elasticnet.

4.3.2 Y depends only on covariates in X_2 (worst case for us)

We now try the method with a response depending only on variables in X_2. The datasets used here were still those from 1. Depending only on X_2 imply sparsity and impossibility to obtain the true model when using the true structure. CorReg is still better than classical methods.

n	p_2	ψ	indicator	OLS	CorReg $\hat{S}$	CorReg S
30	16	U	MSE (sd)	81781.54 (247281.7)	507.1 (269.2)	593.03 (428.3)
			cpl (sd)	30 (0)	24.11 (1)	25 (0)
		HU	MSE (sd)	81781.54 (247281.7)	587.07 (425.6)	593.03 (428.3)
			cpl (sd)	30 (0)	24.82 (0.6)	25 (0)
30	32	U	MSE (sd)	74136.81 (297716.1)	189.45 (48.1)	144.51 (31.4)
			cpl (sd)	30 (0)	13.85 (0.9)	9 (0)
		HU	MSE (sd)	74136.81 (297716.1)	340.53 (169.4)	144.51 (31.4)
			cpl (sd)	30 (0)	21 (0)	9 (0)
50	16	U	MSE (sd)	684.04 (438.1)	190.97 (37.7)	197.32 (40)
			cpl (sd)	41 (0)	24.43 (0.7)	25 (0)
		HU	MSE (sd)	684.04 (438.1)	196.36 (40.5)	197.32 (40)
			cpl (sd)	41 (0)	24.99 (0.5)	25 (0)
50	32	U	MSE (sd)	596.32 (323)	126.94 (15.5)	119.39 (12.4)
			cpl (sd)	41 (0)	11.45 (0.9)	9 (0)
		HU	MSE (sd)	596.32 (323)	168.03 (27.7)	119.39 (12.4)
			cpl (sd)	41 (0)	21 (0)	9 (0)
100	16	U	MSE (sd)	168.35 (20.1)	133.75 (12.4)	135.08 (13.1)
			cpl (sd)	41 (0)	24.53 (0.7)	25 (0)
		HU	MSE (sd)	168.35 (20.1)	135.04 (13.1)	135.08 (13.1)
			cpl (sd)	41 (0)	25.01 (0.4)	25 (0)
100	32	U	MSE (sd)	168.07 (21.4)	109.61 (7.4)	108.64 (7.2)
			cpl (sd)	41 (0)	10.02 (0.8)	9 (0)
		HU	MSE (sd)	168.07 (21.4)	124.23 (11.1)	108.64 (7.2)
			cpl (sd)	41 (0)	21 (0)	9 (0)

Table 6: OLS and OLS combined with constrained CorReg.Y depends on all variables in X_2. Sometimes $\hat{S}$ gives better results than S because S is penalized by the fact that it relies only on covariates not in the true model.

n	p_2	ψ	indicator	LAR	CorReg $\hat{S}$	CorReg S
30	16	U	MSE (sd)	367.45 (198.4)	298.32 (122.4)	292.53 (102.2)
			cpl (sd)	14.37 (5.1)	12.9 (3.8)	12.84 (3.8)
		HU	MSE (sd)	367.45 (198.4)	298.1 (119.1)	292.53 (102.2)
			cpl (sd)	14.37 (5.1)	12.81 (3.8)	12.84 (3.8)
30	32	U	MSE (sd)	273.12 (267.2)	166.93 (49)	152.45 (39.5)
			cpl (sd)	11.7 (5.8)	8.7 (2.5)	7.45 (1.4)
		HU	MSE (sd)	273.12 (267.2)	175.68 (57.5)	152.45 (39.5)
			cpl (sd)	11.7 (5.8)	9.19 (2.9)	7.45 (1.4)
50	16	U	MSE (sd)	189.64 (52.5)	171.97 (45.9)	176.58 (46.8)
			cpl (sd)	14.15 (3.9)	13.4 (2.8)	13.5 (3.1)
		HU	MSE (sd)	189.64 (52.5)	174.36 (45.3)	176.58 (46.8)
			cpl (sd)	14.15 (3.9)	13.57 (3)	13.5 (3.1)
50	32	U	MSE (sd)	163.88 (67.4)	124.53 (17)	121.99 (16.6)
			cpl (sd)	13.06 (7.2)	8.29 (1.6)	7.73 (1.2)
		HU	MSE (sd)	163.88 (67.4)	133.7 (26.4)	121.99 (16.6)
			cpl (sd)	13.06 (7.2)	9.65 (3.3)	7.73 (1.2)
100	16	U	MSE (sd)	129.65 (14.8)	127.11 (12.1)	127.46 (12.2)
			cpl (sd)	14.86 (3.1)	13.97 (2.5)	14.04 (2.5)
		HU	MSE (sd)	129.65 (14.8)	127.46 (12.2)	127.46 (12.2)
			cpl (sd)	14.86 (3.1)	14.02 (2.5)	14.04 (2.5)
100	32	U	MSE (sd)	121.21 (20.1)	109.66 (7.9)	109.04 (7.8)
			cpl (sd)	12.84 (5.2)	8.42 (1.1)	8.11 (0.8)
		HU	MSE (sd)	121.21 (20.1)	112.56 (11.2)	109.04 (7.8)
			cpl (sd)	12.84 (5.2)	9.77 (2.3)	8.11 (0.8)

Table 7: LASSO (with LAR) combined with constrained CorReg.Y depends on all variables in X_2.

n	p_2	ψ	indicator	Elasticnet	CorReg $\hat{S}$	CorReg S
30	16	U	MSE (sd)	499.32 (218.8)	311.41 (137.4)	305.09 (123.5)
			cpl (sd)	14.08 (6.1)	11.75 (4.8)	11.49 (4.7)
		HU	MSE (sd)	499.32 (218.8)	307.6 (128.7)	305.09 (123.5)
			cpl (sd)	14.08 (6.1)	11.63 (4.7)	11.49 (4.7)
30	32	U	MSE (sd)	691.55 (503.2)	193.45 (54.2)	150.17 (35)
			cpl (sd)	12.54 (7.2)	10.26 (2.7)	7.81 (1.4)
		HU	MSE (sd)	691.55 (503.2)	279.21 (117.3)	150.17 (35)
			cpl (sd)	12.54 (7.2)	13 (4.3)	7.81 (1.4)
50	16	U	MSE (sd)	282.05 (134.8)	180.64 (47.6)	181.49 (48.6)
			cpl (sd)	20.37 (6.3)	13.68 (4)	14.05 (4)
		HU	MSE (sd)	282.05 (134.8)	180.84 (47.5)	181.49 (48.6)
			cpl (sd)	20.37 (6.3)	14.03 (4)	14.05 (4)
50	32	U	MSE (sd)	308.13 (142.9)	129.39 (19.6)	122.71 (17.8)
			cpl (sd)	22.94 (7.7)	9.37 (2.1)	7.91 (1.4)
		HU	MSE (sd)	308.13 (142.9)	159.14 (30.4)	122.71 (17.8)
			cpl (sd)	22.94 (7.7)	15.42 (4)	7.91 (1.4)
100	16	U	MSE (sd)	150.14 (19.7)	127.48 (14)	128.21 (15)
			cpl (sd)	25.6 (5.4)	14.73 (3.8)	14.73 (3.9)
		HU	MSE (sd)	150.14 (19.7)	128.21 (14.9)	128.21 (15)
			cpl (sd)	25.6 (5.4)	14.7 (3.9)	14.73 (3.9)
100	32	U	MSE (sd)	157.18 (23.5)	110.24 (8)	109.24 (7.8)
			cpl (sd)	30.41 (6.7)	8.96 (1.3)	8.3 (1)
		HU	MSE (sd)	157.18 (23.5)	123.8 (12.5)	109.24 (7.8)
			cpl (sd)	30.41 (6.7)	17.18 (3.3)	8.3 (1)

Table 8: Elasticnet (with Elasticnet) combined with constrained CorReg.Y depends on all variables in X_2.

n	p_2	ψ	indicator	Stepwise	CorReg $\hat{S}$	CorReg S
30	16	U	MSE (sd)	394.59 (454.6)	341.47 (153.2)	351.14 (148)
			cpl (sd)	13.09 (3.2)	12.9 (2.9)	12.92 (3)
		HU	MSE (sd)	394.59 (454.6)	353.99 (157)	351.14 (148)
			cpl (sd)	13.09 (3.2)	12.83 (2.9)	12.92 (3)
30	32	U	MSE (sd)	410.11 (343.5)	163.05 (45.4)	152.08 (38.1)
			cpl (sd)	15.8 (6.5)	8.29 (2.2)	7.34 (1.3)
		HU	MSE (sd)	410.11 (343.5)	178.94 (69.7)	152.08 (38.1)
			cpl (sd)	15.8 (6.5)	8.89 (3)	7.34 (1.3)
50	16	U	MSE (sd)	180.67 (52.7)	173.02 (38.3)	174.84 (36.8)
			cpl (sd)	13.56 (3.7)	12.78 (2.4)	13.07 (2.4)
		HU	MSE (sd)	180.67 (52.7)	175.3 (38.1)	174.84 (36.8)
			cpl (sd)	13.56 (3.7)	13.01 (2.4)	13.07 (2.4)
50	32	U	MSE (sd)	188.91 (88.5)	124.38 (17.4)	122.38 (16.7)
			cpl (sd)	14.74 (6.6)	8.07 (1.6)	7.65 (1.2)
		HU	MSE (sd)	188.91 (88.5)	134.22 (23.6)	122.38 (16.7)
			cpl (sd)	14.74 (6.6)	9.5 (2.8)	7.65 (1.2)
100	16	U	MSE (sd)	128.18 (13.7)	127.43 (13.7)	127.71 (13.3)
			cpl (sd)	13.78 (2.8)	13.51 (2)	13.59 (2.1)
		HU	MSE (sd)	128.18 (13.7)	127.71 (13.3)	127.71 (13.3)
			cpl (sd)	13.78 (2.8)	13.59 (2.1)	13.59 (2.1)
100	32	U	MSE (sd)	126.15 (20.6)	109.47 (8)	109.04 (7.8)
			cpl (sd)	13.25 (4.4)	8.3 (1)	8.12 (0.8)
		HU	MSE (sd)	126.15 (20.6)	113.96 (11.4)	109.04 (7.8)
			cpl (sd)	13.25 (4.4)	10.08 (2.6)	8.12 (0.8)

Table 9: Stepwise combined with constrained CorReg. Y depends on all variables in X_2.

4.4 Robustess of the model

We have generated non linear structures (Tables 10 and 11). Variables in X_2 depends on the log or the square (randomly choosen with equiprobability) of a variable in X_1 . Dependencies are still real but non linear. CorReg still found dependencies. One example of non-linear structure found with CorReg on real datasets is given in section 5.1.

n	p_2	ψ	Time Mixmod	Time MCMC	TL	WL	ML	Δp_2	$\Delta compl$
30	16	U	0.4313	2.7769	9.96	2.45	5.95	3.5	29.42
			(0.0391)	(0.0883)	(1.3175)	(1.3808)	(1.3056)	(1.453)	(6.2639)
		HU	0.4313	4.9079	7.28	1.25	8.63	7.38	5.77
			(0.0391)	(0.4353)	(1.5769)	(0.9987)	(1.5548)	(1.6316)	(6.1297)

Table 10: Results of the Markov chain for non linear structure (log and square). Mean observed and standard deviation (sd).

n	p_2	ψ	indicator	LASSO	CorReg $\hat{S}$	CorReg S
30	0	U	MSE (sd)	996.06 (323.7)	1100.17 (391.4)	1501.25 (905.9)
			cpl (sd)	17.66 (5.4)	14.7 (4.7)	11.97 (4.9)
		HU	MSE (sd)	996.06 (323.7)	1094.5 (621.8)	1501.25 (905.9)
			cpl (sd)	17.66 (5.4)	16.98 (4.8)	11.97 (4.9)

Table 11: Y depends on all variables in X. CorReg is not too far from LASSO even if it stays behind.

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.

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)

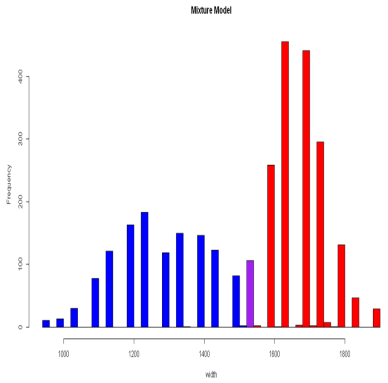

Figure 4: Example of non-Gaussian real variable easily modeled by a Gaussian mixture

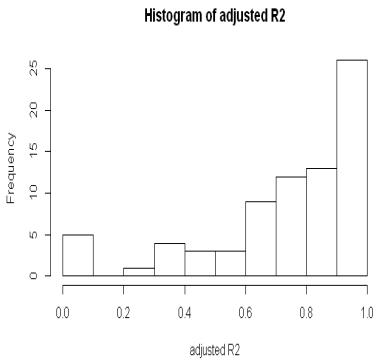

Figure 5: R^2_{adj} of the 76 subregressions.

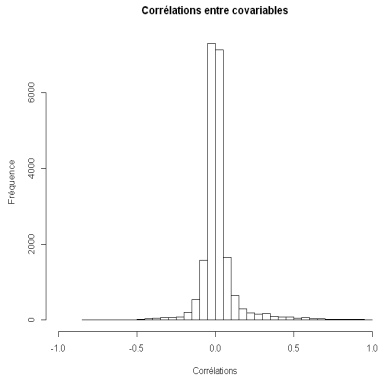

Figure 6: Histogram of correlations in X.

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 $p_2 = 76$ subregressions with a mean of $\bar{p_1} = 5.17$ regressors. In X_1 the number of $\rho > 0.7$ is **79.33%** smaller than in X.

It is now time to look at the predictive results (Figure 6). 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.

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

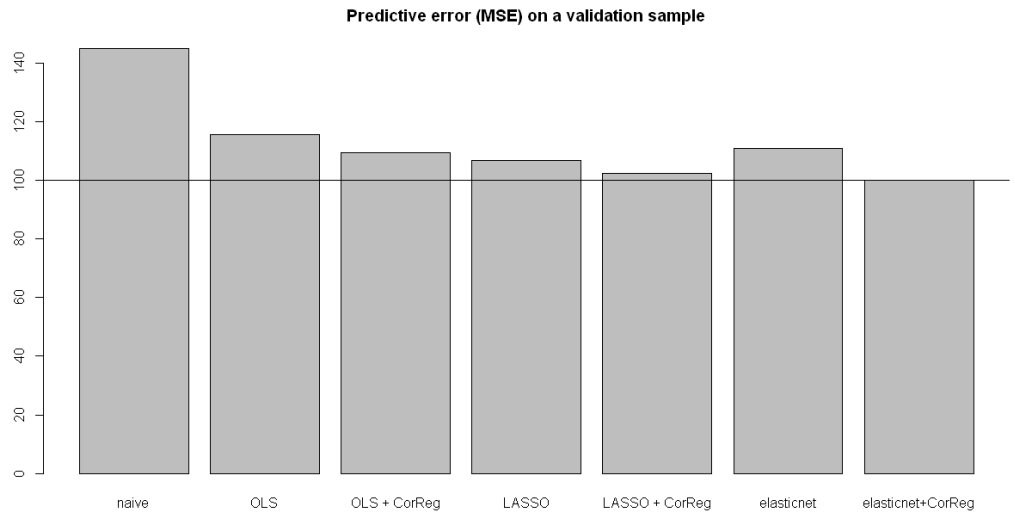

Figure 7: MSE comparison on industrial dataset. Learning set : 3 000 products, validation set : 847 products

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 12: Results obtained on a validation sample.

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

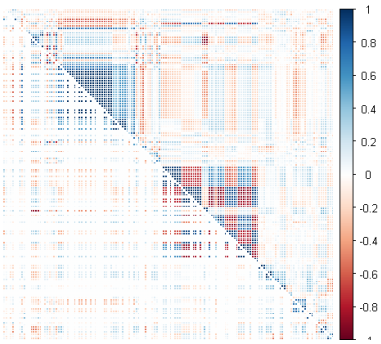

Figure 8: Correlations between the covariates in X (upper) and $\hat{X}_1$ (lower).

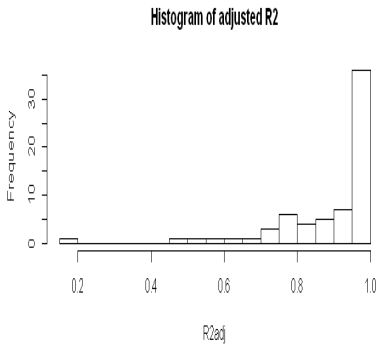

Figure 9: R^2_{adj} of the 67 sub-regressions.

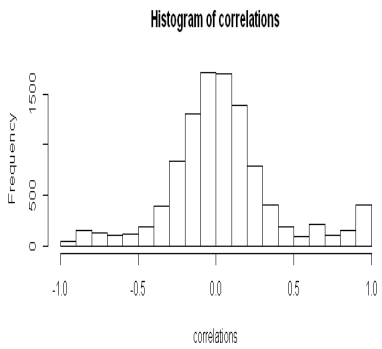

Figure 10: Histogram of correlations in X.

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 13: 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 interprete and predict physical phenomenon efficiently and to help to manage missing values. 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 actual generative model allows such a functionality 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] Howard D Bondell and Brian J Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

[2] Mary-Huard T. Chiquet J. and Robin S. Multi-trait genomic selection via multivariate regression with structured regularization. *Proceedings of MLCB/NIPS'13 workshop*, 2013.

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

[4] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[5] Dean P Foster and Edward I George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975, 1994.

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

[7] E.I. George and R.E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, pages 881–889, 1993.

[8] Walter R Gilks, Sylvia Richardson, and David J Spiegelhalter. *Markov chain Monte Carlo in practice*, volume 2. CRC press, 1996.

[9] Charles Miller Grinstead and James Laurie Snell. *Introduction to probability*. American Mathematical Soc., 1997.

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

[11] Vasilios N Katsikis and Dimitrios 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] D.W. Marquardt and R.D. Snee. Ridge regression in practice. *American Statistician*, pages 3–20, 1975.

[14] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[15] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 936. John Wiley & Sons, 2012.

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

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

[18] Sijian Wang, Bin Nan, Saharon Rosset, and Ji Zhu. Random lasso. *The annals of applied statistics*, 5(1):468, 2011.

[19] Loic Yengo, Julien Jacques, Christophe Biernacki, et al. Variable clustering in high dimensional linear regression models. 2012.

[20] 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.

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

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

[23] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[24] Hui Zou and Trevor 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 discrete 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? The answer is no. Thus S is identifiable.

If there are exact regressions ($\sigma_j^2 = 0$) in (8), 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. In the followings we suppose $\sigma_j^2 \neq 0$.

8.2 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_1 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 $p_2 < \frac{p}{2}$ and $p_1^j < \frac{p}{2}$ so candidates may be rejected to satisfy this hypothesis. Stronger constraints on p_2 and/or p_1 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_1 if the criterion ψ 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.