CorReg pretreatment : Regression for correlated variables and application in steel industry

Clément THERY

March 12, 2014

Abstract. Linear regression generally suppose to have decorrelated covariates. This hypothesis is often irrealist with industrial datasets that contains many highly correlated covariates due to the process, physcial laws, *etc.* The proposed generative model consists in explicit modeling of the correlations with a family of linear regressions between the covariates permitting to obtain by marginalization a parsimonious correlation-free regression model, easily understandable and compatible with variable selection methods. 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.

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 and is simple to learn for non-statistician. 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 that faces an also classical problem : the variance of the estimators. When estimating the parameters of the regression we have to compute the inverse of a matrix[14] which will be ill-conditioned or even singular if some covariates depend linearly from each other. For a model defined by

$$Y_{|X} = X\beta + \varepsilon \tag{1}$$

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

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

With variance

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

And when correlations between covariates are strong, the matrix to invert is ill-conditioned and the variance explodes. This variance 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. For example, this work takes place in an

industrial (steel industry) context with a big set of covariates (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, with a preference for interpretation in industrial context (better to improve the process when possible than to 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 implies many irrelevant variables (datasets based on the whole process without any a priori) so we have to use variable selection methods.

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

Ridge regression[12] proposes a biased estimator that can be written in terms of a parametric L_2 penalty:

$$\hat{\beta} = \text{argmin} \left\{ \| Y - X\beta \|_2^2 \right\} \text{ subject to } \| \beta \|_2^2 \leq k \tag{4}$$

But Ridge regression is not efficient to select covariates (it's an assumed choice) because coefficients tends to 0 but don't reach 0. So it gives difficult interpretations for large values of p and is not adapted for our industrial context. We need to reduce the dimension of the model. Our goal is not just to predict but also to understand the response variable.

The Least Absolute Shrinkage and Selection Operator (LASSO)[15] consists in a shrinkage of the regression coefficients based on a λ parametric L_1 penalty.

$$\hat{\beta} = \text{argmin} \left\{ \| Y - X\beta \|_2^2 \right\} \text{ subject to } \| \beta \|_1 \leq \lambda \tag{5}$$

The Least Angle Regression[4] (LAR) 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 coeffients set exactly to 0. But LASSO also faces consistency problems [21] when confronted with correlated covariates. Another limitation of the LASSO is that it preserves at most n predictors (troublesome when in high dimension). Some recent variants of the LASSO does exist for the choice of the penalization coefficient like the adaptative LASSO [22] or the random LASSO [17].

Elastic net[23] is a method developped to be a compromise between Ridge regression and the LASSO. Elastic net can be written:

$$\hat{\beta} = (1 + \lambda_2) \, \text{argmin} \left\{ \| Y - X\beta \|_2^2 \right\}, \text{ subject to } (1 - \alpha) \| \beta \|_1 + \alpha \| \beta \|_2^2 \leq t \text{ for some } t \tag{6}$$

where $\alpha = \frac{\lambda_2}{(\lambda_1 + \lambda_2)}$. It seems to give good predictions. But it is based on the grouping effect and if the dataset contains two identical variables they will obtain the same coefficient 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[18] (CLERE) describes the β_j no longer as fixed effect parameters but as unobserved independant random variables whith 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 $\hat{\beta}$. 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.

When some try to reduce the dimension and then just hope to have small correlations in the remaining dimensions, we propose to focus on the correlations, giving a model with orthogonal covariates and an explicit structure between covariates. 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 dimension reduction tools without suffering from correlations. We only consider strong correlations (i.e. : problematic ones) thus we keep most of the information contained in the dataset.

Our work is based on the assumption that if we know that correlations are a problem and if we precisely know the correlations, we could use this knowledge to avoid the problem. The idea is to suppose explicitly a linear structure between the covariates. It 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 [16] when the structure is straight forward (recursive SEM). Other methods already rely on linear structure and start to take into account correlations but they only consider covariances between the residuals (SUR) [19] or covariances between the endogenous variables like SPRING (Structured selection of Primordial Relationships IN the General linear model) [2].

In this work, we decide to distinguish the response variable from the other endogenous variables (that are on the left of a regression). Thus we don't have a system of regressions but one regression on our response variable and a system of subregressions (without the response variable). And 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 reduced set of independent covariates. Thus we reduce dimension and correlations in the same time. The structure justifies the eviction of the redundant covariates without significant information loss. It can be seen as a pretreatment on the dataset based on the hypothesis of a strong structure between the covariates (i.e. : small ε_X).

We can use any variable selection method on the reduced dataset with improved efficiency (reduced variance) due to dimension reduction and correlation suppression. So we obtain two kinds of zeros in our first 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.

But to work, we need an explicit structure between the covariates. SEM are often used in social sciences and economy where a structure is supposed "by hand" but here we want to be able to find a structure without any a priori (possibility to include some known structure remains). 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 graph for our SEM. So we developped an MCMC algorithm to find it (R package CorReg on CRAN). However, Graphical LASSO can be used in the initialization step of our MCMC. This structure is based on gaussian mixture models to fit better real datasets and to allow identifiability of the structure in terms of complexity (number of parameters).

This paper will first present the reduced model and its properties before describing in Section 3 the algorithm 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 sixth part.

2 Model to decorrelate the covariates

2.1 The generative model

Let $Y \in \mathcal{R}^n$ be a response variable we want to explain with a set $X \in \mathcal{R}^{n \times p}$ of p correlated covariates. We propose to explicitly define a family of p_2 internal regressions between covariates whith I_2 the set of indices of endogenous variables in 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 $p_1 = (p_1^1, \ldots, p_1^{p_2})$ 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 X^j the j^{th} column of a matrix X. For lighter notation we define $X_2 = X^{I_2}$ the matrix of the endogenous covariates and $X_1 = X \setminus X_2$ the matrix of the remaining exogenous covariates.

We can now write the generative model:

- Main regression between Y and X:

$$Y_{|S} = XA + \varepsilon_Y = X_1 A_1 + X_2 A_2 + \varepsilon_Y \text{ whith } \varepsilon_Y \sim \mathcal{N}(0, \sigma_Y^2 I_n); \tag{7}$$

  where $A = (A_1, A_2) \in \mathcal{R}^p$ is the vector of the regression coefficients and I_n the identity matrix,

- Family of p_2 regressions within correlated covariates in X:

$$\forall j \in I_2 : X^j_{|X_1,S} = X_1 B_1^j + \varepsilon_j \text{ with } \varepsilon_j \sim (0, \sigma_j^2 I_n); \tag{8}$$

  where $B_1^j \in \mathcal{R}^{(p-p_2)}$ are the vectors of the regression coefficients between the covariates (containing some zeros according to I_1^j).

- $p - p_2$ remaining independent exogenous Gaussian mixtures:

$$\forall j \notin I_2 : X^j_{|S} \sim f(\theta_j) \tag{9}$$

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

This generative model is conditionnal to S, the discrete structure model that is identifiable because we can't permute some regressions in (8) and obtain the same joint distribution $P(X, Y)$, the residuals (ε_j) would not stay Gaussian. These residuals are in the following supposed independent but one can suppose dependencies between them and then use appropriate tools to estimate them and the B_1^j like SUR with Feasible Generalized Least Squares (FGLS) by Zellner [19] or SPRING [2] as mentioned in the introduction.
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 subregression 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.

2.2 Estimator and properties

We note that (7) and (8) also give by simple integration on X_2 a regression model on Y *depending only on uncorrelated covariates* X_1:

$$Y = X_1(A_1 + \sum_{j \in I_2} B_1^j A_j) + \sum_{j \in I_2} \varepsilon_j A_j + \varepsilon_Y = X_1 \alpha_1 + \varepsilon_\alpha \tag{10}$$

So we have the unbiased estimator:

$$\hat{\alpha}_1 = (X_1' X^1)^{-1} X_1' Y \tag{11}$$

with variance:

$$\text{Var}[\hat{\alpha}_1 | X] = (\sigma_Y^2 + \sum_{j \in I_2} \sigma_j^2 A_j^2)(X_1' X^1)^{-1} \tag{12}$$

We see that the variance is reduced (no correlations and smaller matrix give better conditioning) for small values of σ_j *i.e.* strong correlations. Moreover, 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. This new model is reduced even without variable selection and is just a linear regression so every method for variable selection in linear regression can be used. Hence we hope to obtain a parsimonious model with *two kinds of zeros*: those from decorrelating step and those from selection step, with different meanings.
The CorReg package proposes to make selection with both LASSO (with LAR), elasticnet, adaptative lasso, and others more classical methods. 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 [20]).

2.3 Better interpretation

Grouping effect is the fact that correlated covariates get similar coefficient and are selected together. It's the case with elasticnet [23] for example. If $X_2 = X_1 + e$ and we have the grouping effect, we will obtain a model like $\hat{Y} = \hat{a}X_1 + \hat{a}X_2$. Then, if you try to modify the response value, you will modify one of the covariates and both will change so you won't get expected results.

At the opposite, methods like the LASSO with LAR will only keep one covariate for each group of correlated covariates so we get $\hat{Y} = 2aX_1$ and think that $X_2 \perp Y$ but Y depends on X_2.

Nothing constrains us to give only one equation. It is clearly better to give the user another equation (or system for more complex models) describing the correlations. So you get the following model : $Y = aX_1 + aX_2$ and $X_1 = X_2 + e$. Then you have more information and are able to decide better actions. With such a model, grouping effect is no more useful because when saying $Y = 2aX_1$ and $X_2 = X_1 + e$ with the information that X_2 has been removed locally because of its correlation with X_1, you don't get misleading interpretations anymore. So it is possible to combine the advantages of grouping effect and selection just giving several equations. Each equation here is very simple (only linear regressions) so you don't really increase complexity of the model. Moreover, the uncrossing constraint ($I_1 \cap I_2 = \emptyset$) guarantee to keep a simple structure easily interpretable.

2.4 More about the generative model

2.4.1 Structure comparison

All our work is based on S, the linear structure between the covariates. Our 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 [11]. 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

$$P(S|X) \propto P(X|S)P(S) \tag{13}$$

$$BIC = -2\mathcal{L}(X, S, \Theta) + |\Theta| \log(n) \approx -2 \log P(X|S) \tag{14}$$

But BIC tends to give too complex structures because we test a great range of models. Thus CorReg allows to 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 hypothesis are made and will become constraints in the MCMC when hierarchical hypothesis is made (it will be the case 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.

2.4.2 Detailed generative model

To better fit industrial variables (Figure 1), we suppose that variables in X_1 follow Gaussian mixtures. The great flexibility [13] 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.

$$\forall j \notin I_2 : X^j_{|S} \sim f(\theta_j) = \mathcal{GM}(\pi_j; \mu_j; \sigma_j^2) \text{ with } \pi_j, \mu_j, \sigma_j^2 \text{ vectors of size } k_j; \tag{15}$$

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 compute corresponding ψ separately, give it as an input of CorReg and then improve the efficiency of the algorithm (it will find a structure only if it is really relevant).

3 Estimating structure of sub-regressions with a Markov chain

3.1 The neighbourhood

3.1.1 Definition

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:

$$\mathcal{V}_{S,j} = \{S\} \cup \{S^{(i,j)} | 1 \leq i \leq p, i \neq j\} \tag{16}$$

$$\text{where } j \sim \mathcal{U}(\{1, \ldots, p\}) \tag{17}$$

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 $\forall 1 \leq j \leq p : p_1^j = |I_1^j|$, $I_2 = \{j | p_1^j > 0\}$, $p_2 = |I_2|$, .

3.1.2 Alternatives

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. Strongest constraints on p_2 and/or p_1 can be given in CorReg if relevant.

If the algorithm did not have time to converge, 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.

3.2 The walk

3.2.1 Transition probabilities

We first make the approximation

$$P(S|X) \approx exp(\psi(S)). \tag{18}$$

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:

$$\forall (S, \tilde{S}) \in \mathcal{S}^2 : \mathcal{P}(S, \tilde{S}) = \sum_{j=1}^{p} \mathbf{1}_{\{\tilde{S} \in \mathcal{V}_{S,j}\}} \frac{\exp(-\frac{1}{2}\psi(\tilde{S}))}{\sum_{S_l \in \mathcal{V}_{S,j}} \exp(-\frac{1}{2}\psi(S_l))} \tag{19}$$

$$\tag{20}$$

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 3)).

3.2.2 Initialisation

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 can be reduced by the hadamard product with the binary matrix obtained by Graphical Lasso[6]. Graphical LASSO is time consuming so we only used it on real datasets (each simulation was computed 100 times).

3.2.3 Reduced neighbourhood

3.2.4 Multiple try

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. Each configuration has been tested 100 times and the Mean Squared Error (MSE) given were computed on a validation sample of 1000 individuals. Each simulated dataset presented here had $p = 40$ covariates to explain the response variable. Section 4.2.2 show the results obtained in terms of $\hat{S}$. Sections 4.3 et 4.4 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 B_1^j are generated according to the same Poisson law but with a random sign. S only contains binary relationships but CorReg was not constrained. We used Rmixmod to estimate the densities of each covariate.

4.2 Finding the structure

4.2.1 How to evaluate found structure ?

The first criterion is ψ which is minimised in the MCMC. We can compare this criterion for both the true Structure and the found one. But it won't show how far the found structure is from the true one in terms of S. So 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 $BIC(S)$ is the BIC of the True generative model S, $BIC(\emptyset)$ is the BIC of the trivial model (no regression, $I_2 = \emptyset$) and $BIC(\hat{S})$ is the BIC of the found structure.

n	p_2	$BIC(S)$	$BIC(\emptyset)$	$BIC(\hat{S})$	TL	WL	ML	Δp_2	$\Delta compl$
30	0	10 291.85	10 291.85	-2 354.09	0	7.3	0	-7.3	200.7
		(221.1)	(221.1)	(1 512.3)	(0)	(1)	(0)	(1)	(23.5)
30	16	874.58	11 414.72	37.87	7.78	5.85	8.08	2.23	124.82
		(164.5)	(387.7)	(1 546.7)	(2.7)	(2.1)	(2.7)	(3.1)	(58.1)
30	32	-8419.17	12248.93	-4043.98	22.86	1.75	8.92	7.17	40.38
		(74.6)	(473.7)	(1041.8)	(3.2)	(1.2)	(3.2)	(2.9)	(30.9)
50	0	16 770.33	16 770.33	16 742.84	0	3.91	0	-3.91	17.12
		(313.9)	(313.9)	(318.2)	(0)	(1.5)	(0)	(1.5)	(8.5)
50	16	1 173.65	18 662.67	1 158.82	11.15	5.37	4.74	-0.63	27.14
		(241.3)	(600.5)	(289.4)	(2)	(2.1)	(1.9)	(0.8)	(7.9)
50	32	-14176.34	20106.95	-11011.65	27.54	1.26	4.21	2.95	14.8
		(125.7)	(785.9)	(1329.5)	(1.7)	(1)	(1.7)	(1.2)	(4.3)

Table 1: Results of the Markov chain with no constraint on $\hat{p}_1$. Mean observed and standard deviation (sd).

With no constraints on $\hat{p}_1$ when $n < p$ we can have a perfect overlearning, so it is recommended to constrain each $\hat{p}_1^j < n$. Table 12 shows the impact on the research of such a constraint.

n	p_2	$BIC(S)$	$BIC(\emptyset)$	$BIC(\hat{S})$	TL	WL	ML	Δp_2	$\Delta compl$
30	0	10 288.64	10 288.64	10 240.77	0	5.3	0	-5.3	21.44
		(229.3)	(229.3)	(235.8)	(0)	(1.8)	(0)	(1.8)	(7.1)
30	16	8 865.52	11 339.8	793.46	10.71	6.08	5.12	-0.96	39.32
		(172.6)	(456.5)	(217.8)	(2.2)	(2.3)	(2.2)	(0.8)	(6.6)
30	32	-8 405.13	12 373.96	-5 482.67	25.25	2.1	6.63	4.53	26.66
		(86.4)	(454)	(572.4)	(1.5)	(1.2)	(1.5)	(0.9)	(6.3)
50	0	16 870.51	16 870.51	16 848.87	0	4.02	0	-4.02	12.74
		(380.8)	(380.8)	(386.5)	(0)	(1.7)	(0)	(1.7)	(5.8)
50	16	1 215.15	18 737.71	1 182.58	11.4	5.14	4.51	-0.63	23.62
		(252)	(605.6)	(279.1)	(1.9)	(2)	(1.8)	(0.7)	(5.2)
50	32	-14 153.99	20 231.62	-11 599.72	28.04	1.37	3.75	2.38	13.79
		(124.5)	(853.1)	(1 035.3)	(1.5)	(0.8)	(1.4)	(1)	(4.5)

Table 2: Results of the Markov chain with constraint $\hat{p}_1^j \leq 5$. Mean observed and standard deviation (sd).

n	p_2	$BIC(S)$	$BIC(\emptyset)$	$BIC(\hat{S})$	TL	WL	ML	Δp_2	$\Delta compl$
30	16	?	?	?	?	?	?	?	?
		(?)	(?)	(?)	(?)	(?)	(?)	(?)	(?)

Table 3: Results of the Markov chain for non linear structure (??). Mean observed and standard deviation (sd).

4.3 Y depends on all variables in X

4.3.1 No constraints on p_1

n	p_2	indicator	OLS	CorReg $\hat{S}$	CorReg S
30	0	MSE (sd)	5.42 10^{10} (2.99 10^{11})	1.01 10^{10} (6.49 10^{10})	5.42 10^{10} (2.99 10^{11})
		complexity (sd)	30 (0)	30 (0)	30 (0)
30	16	MSE (sd)	1.1 10^{11} (6.97 10^{11})	1.1 10^{10} (1.03 10^{10})	659.24 (564.6)
		complexity (sd)	30 (0)	27.09 (2.8)	25.14 (0.4)
30	32	MSE (sd)	1.15 10^{8} (7.00 10^{8})	275.37 (509.7)	138.7 (20.7)
		complexity (sd)	30 (0)	16.38 (2.9)	9.22 (0.4)
50	0	MSE (sd)	578.76 (304.1)	64 513.24 (121 388.1)	578.76 (304.1)
		complexity (sd)	41 (0)	37.09 (1.5)	41 (0)
50	16	MSE (sd)	659.55 (450.3)	1 704.51 (2996.7)	203.67 (49)
		complexity (sd)	41 (0)	24.48 (0.8)	25.11 (0.3)
50	32	MSE (sd)	595.54 (314.3)	129.7 (15.2)	120.18 (12.8)
		complexity (sd)	41 (0)	12.2 (1.2)	9.25 (0.5)

Table 4: OLS and OLS combined with CorReg. Y depends on all variables in X. 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	indicator	LAR	CorReg $\hat{S}$	CorReg S
30	0	MSE (sd)	2.32 10^{7} (9.475 10^{6})	2.62 10^{7} (1.56 10^{7})	2.32 10^{7} (9.475 10^{6})
		complexity (sd)	18.6 (5)	18.28 (4.5)	18.6 (5)
30	16	MSE (sd)	4.44 10^{6} (7.22 10^{6})	7.72 10^{5}(2.03 10^{6})	723.52 (657.4)
		complexity (sd)	23.16 (3.1)	24.13 (2.4)	24.95 (0.5)
30	32	MSE (sd)	235.7 (142.1)	168.66 (44)	138.4 (20.8)
		complexity (sd)	14.65 (3.5)	11.43 (2)	9.08 (0.3)
50	0	MSE (sd)	583.84 (319.3)	71 360.24 (145 938.8)	583.84 (319.3)
		complexity (sd)	40.77 (0.6)	36.2 (2.1)	40.77 (0.6)
50	16	MSE (sd)	344.42 (191.1)	1 704.51 (3006.4)	203.54 (50.5)
		complexity (sd)	31.98 (2.9)	24.3 (0.8)	24.93 (0.4)
50	32	MSE (sd)	162.1 (75.8)	124.53 (15.2)	119.85 (12.7)
		complexity (sd)	15.23 (5)	9.98 (1.1)	9.07 (0.3)

Table 5: LASSO (with LAR) combined with CorReg.Y depends on all variables in X. CorReg logically wins

4.3.2 Constrained p_1

n	p_2	indicator	OLS	CorReg $\hat{S}$	CorReg S
30	0	MSE (sd)	$4.37\ 10^{10}$ ($1.63\ 10^{10}$)	$1.32\ 10^{10}$ ($8.24\ 10^{10}$)	$4.37\ 10^{10}$ ($1.63\ 10^{10}$)
		complexity (sd)	30 (0)	30 (0)	30 (0)
30	16	MSE (sd)	$4.24\ 10^{13}$ ($4.23\ 10^{14}$)	15 405 (54 420)	646.55 (572.6)
		complexity (sd)	30 (0)	24.21 (0.8)	25.17 (0.4)
30	32	MSE (sd)	$3.20\ 10^{8}$ ($2.85\ 10^{9}$)	172.71 (33.4)	137.35 (21.1)
		complexity (sd)	30 (0)	13.65 (0.9)	9.12 (0.3)
50	0	MSE (sd)	553.6 (254.7)	60 218.38 (88 854.3)	553.6 (254.7)
		complexity (sd)	41 (0)	36.98 (1.7)	41 (0)
50	16	MSE (sd)	545.36 (248.6)	3 285.63 (10 979.7)	192.4 (38.3)
		complexity (sd)	41 (0)	24.46 (0.7)	25.09 (0.3)
50	32	MSE (sd)	597.77 (378.2)	126.75 (13.4)	119.2 (10.4)
		complexity (sd)	41 (0)	11.59 (1.1)	9.21 (0.5)

Table 6: 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	indicator	LAR	CorReg $\hat{S}$	CorReg S
30	0	MSE (sd)	$2.4\ 10^{7}$ ($1.05\ 10^{7}$)	$2.16\ 10^{7}$ ($1.02\ 10^{7}$)	$2.4\ 10^{7}$ ($1.05\ 10^{7}$)
		complexity (sd)	18.6 (5)	18.28 (4.5)	18.6 (5)
30	16	MSE (sd)	$6.0\ 10^{6}$ ($8.35\ 10^{6}$)	29 052.99 (169 828.6)	725.14 (767.8)
		complexity (sd)	22.91 (3.2)	23.75 (1.4)	24.95 (0.4)
30	32	MSE (sd)	260.67 (210.8)	153.64 (35.9)	137.23 (21.1)
		complexity (sd)	14.58 (3.3)	10.56 (1.4)	9.06 (0.2)
50	0	MSE (sd)	604.2 (600.9)	76 213.71 (122 663.8)	604.2 (600.9)
		complexity (sd)	40.82 (0.5)	36.17 (2.2)	40.82 (0.5)
50	16	MSE (sd)	312.6 (194.5)	3 547.33 (12 534.3)	192.13 (38.9)
		complexity (sd)	31.28 (2.7)	24.26 (1)	24.97 (0.5)
50	32	MSE (sd)	157.9 (41.6)	122.76 (13.6)	118.69 (10.6)
		complexity (sd)	14.56 (4.6)	9.74 (1)	9.01 (0.1)

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

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

n	p_2	Indicator	OLS	CorReg $\hat{S}$	CorReg S
30	0	MSE (sd)	?	?	?
		complexity (sd)	?	?	?
30	16	MSE (sd)	?	?	?
		complexity (sd)	?	?	?
30	32	MSE (sd)	?	?	?
		complexity (sd)	?	?	?
50	0	MSE (sd)	502.05 (227.4)	108891.58 (416491.9)	502.05 (227.4)
		complexity (sd)	41 (0)	36.97 (1.6)	41 (0)
50	16	MSE (sd)	580.25 (297.7)	188.82 (36.2)	196.54 (41.2)
		complexity (sd)	41 (0)	24.26 (0.8)	25.08 (0.3)
50	32	MSE (sd)	592.65 (275.5)	129.11 (14)	122.64 (12.9)
		complexity (sd)	41 (0)	11.49 (0.9)	9.29 (0.6
400	0	MSE (sd)	?	?	?
		complexity (sd)	?	?	?
400	16	MSE (sd)	?	?	?
		complexity (sd)	?	?	?
400	32	MSE (sd)	?	?	?
		complexity (sd)	?	?	?

Table 8: OLS and OLS combined with constrained CorReg. Y depends only on variables in X_2. CorReg does not allow the true model but still wins.

n	p_2	Indicator	LAR	CorReg $\hat{S}$	CorReg S
30	0	MSE (sd)	?	?	?
		complexity (sd)	?	?	?
30	16	MSE (sd)	?	?	?
		complexity (sd)	?	?	?
30	32	MSE (sd)	?	?	?
		complexity (sd)	?	?	?
50	0	MSE (sd)	599.19 (427.2)	124506.64 (436522.1)	599.19 (427.2)
		complexity (sd)	40.78 (0.5)	36.2 (1.9)	40.78 (0.5)
50	16	MSE (sd)	195.2 (68.8)	157.06 (29.4)	160.37 (31.2)
		complexity (sd)	19.66 (4.2)	15.62 (2.6)	15.9 (2.8)
50	32	MSE (sd)	161.07 (64.2)	124.88 (13.8)	122.12 (12.5)
		complexity (sd)	14.42 (4.8)	9.48 (0.9)	8.9 (0.4
400	0	MSE (sd)	?	?	?
		complexity (sd)	?	?	?
400	16	MSE (sd)	?	?	?
		complexity (sd)	?	?	?
400	32	MSE (sd)	?	?	?
		complexity (sd)	?	?	?

Table 9: LASSO (with LAR) combined with constrained CorReg.Y depends only on variables in X_2. CorReg does not allow the true model but still wins.

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)

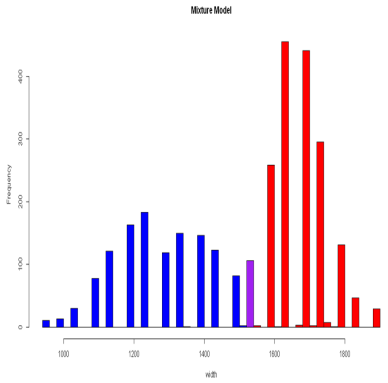

Figure 1: Example of non-Gaussian real variable easily modeled by a Gaussian mixture

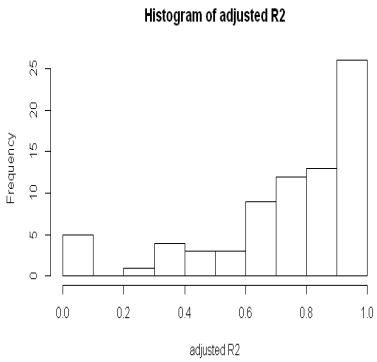

Figure 2: R^2_{adj} of the 76 subregressions.

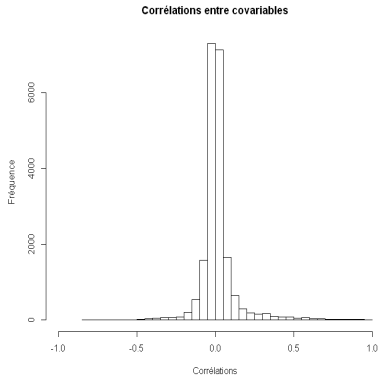

Figure 3: Histogram of correlations in X.

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). Regulation models used to optimize the process were also found. 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 5.1). 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

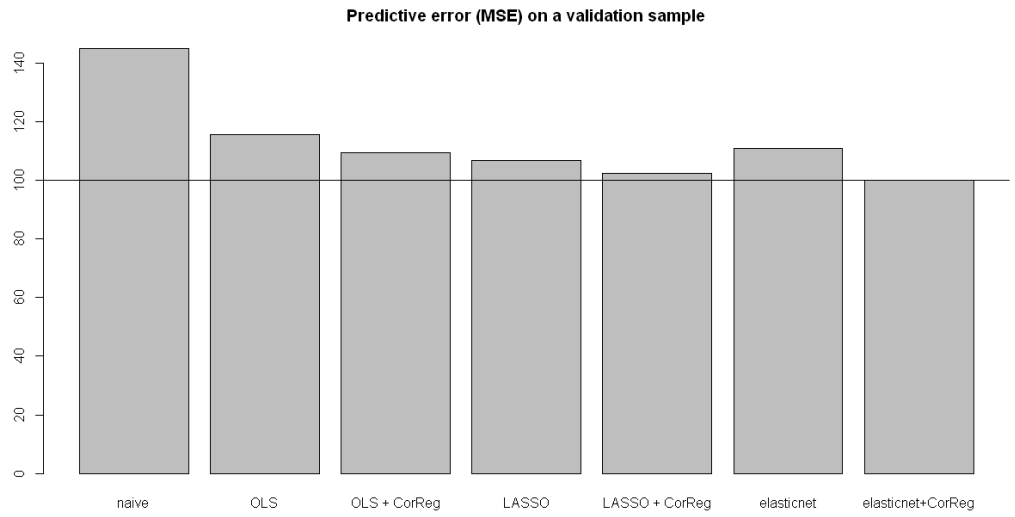

Figure 4: MSE comparison on industrial dataset. Learning set : 3 000 products, validation set : 847 products

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 10: 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

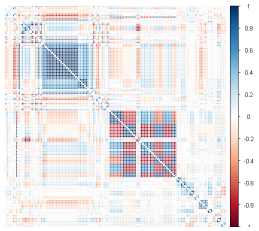

Figure 5: Correlations between the covariates

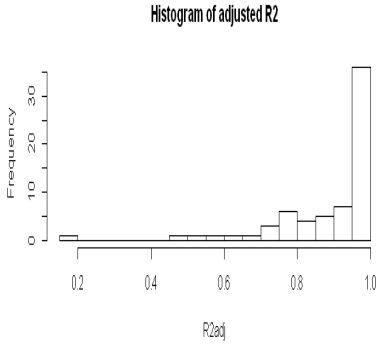

Figure 6: R^2_{adj} of the 67 sub-regressions.

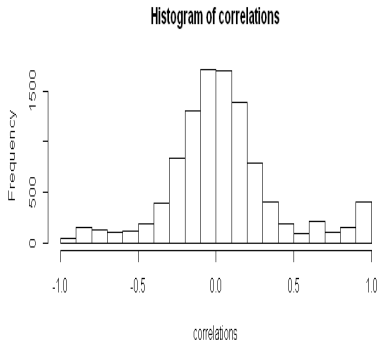

Figure 7: Histogram of correlations in X.

Here $n < p$ so we only have 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 11: 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

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.

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

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

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

[14] G. Saporta. *Probabilités, analyse des données et statistique*. Editions Technip, 2006.

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

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

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

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

[19] 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.

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

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

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

[23] 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.

7 Etoile ou pas ?

qui	p_2	$BIC(S)$	$BIC(\emptyset)$	$BIC(\hat{S})$	TL	WL	ML	Δp_2	$\Delta compl$
BIC	16	1 215.15	18 737.71	1 182.58	11.4	5.14	4.51	-0.63	23.62
		(252)	(605.6)	(279.1)	(1.9)	(2)	(1.8)	(0.7)	(5.2)
BIC^*	16	1 322.43	18 711.38	1 309.38	11.62	4.49	4.27	-0.22	7.6
		(250.9)	(564.2)	(252.5)	(1.8)	(1.9)	(1.8)	(0.4)	(3.2)

Table 12: n=50.Results of the Markov chain with constraint $\hat{p}_1^j \leq 5$. Mean observed and standard deviation (sd).

qui	p_2	Indicator	complet	CorReg $\hat{S}$	CorReg S
BIC OLS	16	MSE (sd)	580.25 (297.7)	188.82 (36.2)	196.54 (41.2)
		complexity (sd)	41 (0)	24.26 (0.8)	25.08 (0.3)
BIC LAR	16	MSE (sd)	195.2 (68.8)	157.06 (29.4)	160.37 (31.2)
		complexity (sd)	19.66 (4.2)	15.62 (2.6)	15.9 (2.8)
BIC* OLS	16	MSE (sd)	624.09 (359.9)	196.5 (34)	198.96 (35.1)
		complexity (sd)	41 (0)	24.89 (0.5)	25.11 (0.3)
BIC* LAR	16	MSE (sd)	179.63 (44.2)	159.74 (31.2)	160.24 (31.3)
		complexity (sd)	19 (3.6)	15.74 (2.4)	15.86 (2.4)

Table 13: p1max=5. n=50. Y dépend de X2.