

# Regression for correlated variables : application in steel industry

Clément THERY

January 22, 2014

- resume
- keywords : linear regression, correlations, sem, selection, graphs, pretreatment, plugin

## 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. 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 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).

When estimating the parameters of the regression we have to compute the inverse of a matrix[11] which will be ill-conditioned or even singular if some covariates depend linearly from each other. For a model defined by

$$Y = X\beta + \varepsilon \quad (1)$$

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 :

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

Because it 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. Ridge regression[9] proposes a biased estimator :

$$\hat{\beta} = (X'X + kI)^{-1} X'Y \text{ with } k \geq 0 \quad (3)$$

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 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.

Real datasets implies many irrelevant variables (datasets based on the whole process without any a priori) so we have to use variable selection methods.

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

The Least Absolute Shrinkage and Selection Operator (LASSO)[12] consists in a shrinkage of the regression coefficients based on a  $\lambda$  parametric  $L_1$  penalty.

$$\hat{\beta} = \operatorname{argmin} \left\{ \|Y - X\beta\|_2^2 \right\} \text{ subject to } \|\beta\|_1 \leq \lambda \quad (4)$$

The Least Angle Regression[5] (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 coefficients set exactly to 0. But LASSO also faces consistency problems[17] when confronted with correlated covariates. This point will be developed further (see 5.5) with numerical results. Another limitation of the LASSO is that it preserves at most  $n$  predictors (troublesome when in high dimension).

Elastic net[18] is a method developed to be a compromise between Ridge regression and the LASSO. Elastic net can be written:

$$\hat{\beta} = (1 + \lambda_2) \operatorname{argmin} \left\{ \|Y - X\beta\|_2^2 \right\}, \text{ subject to } (1 - \alpha) \|\beta\|_1 + \alpha \|\beta\|_2^2 \leq t \text{ for some } t \quad (5)$$

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.

The CLusterwise Effect REgression[14] (CLERE) tries to reduce the dimension by considering the  $\beta_j$  no longer as fixed effect parameters but as unobserved independant random variables whith  $\beta$  following a Gaussian Mixture distribution.

$$\beta_j | \mathbf{z}_j \sim \mathcal{N} \left( \sum_{k=1}^g b_k z_{jk}, \gamma^2 \right) \text{ with} \quad (6)$$

$$\mathbf{z}_j = (z_{j1}, \dots, z_{jg}) \sim \mathcal{M}(\pi_1, \dots, \pi_g) \text{ and} \quad (7)$$

$$\forall k = 1, \dots, g \quad \sum_{j=1}^p z_{jk} \geq 1 \quad (8)$$

The idea is to hope that this 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 we need to suppose having many covariates with the same level of effect on the response variable and seems to be less efficient in prediction than elastic net.

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. In fact we search the greatest set of orthogonal covariates to keep the maximum but with an orthogonality constraint. This can be viewed as a pretreatment on the dataset allowing to use then 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. We will in a second time be able to use the remaining part of the information (sequential approach).

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 recursive Structural Equation Model (SEM)[3]. That can be viewed as a system of linear regression.

$$Y = X\beta_Y + \varepsilon_Y \quad (9)$$

$$X = XB + \varepsilon_X \quad (10)$$

Recursive sem don't really have a specific estimator because general system estimators (Seemingly Unrelated Regression (SUR)[15] and Two-Stage Least Squares (2SLS)) are equivalent to independent Ordinary Least Squares when applied to recursive SEM [13]. Our work can be viewed as a new way of estimating recursive SEM based on their own structure. In this work, we decide to distinguish the response variable from the other 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). 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  $\varepsilon_X$ ).

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.

We then observe that the reduced model only uses the partition given by the structure (who is explained and who explains) but not the structure itself (how they interacts). We also notice that even if variables are highly correlated, each can have a specific effect ( $\varepsilon_X$ ) on the response variable.

So in a second step we propose further usage of the structure, taking back correlated variables to estimate the residuals of the reduced model. This estimation is also a linear regression that can take profit from any variable selection method. The idea is to estimate the complete model under the constraint of the structure. It can be done with Constrained Least Squares [1] or sequentially with classical OLS (non optimal but easier to use selection methods).

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 developed an MCMC algorithm to find it (R package CorReg on Rforge). 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, its properties and the algorithm used to find the structure. Then we talk about further usage of the method, both estimating residuals of the reduced model and managing missing values in the dataset. We will finish with some numerical results on simulated and real industrial datasets before concluding and giving some perspectives.

## 2 Model to decorrelate the covariates

We have a  $p$  correlated covariates  $X$  to explain a response variable  $Y$ . Let  $Z$  be the adjacency matrix that defines which covariates is depending on which others. That is  $Z_{i,j} = \mathbf{1}_{(X^j \text{ depends on } X^i)}$ . We consider dependencies in a generative point of view ("depends on" means "is generated according to") so  $Z$  is not symmetric and has no cycles.

We can describe the structure  $Z$  by  $S = (p_2, I_2, p_1, I_1)$  defined by :

$$p_2 = \sum_{j=1}^p \mathbf{1}_{(\exists i, Z_{i,j} \neq 0)} \text{ the number of sub-regressions} \quad (11)$$

$$I_2 = (I_2^1, \dots, I_2^{p_2}) \text{ vector of the indices of the dependent covariates} \quad (12)$$

$$I_1 = (I_1^1, \dots, I_1^{p_1}) \text{ with} \quad (13)$$

$$I_1^j = \{i | Z_{i,j} = 1\} \text{ indices of the covariates explaining } X^j \quad (14)$$

$$p_1 = (p_1^1, \dots, p_1^{p_1}) \text{ where } p_1^j = \#I_1^j \quad (15)$$

We suppose  $I_1 \cap I_2 = \emptyset$ , i.e. dependent variables don't explain other variables in  $X$ .

We note  $I_2^c = \{1, \dots, p\} \setminus I_2$  Then our generative model can be written :

$$Y_{|X,S} = Y_{|X} = XA + \varepsilon_Y = X^{I_2^c} A_{I_2^c} + X^{I_2} A_{I_2} + \varepsilon_Y \text{ with } \varepsilon_Y \sim \mathcal{N}(0, \sigma_Y^2) \quad (16)$$

$$\forall j \in I_2 : X_{|X^{I_1^j}, S}^j = X^{I_1^j} B_{I_1^j}^j + \varepsilon_j \text{ with } \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2) \quad (17)$$

$$\forall j \notin I_2 : X^j = f(\theta_j) \text{ free law} \quad (18)$$

Where  $B_{I_1^j}^j$  is the  $p_1^j$ -sized vector of the coefficients of the subregression.

We can also use a matricial notation for the subregressions :

$$X^{I_2} = X^{I_2^c} B_{I_2^c}^{I_2} + \varepsilon_{I_2} \quad (19)$$

Where  $B_{I_2^c}^{I_2}$  is a  $(p - p_2) \times p_2$  matrix and  $\varepsilon_{X^{I_2}}$  a  $n \times p_2$  noise matrix. Column  $i$  :  $\varepsilon_X(i) \sim \mathcal{N}(0, d_i^2 I_n)$ . And  $\varepsilon_X(i) \perp \varepsilon_X(j), i \neq j$ .

Classical methods estimate  $Y|X$  and obtain (Maximum Likelihood Estimation):

$$\hat{A} = (X'X)^{-1} X'Y \text{ (ill-conditioned matrix to inverse)} \quad (20)$$

Classical model MSE :

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \underbrace{\sigma^2}_{small} \text{Tr}(I_n - \underbrace{X(X'X)^{-1}X'}_{ill-conditioned}) \frac{1}{n} = \frac{(n-p)\sigma^2}{n}. \quad (21)$$

Our idea is to first explain  $Y$  only with  $X^{I_1}$  :

$$Y = X^{I_1} \underbrace{(A_{I_1} + B_{I_1}^{I_2} A_{I_2})}_{\tilde{A}^{I_1}} + \underbrace{\varepsilon_{X^{I_2}} A_{I_2} + \varepsilon_Y}_{\tilde{\varepsilon}} = X^{I_1} \tilde{A}^{I_1} + \tilde{\varepsilon} \quad (22)$$

So the new estimator simply is :

$$\hat{\tilde{A}}^{I_1} = (X_{I_1}' X^{I_1})^{-1} X_{I_1}' Y \quad (23)$$

$$\hat{\tilde{A}}^{I_2} = 0 \quad (24)$$

and we get the following MSE :

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \left( \underbrace{\sigma^2 + \sum_{i=1}^{p_2} A_{I_2}^2(i) d_i^2}_{increased \text{ variance}} \right) \text{Tr}(I_n - \underbrace{X^{I_1} (X_{I_1}' X^{I_1})^{-1} X_{I_1}'}_{small \text{ and well-conditioned}}) \frac{1}{n} = \frac{(n-p_1)}{n} (\sigma^2 + \sum_{i=1}^{p_2} A_{I_2}^2(i) d_i^2). \quad (25)$$

both classical and our new estimators of  $Y$  are unbiased (true model)[11].

There is no theoretical guarantee that our model is better. It's just a compromise between numerical issues caused by correlations for estimation and selection versus increased variability due to

structural hypothesis. Therefore we made some simulations to compare both methods (see the end of this paper). 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. The explicit structure between the covariates helps to understand the model and the complex link between the covariate and the response variable so we call this model explicative.

When we use a variable selection method on it we obtain two kinds of 0 :

1. Because of the structure we coerce  $\hat{A}^{I_2} = 0$ . This kind of zero means redundant information but the covariate can be correlated with the response variable. So we don't have the grouping effect (so we are more parsimonious ) and we don't suffer from false interpretation (LASSO would).
2. Variable selection methods can lead to get some exact zeros in  $\hat{A}^{I_1}$ . This kind of zero means that implied covariate has no significant effect on the response variable. And because variables in  $X^{I_1}$  are orthogonal, we know that it is not misleading interpretation due to correlations.

We will see later how to improve this model.

## 2.1 Why grouping effect is not that wonderful

In industrial context, when a model explain why things go wrong, one will try to fix the problem. If  $X_1 = X_2 + e$  and we have the grouping effect, we will obtain a model like  $Y = aX_1 + aX_2$ . Then when one will try to modify Y he will modify one of the covariates and both will change so he won't get expected results. 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  $X_1 = X_2 + e$ . So you have more information and are able to decide better actions. With such a model, grouping effect is no more useful because when saying  $Y = 2aX_2$  AND  $X_1 = X_2 + e$  you clearly show that  $X_2$  is correlated with both  $Y$  and  $X_1$ . So it is possible to combine the advantages of grouping effect and selection just giving several equations. Each equation here is very simple so you don't really increase complexity of the model. Uncrossed model (nilpotent Z) guarantee to keep a simple structure easily interpretable.

## 3 Estimating subregressions

All our work is based on a linear structure between the covariates. Let's define it and a manner to estimate it.

### 3.1 Structure definition

We note  $Z \in \mathcal{M}(p, p)$  with  $Z_{i,j} = \mathbf{1}_{(B_{i,j} \neq 0)} \forall (i, j) \in I_1 \times I_2$ , and  $Z_{i,j} = 0$  elsewhere. Let's define  $\mathcal{Z}$  the ensemble of feasible structures.

$Z \in \mathcal{Z} \Leftrightarrow$ :

- $Z$  is binary
- $ZZ = 0$  ( $Z$  is not crossed)

So  $\mathcal{Z}$  is just the set of the binary square nilpotent matrices of size  $p$ .

$Z$  is an adjacency matrix and we know [2] that  $Z^p$  shows the number of paths of length  $p$  (linking  $p + 1$  vertices) and with only  $p$  vertices available, it implies to contain almost one cycle. So in fact, we only need  $Z^p = 0$  to guarantee we have no cycles but, for the moment, we don't want to obtain a model with too much levels of estimation (sequential estimation in two steps is already difficult so we don't want to add other layers on it). So we suppose that  $Z$  is nilpotent, meaning it does not contain any non-trivial path. This strong hypothesis also strongly reduces the size of  $\mathcal{Z}$ . Another reason to make this assumption is that long paths may lead to link independent covariates and to obtain models that are not identifiable and with difficulties to interpret them.

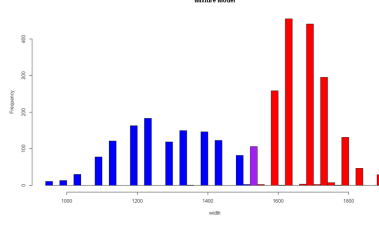


Figure 1: Industrial variables seem to follow gaussian mixture models

It directly implies (simple algebra) that  $\forall i : Z_{i,i} = 0$ . And we also have  $ZZ = 0 \Leftrightarrow \forall (i, j)$ , if  $Z_{i,j} = 1$  then  $\forall k$  we have  $Z_{j,k} = 0$  and  $Z_{k,i} = 0$  (no path of length  $> 1$ ). Thus variables on the left ( $X^{I_2}$ ) and variables on the right ( $X^{I_1}$ ) form a partition of  $X$ . We note  $p_1 = |I_1|$  and  $p_2 = |I_2|$ .  $B$  can be seen as the weight of each path.  $Z$  is the structure and  $B$  the associated parameter.

### 3.2 Structures comparison

$$\mathbf{P}(X) = \mathbf{P}(X^{I_1}, X^{I_2}) = \mathbf{P}(X^{I_2}|X^{I_1})\mathbf{P}(X^{I_1}) \quad (26)$$

When estimating the quality of the model we look at the likelihood (to compute the Bayesian Information Criterion (BIC)):

$$\mathcal{L}_B(X) = \prod_{\substack{1 \leq i \leq n \\ j \in I_2}} \mathbf{P}(X_{i,j}|X^{I_1}) \prod_{\substack{1 \leq i \leq n \\ j \in I_1}} \mathbf{P}(X_{i,j}) \quad (27)$$

Where  $X_{|X^{I_1}}^{I_2} \sim \mathcal{N}(X^{I_1} B_{I_1}^{I_2}, d^2 I_{n \times p_2})$  is Gaussian (it is the residual of the sub-regressions) and  $X^{I_1}$  follow gaussian mixture models (to better fit industrial variables) estimated separately (we use the Rmixmod package for R). In the following,  $\otimes$  and  $\oplus$  denote respectively the kronecker product and sum.

$$\forall j \in I_2 : X_{|X^{I_1}}^j \sim \mathcal{N}(X^{I_1} B_{I_1}^j, d_j^2) \quad (28)$$

$$\forall j \in I_1 : X^j \sim \mathcal{GM}(\pi_j; \mu_j; \sigma_j^2) \text{ with } \pi_j, \mu_j, \sigma_j^2 \text{ vectors of size } k_j \quad (29)$$

$$\text{And we obtain, } \forall j \in I_2 : X^j \sim \mathcal{GM}(\bigotimes_{\substack{i \in I_1 \\ Z_{i,j}=1}} \pi_i ; \bigoplus_{\substack{i \in I_1 \\ Z_{i,j}=1}} B_{i,j} \mu_i ; d_j^2 + \bigoplus_{\substack{i \in I_1 \\ Z_{i,j}=1}} B_{i,j}^2 \sigma_i^2) \quad (30)$$

So when we compare (29) and (30) we see that the number of classes in  $I_2$  variables differs when subregressions are of length  $> 1$  (almost 2 predictors) with multiple-class predictors (so the kronecker product is effective). We call this difference in component number "identifiability" in the sense that we try to find the model with the fewest component in  $X^{I_1}$  (we use the BIC so when comparing two models with the same likelihood, the one with lesser parameters will win). So in this case we have a "better" model (the simplest one) in the group of equivalent models.

Remark : the uncrossing constraint ( $I_2 \cap I_1 = \emptyset$ ) significantly reduces the number of feasible models and thus increases identifiability (it also reduces the number of models equivalent in interpretation).

And if we try to permute one subregression we obtain for  $k \in I_2$  and  $j \in I_1$ :

$$X^k - \varepsilon_X^k - \sum_{\substack{i \in I_1 \\ i \neq j}} X^i B_{i,k} \sim \mathcal{GM}(\tilde{\pi}_j; \tilde{\mu}_j; \tilde{\sigma}_j^2) \text{ where} \quad (31)$$

$$\tilde{\pi}_j = \left( \bigotimes_{i \in I_1} \pi_i \right) \otimes \bigotimes_{\substack{i \in I_1 \\ i \neq j}} \pi_i \quad (32)$$

$$\tilde{\mu}_j = \frac{(\bigoplus_{i \in I_1} B_{i,k} \mu_i) \oplus \bigoplus_{\substack{i \in I_1 \\ i \neq j}} (-B_{i,k} \mu_i)}{B_{j,k}} \quad (33)$$

$$\tilde{\sigma}_j^2 = \frac{2d_k^2 + (\bigoplus_{i \in I_1} B_{i,k}^2 \sigma_i^2) \oplus \bigoplus_{\substack{i \in I_1 \\ i \neq j}} B_{i,k}^2 \sigma_i^2}{B_{j,k}^2} \quad (34)$$

Thus if we call identifiability the ability to find the model with the smallest variance on the subregression residuals, we have identifiability on the structure if  $X^k - \varepsilon_X^k - \sum_{\substack{i \in I_1 \\ i \neq j}} X^i B_{i,k} \approx X^j$ . That is

the case if  $d_k^2 > 0$  (non-exact subregression) because the variance of the noise of the subregression is doubled when the subregression is permuted. So even with gaussian variables, non-exact subregression are a sufficient condition for identifiability (in the meaning of smallest-variance subregressions). But you need to look the marginal laws because the joint one relies on the structure and does not use marginal laws of the left-sided covariates.

If there are exact subregressions, classical methods will fail (singular matrix) and 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 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. Our R package prints a warning to point out exact subregressions when found.

Now we have made hypothesis on the distribution, we can use them to compare the structures with the Bayesian Information criterion (BIC) [8].

$$BIC(Z|X) = \sum_{j=1}^p BIC(Z^j|X^j) \quad (35)$$

Where  $BIC(Z^j|X^j)$  is defined this way :

$$BIC(Z^j|X^j) = -2\mathcal{L}(X^j, \theta_j) + K_j \log(n) \quad (36)$$

Where  $K_j$  is the number of parameters to estimate. We will now use the following notation :  $BIC(Z) = BIC(Z|X)$

If we have some hypothesis on the distribution of some variables (exponentially distributed for example) we can compute corresponding  $BIC$  separately and then improve the efficiency of the algorithm (it will find a structure only if it is really relevant).

### 3.3 The Markov chain

First we find that  $S$  is completely described with  $I_1$  :

$$I_2 = \{j | \#I_1^j > 0\} \quad (37)$$

$$p_2 = \#I_2 \quad (38)$$

$$\forall j p_1^j = \#I_1^j \quad (39)$$

So we will only describe the variations in  $I_1$  at each step and other parts of  $S$  will follow according to the previous definition. for each step, starting from  $S \in \mathcal{S}$  we define a neighbourhood  $\mathcal{V}_{S,j}$  with

$j \sim \mathcal{U}(\{1, \dots, p\})$  like this :

$$\begin{aligned} \mathcal{V}_{S,j} = \{ \tilde{S} \in \mathcal{S} \mid & \exists ! i, \tilde{Z}_{i,j} = 1 - Z_{i,j}, \text{ and } \forall (k,l) \neq (i,j) : \\ & \tilde{Z}_{j,l} = (1 - \tilde{Z}_{i,j}) Z_{j,l} \text{ (row-wise relaxation),} \\ & \tilde{Z}_{k,i} = (1 - \tilde{Z}_{i,j}) Z_{k,i} \text{ (column-wise relaxation)} \\ & \tilde{Z}_{k,l} = Z_{k,l} \} \cup \{Z\} \end{aligned}$$

We have  $|\mathcal{V}_{Z,j}| = p$  but some other constraints can be added on the definition of  $\mathcal{Z}$  and will consequently modify the size of the neighbourhood (for example a maximum complexity for the subregressions or the whole structure).

The algorithm follows a time-homogeneous markov chain whose transition matrix  $\mathcal{P}$  has  $|\mathcal{Z}|$  rows and columns (combinatory so we'll just compute the probabilities when we need them). And  $\mathcal{Z}$  is the finite state space.

We want

$$\mathcal{P}(Z, \tilde{Z}) = \mathbf{1}_{[\exists j, \tilde{Z} \in \mathcal{V}_{Z,j}]} P(\tilde{Z}|X) \quad (40)$$

Because the walk follows a regular and thus ergodic markov chain with a finite state space, it has exactly one stationary distribution [7] :  $\pi$  and every rows of  $\lim_{k \rightarrow \infty} \mathcal{P}^k = W$  equals  $\pi$ .

With  $\forall Z \in \mathcal{Z}$  :

$$0 \leq \pi(Z) \leq 1 \quad (41)$$

$$\sum_{Z \in \mathcal{Z}} \pi(Z) = 1 \quad (42)$$

$$\pi(Z) = \sum_{\tilde{Z} \in \mathcal{Z}} \pi(\tilde{Z}) \mathcal{P}(\tilde{Z}, Z) \quad (43)$$

$$= \sum_{\tilde{Z} \in \mathcal{Z}} \pi(\tilde{Z}) \mathcal{P}(\tilde{Z}, Z) \quad (44)$$

$$P(Z|X) = P(X|Z)P(Z) \propto P(X|Z) \quad (45)$$

We make a first approximation :

$$P(X|Z) \approx \exp(BIC(Z)) \quad (46)$$

We define [8], :

$$q(\tilde{Z}, \mathcal{V}_{Z,j}) = \mathbf{1}_{\{\tilde{Z} \in \mathcal{V}_{Z,j}\}} \frac{\exp(\frac{-1}{2} \Delta BIC(\tilde{Z}, \mathcal{V}_{Z,j}))}{\sum_{Z_l \in \mathcal{V}_{Z,j}} \exp(\frac{-1}{2} \Delta BIC(Z_l, \mathcal{V}_{Z,j}))} \quad (47)$$

where  $\Delta BIC(Z, \mathcal{V}_{Z,j}) = BIC(Z) - \min\{BIC(\tilde{Z}) | \tilde{Z} \in \mathcal{V}_{Z,j}\}$  is the gap between a structure and the worst structure in the neighbourhood in terms of BIC.

And then we can note  $\forall (Z, \tilde{Z}) \in \mathcal{Z}^2$  :

$$\mathcal{P}(Z, \tilde{Z}) = \frac{1}{p} \sum_{j=1}^p q(\tilde{Z}, \mathcal{V}_{Z,j})$$

The output will be the best structure seen in terms of BIC. If we have some knowledge about some sub-regressions (physical models for example) we can add them in the found structure. So the model is really expert-friendly.

Because the walk follows a regular and thus ergodic markov chain with a finite state space, it has exactly one stationary distribution [7] :  $\pi$  and every rows of  $\lim_{k \rightarrow \infty} \mathcal{P}^k = W$  equals  $\pi$ .



With  $\forall Z \in \mathcal{Z}$  :

$$0 \leq \pi(Z) \leq 1 \quad (48)$$

$$\sum_{Z \in \mathcal{Z}} \pi(Z) = 1 \quad (49)$$

$$\pi(Z) = \sum_{\tilde{Z} \in \mathcal{Z}} \pi(\tilde{Z}) \mathcal{P}(\tilde{Z}, Z) \quad (50)$$

$$= \sum_{\tilde{Z} \in \mathcal{Z}} \pi(\tilde{Z}) \frac{1}{p} \sum_{j=1}^p q(Z, \mathcal{V}_{\tilde{Z},j}) \quad (51)$$

$$= \sum_{\tilde{Z} \in \mathcal{Z}} \pi(\tilde{Z}) \frac{1}{p} \sum_{j=1}^p \mathbf{1}_{\{Z \in \mathcal{V}_{\tilde{Z},j}\}} \frac{\exp(\frac{-1}{2} \Delta BIC(Z, \mathcal{V}_{\tilde{Z},j}))}{\sum_{Z_l \in \mathcal{V}_{\tilde{Z},j}} \exp(\frac{-1}{2} \Delta BIC(Z_l, \mathcal{V}_{\tilde{Z},j}))} \quad (52)$$

$$= \sum_{\tilde{Z} \in \mathcal{Z}} \pi(\tilde{Z}) \frac{1}{p} \sum_{j=1}^p \mathbf{1}_{\{Z \in \mathcal{V}_{\tilde{Z},j}\}} \frac{\exp(\frac{-1}{2} BIC(Z))}{\sum_{Z_l \in \mathcal{V}_{\tilde{Z},j}} \exp(\frac{-1}{2} BIC(Z_l))} \quad (53)$$

The initial structure is based on a first warming algorithm taking the correlations into account. Ones are randomly placed into  $Z$ , weighted by the absolute value of the correlations. Then this structure is reduced by the hadamard product with the binary matrix obtained by Graphical Lasso[6].

## 4 Further use of the subregression structure

At this step we only used the partition implied by the structure (who explains and who is explained) but not the structure  $B$  itself. The subregression structure is full of information and we want to use it.

### 4.1 Noise reduction

Because the structure is not exact and contains some noise (noted  $\varepsilon_{X^{I_2}}$ ), we lost some information using only  $X^{I_1}$ . So we use the structure to take back this information into the model.

#### 4.1.1 Sequential approach

It's just about estimating the part of the residuals due to our hypothesis.

Knowing (22) we can estimate :

$$\hat{\varepsilon} = Y - X^{I_1} \hat{A}_{I_1} \quad (54)$$

and knowing (??) we can estimate :

$$\hat{\varepsilon}_{X^{I_2}} = X^{I_2} - X^{I_1} \hat{B}_{I_1}^{I_2} \quad (55)$$

Then we obtain a classical linear model :

$$\tilde{\varepsilon} = \varepsilon_{X^{I_2}} A_{I_2} + \varepsilon_Y \quad (56)$$

So we have :

$$\hat{A}_{I_2} = (\hat{\varepsilon}_{X^{I_2}}' \hat{\varepsilon}_{X^{I_2}})^{-1} \hat{\varepsilon}_{X^{I_2}}' \hat{\varepsilon} \quad (\text{OLS estimation}), \text{ and because of (22):} \quad (57)$$

$$\hat{A}_{I_1} = \hat{A}_{I_1} - \hat{B}_{I_1}^{I_2} \hat{A}_{I_2} \quad (\text{Identification}), \text{ and then :} \quad (58)$$

$$\hat{Y} = X^{I_1} \hat{A}_{I_1} + X^{I_2} \hat{A}_{I_2} \quad (\text{called the predictive model}) \quad (59)$$

When time comes to explain interpret the model, we come with different and meaningful kinds of zeros. Two kinds from the reduced model :

- 0 in  $\tilde{A}_{I_1}$  mean that the variable have no effect on  $Y$

- 0 in  $\tilde{A}_{I_2}$  are coerced to 0 by the structure and mean we will estimate them later (predictive model).

and three kinds from the predictive model :

- 0 in  $A_{I_2}$  mean that knowing  $X^{I_1}$ , the associated variable in  $X^{I_2}$  have no effect on  $Y$  (no specific effect).  $X^{I_2}$  can be correlated to the response but is only due to its correlation with  $X^{I_1}$ .
- 0 in  $\tilde{A}_{I_1}$  that becomes non-zero in  $A_{I_1}$  mean that only the associated  $\varepsilon_{X^{I_2}}$  have an effect on  $Y$  but corresponding variables in both  $X^{I_1}$  and  $X^{I_2}$  don't. So it kind of points a latent variable (that is  $\varepsilon_{X^{I_2}}$ ).
- 0 in  $A_{I_1}$  that were non-zero in  $\tilde{A}_{I_1}$  mean that the effect of corresponding variable in  $X^{I_1}$  on the response is entirely contained in the effect of  $X^{I_2}$ .

Each model can use any estimator and selection method. They are only linear models but we consider them together and estimate them sequentially.

#### 4.1.2 Constrained approach

The idea is the following. After computing the explicative model, we have an estimation of the value of  $\tilde{A}_{I_1}$  and we know that  $\tilde{A}_{I_1} = A_{I_1} + B_{I_1}^{I_2} A_{I_2}$ . So we just have to resolve

$$\min_A (\|Y - XA\|^2) \quad \text{constrained by} \quad (60)$$

$$\hat{\tilde{A}}_{I_1} = A_{I_1} + B_{I_1}^{I_2} A_{I_2} \quad (61)$$

We can either use  $\hat{B}$  (faster) or compute a new estimation of  $B$  with alternated optimisation (optimal). Besides being faster, reusing  $\hat{B}$  allows to have the same values within two studies on the same dataset but with different response variables. Hence, it is easier to explain for the final users (metallurgists in our case).

The constraint can be noted  $B_{I_1} A = \hat{\tilde{A}}_{I_1}$ . Whith  $B_{I_1}^{I_1}$  the identity matrix,  $B_{I_1}^{I_2}$  the structure, and  $B_{I_2} = 0$  (since  $X = XB$ ).

Equality Constrained Least squares [1] estimator (noted  $\bar{A}$ ) can be written in two ways :

$$\bar{A} = \hat{A}_{OLS} - (X'X)^{-1} B_{I_1}' (B_{I_1} (X'X)^{-1} B_{I_1}')^{-1} (B_{I_1} \hat{A}_{OLS} - \hat{\tilde{A}}_{I_1}) \quad (62)$$

but this formulation is only valid for  $(X'X)$  non-singular. so if  $n < p$  we can use the more general (and equivalent for  $n \geq p$ ) form :

$$\bar{A} = R(R'X'XR)^{-1} R'XY + [I_p - R(R'X'XR)^{-1} R'X'X] B_{I_1}' (B_{I_1} B_{I_1}')^{-1} \hat{\tilde{A}}_{I_1} \quad (63)$$

with  $R$  such that  $(B_{I_1}', R)$  is not singular and  $R'B_{I_1}' = 0$ .  $R$  is not unique and can always be found (but how ???? don't want to compute gaussian elimination). In practice, a fast approximation is to use the first formula with Moore-Penrose pseudo-inverse [10] when confronted to  $(X'X)^{-1}$ . This approximation is implemented in the CorReg package.

#### 4.1.3 Variable selection

Variable selection in the naive explicative model shows independence between some covariates in  $X^{I_1}$  and the response variable  $Y$ .

Variable selection in noise reduction helps to take back only relevant information (relevant for  $Y$  and not already explained by the reduced model).

Identification phase may not be sufficient to find zeros in  $A_{I_1}$  that where non-zeros in  $\tilde{A}_{I_1}$  because it is not an estimation but just an identification based on two different estimations (Cases where  $Y$  is explained with only  $X^{I_2}$ ). So we need a selection step here. We define a new model to replace simple identification.

$$\hat{\tilde{Y}} = Y - X^{I_2} \hat{A}_{I_2} = X^{I_1} A_{I_1} + \varepsilon \quad (64)$$

This model is also a basic linear regression and we can use any variable selection method to estimate its parameter and find zeros in it.

After this last selection step, we know which covariates to use in both  $X^{I_1}$  and  $X^{I_2}$  but this hypothesis was not the same when estimating the values of non-zero parameters. So we add a final estimation phase in two step to obtain final values of the parameter [16]. It only consists in computing both explicative and predictive models but without the covariates rejected by the selection step .

We note  $I_1^*$  the set of covariates associated with non-zeros in  $\hat{A}_{I_1}$  and  $I_2^*$  the set of covariates associated with non-zeros in  $\hat{A}_{I_2}$

$$\hat{A}_{I_1^*} = (X_{I_1^*}' X_{I_1^*}^{I_1^*})^{-1} X_{I_1^*}' Y \quad (65)$$

$$\hat{\varepsilon}^* = Y - X_{I_1^*}^{I_1^*} \hat{A}_{I_1^*} \quad (66)$$

$$\hat{\varepsilon}_{X^{I_2^*}} = X^{I_2^*} - X^{I_1} \hat{B}_{I_1}^{I_2^*} \quad (67)$$

$$\hat{A}_{I_2^*} = (\hat{\varepsilon}_{X^{I_2^*}}' \hat{\varepsilon}_{X^{I_2^*}})^{-1} \hat{\varepsilon}_{X^{I_2^*}}' \hat{\varepsilon}^* \quad (68)$$

$$\hat{A}_{I_1^*} = \hat{A}_{I_1^*} - \hat{B}_{I_1}^{I_2^*} \hat{A}_{I_2^*} \quad (69)$$

Predictive model is asymptotically the same as classical one (we also use all the variables) but it is sequential and relies on the results of the reduced model so with a finite dataset it is a compromise between the cost of correlations and the cost of this sequential aspect. One more time, only numerical results can say which model is the best. For the constrained predictive model we need constrained selection methods. A paper called "The Constrained LASSO" was recently submitted by Gareth JAMES *et al.* with the description of an algorithm for equality constrained LASSO but we have not yet implemented it to test its efficiency.

## 4.2 Missing values

Industrial datasets often contains missing values (see Figure 2). Let  $M$  (for Missing) be the binary  $n \times p$  matrix that indicates where the missing values are (that allows sparse storage). We have :

$$M_{i,j} = \mathbf{1}_{[X_{i,j}=\emptyset]} \quad (70)$$

We suppose here that  $Z$  has already been learnt. We then try to use this structure to deal with missing values using all information we have.

$\forall(i, j), M_{i,j} = 1$  : we have 4 different situations :

1. if  $j \in I_2$  and  $M_i Z^j = 0$  (no missing values for this individual in the dependent variables). Can also be written  $\sum_{k=1}^p M_{i,k} Z_{k,j} = 0$ . We can directly use the subregression to obtain  $\hat{X}_{i,j} = X_i^{I_1} B_{I_1}^j$ . But it has no utility : the explicative model does not depend on  $X^{I_2}$  and the noise estimated with imputed value will be null so it can not be reduced.
2. if  $j \in I_1$  and  $\forall 1 \leq k \leq p, Z_{j,k} = 0$  then the variable is totally independent so we just use density hypothesis (29).
3. if  $j \in I_2$  and  $M_i \times Z^j \neq 0$  we can't directly use the structure because of missing values so we will use a sem-like algorithm, making imputations on  $X^{I_1}$  like explained in the previous case and then using the structure to estimate  $X_{i,j}$ . The idea is just to do so many times and then take the mean of all imputations for  $X_{i,j}$ .
4. if  $j \in I_1$  and  $\exists k, Z_{j,k} \neq 0$  we have to choose between several methods (using cross-validation for example).

- Estimate the subregression given by the line  $j$  of  $Z$  (may imply correlated variables). On the largest dataset without missing values (if there is one).

- Permute subregression that imply  $X^j$ . If  $Z_{j,k} \neq 0$  then

$$\hat{X}_{i,j|X} = -\frac{1}{B_{j,k}} \left[ -X_{i,k} + \sum_{\substack{l \in I_1 \\ l \neq j}} X_{i,l} B_{l,k} \right] \quad (71)$$

If not possible to obtain full datasets for any of these models, density hypothesis remains. At the opposite, if several subregressions are implied, we can choose the best model (e.g. : cross-validation) or compute the mean of the estimated values.

Missing values imputation aims to keep more individuals in the dataset. This approach can fill every blanks but it is also possible to only fill a part of the dataset, keeping away individuals with too many missing values where imputation are less reliable. It may be better to delete some individuals than to make too many imputations (increased variability).

#### 4.2.1 Dealing with missing values in the MCMC algorithm

If there are missing values in the dataset, then the likelihood (27) just becomes

$$\mathcal{L}_Z(B; X) = \prod_{\substack{1 \leq i \leq n \\ j \in I_2 \\ M_{i,j}=0}} \mathbf{P}(X_{i,j}|X^{I_1}, B) \prod_{\substack{1 \leq i \leq n \\ j \in I_1 \\ M_{i,j}=0}} \mathbf{P}(X_{i,j}|B) \quad (72)$$

with,  $\forall(i, j), 1 \leq i \leq n, j \in I_2$  and  $M_{i,j} = 0$  :

$$X_{i,j|X^{I_1}, B} = \sum_{\substack{k \in I_1 \\ Z_{k,j}=1 \\ M_{i,k}=0}} x_{i,k} B_{k,j} + \varepsilon_X^j + \sum_{\substack{k \in I_1 \\ Z_{k,j}=1 \\ M_{i,k}=1}} X_{i,k} B_{k,j} \quad (73)$$

$$X_{i,j|X^{I_1}, B} \sim \mathcal{GM} \left( \bigotimes_{\substack{k \in I_1 \\ Z_{k,j}=1 \\ M_{i,k}=1}} \pi_k ; \left( \sum_{\substack{k \in I_1 \\ Z_{k,j}=1 \\ M_{i,k}=0}} x_{i,k} B_{k,j} \right) + \bigoplus_{\substack{k \in I_1 \\ Z_{k,j}=1 \\ M_{i,k}=1}} B_{k,j} \mu_k ; d_j^2 + \bigoplus_{\substack{k \in I_1 \\ Z_{k,j}=1 \\ M_{i,k}=1}} B_{k,j}^2 \sigma_k^2 \right) \quad (74)$$

So we just have to maximize this explicit likelihood on observed values to be able to compare different models. To achieve this maximisation, we use a SEM-like algorithm (Stochastic Expectation-Maximization) [4] for  $B$  and use Mixmod (only once) to estimate the parameters of marginal gaussian mixtures (always the same).

The algorithm is as follows :

- For each step ( $k$ ) :
  - Generate imputations for missing values in  $X^{I_1}$  based on the gaussian mixture hypothesis (obtained once by Mixmod)
  - Use the completed dataset to estimate the parameters of the density ( $\hat{B}^{(k)}$  and  $\hat{d}^{(k)}$ ).
- After a fixed number  $N$  of steps,  $\hat{B} = \frac{1}{N} \sum_{k=1}^N \hat{B}^{(k)}$  and  $\hat{d} = \frac{1}{N} \sum_{k=1}^N \hat{d}^{(k)}$
- compute the likelihood of the model based on these parameters.

Estimated densities are just gaussian mixtures so it's easy to generate values on it. We just have to randomly choose a component (according to the weights of each component) and then generate a value from a simple gaussian law.

To compare different models, we want to penalize complexity (like BIC does) but here we can't use the BIC because for one variable, different individuals follows different conditional laws with different complexity. A feasible penalty would be to use a modified BIC penalty with the number of parameters weighted by the number of individuals implied. So for a dataset with  $n_1$  individuals following a law with  $K_1$  parameters and the others  $n_2$  a law with  $K_2$  parameters, we would take  $k = \frac{n_1 K_1 + n_2 K_2}{n}$  and then  $k$  may not be an integer.

### 4.2.2 Imputation or deletion

Using SEM to estimate leads to make imputations on the missing values. Because we use the mean parameter after many imputations, we don't have bias problems but it still introduce some variance in the model. If  $n$  is strongly greater than  $p$ , and if missing values represents only few individuals, it can be better to just delete incomplete individuals. But sometimes, even with a small number of missing values, many individuals are incomplete. Just imagine a dataset with one missing value for each individual (but on different variables). Then you can't reduce the dataset (it would become empty) so the question is not just to decide if imputation is a good idea but also to see if it is necessary. Great variability remains better than no result at all.

## 5 Numerical results on simulated datasets

Here are some results on simulated datasets.

### 5.1 Finding the structure

Tableau de la forme :

n	time	trueBIC	BICempty	BIC_opt	True1	False1	missing1	$\Delta p2$	True_left	False_left
40	??	??	??	??	??	??	??	??	??	??
60	??	??	??	??	??	??	??	??	??	??
100	??	??	??	??	??	??	??	??	??	??

Table 1:  $p$  variables. Markov chain was XX seconds long for  $n = 100$ (mean observed).

Ordre des critères de comparaison : MSE sur X, Vraigauche, fauxgauche, bics (les 3), vrais1, faux1, missing1,deltap2

L'idée serait de n'avoir qu'une seule configuration (vu qu'ici on ne dépend pas de Y) et garder la même pour tous les tableaux suivants (pour pouvoir s'appuyer sur celui-ci dans l'interprétation). Tous les tableaux seraient générés en même temps. Pour chaque base générée, on génèrerait plusieurs Y de plusieurs manières pour avoir tous les cas sur les mêmes données. La parallélisation des expériences se ferait alors sur le nombre de réplifications. Les résultats seraient toujours basés sur Zchapeau (et donc Bchapeau).

On devrait constater que quand l'algo a le temps de converger, on trouve pour  $n$  petit des BICs meilleurs que le vrai modèle (d'où un bruit sur la structure). quand  $n$  augmente, ce surapprentissage devrait disparaître et on devrait donc converger vers le vrai Z (et le vrai BIC).

### 5.2 Y depends only on some covariates in $X^{I_1}$

#### 5.2.1 without selection

n	OLS	(sd)	explicative	(sd)	predictive	(sd)
40	NA	NA	??	??	??	??
60	??	??	??	??	??	??
100	??	??	??	??	??	??

Table 2:  $Y$  only depends on  $X^{I_1}$ .  $p = 50$  and  $\text{Var}(Y) \simeq 3.10^8$

On doit constater qu'on est meilleurs que OLS, que l'explicatif gagne (vrai modèle possible) mais que le prédictif reste bon. On doit aussi voir que quand  $n$  grandit OLS commence à redevenir correct.

#### 5.2.2 with selection

Avec le même Y que pour le cas sans sélection, (et les mêmes données) on teste simplement d'autres modèles :

- package lm
- modèle complet avec lasso (et LAR)
- modèle complet avec elastic net
- modèle complet ridge
- modèle explicatif elastic net
- modèle prédictif elastic net
- modèle explicatif LASSO
- modèle prédictif LASSO

Il y aurait un tableau par valeur de  $n$  pour pouvoir donner en plus des MSE des valeurs de sparsité et de validité du modèle (comparaison des positions des 0). Je n'ai pas pour l'instant de quoi utiliser CLERE mais l'article CLERE montre qu'elastic net est meilleur en prédiction donc pour les MSE ce n'est pas trop un problème.

### 5.3 $Y$ depends only on some covariates in $X^{I_2}$

même chose qu'avant mais on part avec un handicap. les notions d'explicatif et prédictif finaux devraient alors prendre tout leur sens.

### 5.4 global case

$Y$  dépend un peu de tout le monde... me semble trop compliqué car beaucoup trop de cas possibles. la conclusion étant de toute manières qu'on sera quelque part entre les deux cas précédents. Je mettrais bien des exemples simples et poussés (3 variables explicatives comme dans l'article sur la consistance du lasso) pour que les gens puissent facilement refaire le test chez eux, même sans notre package (hypothèse du vrai  $Z$ ). la simplicité de l'exemple permettrait aussi de voir ce qui se passe si  $Z$ chapeau est une version permutée du vrai  $Z$ . On testerait là aussi tous les modèles concurrents abordés plus haut.

Attention : on a une variabilité due à la validation croisée. Sur les mêmes données, quand on lance plusieurs fois la sélection on ne trouve pas toujours exactement les mêmes 0 (tout de même relativement stable, peut s'arranger en choisissant un meilleur  $K$  pour la validation croisée).

### 5.5 Consistency Issues

Consistency issues of the LASSO are well known and Zhao [17] gives a very simple example to illustrate it. We have taken the same example to show how our method is more consistent. Here  $p = 3$  and  $n = 1000$ . We define  $X_1, X_2, \varepsilon_Y, \varepsilon_X i.i.d. \sim \mathcal{N}(0, 1)$  and then  $X_3 = \frac{2}{3}X_1 + \frac{2}{3}X_2 + \frac{1}{3}\varepsilon_X$  and  $Y = 2X_1 + 3X_2 + \varepsilon_Y$ . We compare consistencies of complete, explicative and predictive model with LASSO (and LAR) for selection. It happens that the algorithm don't find the true strucure but a permuted one so we look at the results obtained with the true  $Z$  (but  $\hat{B}$  is used) and with the structure found by the Markov chain after a few seconds.

True  $Z$  is found 340 times on 1000 tries.

	Classical LASSO	Explicative	Predictive
True $Z$	1.006479	<b>1.005468</b>	<b>1.006093</b>
$\hat{Z}$	<b>1.006479</b>	1.884175	1.006517

Table 3: MSE observer on a validation sample (1000 individuals)

We observe as we hoped that explicative model is better when using true  $Z$  (coercing real zeros) and that explicative with  $\hat{Z}$  is penalized (coercing wrong coefficients to be zeros). But the main point

is that the predictive model stay better than the classical one with the true  $Z$  and corrects enough the explicative model to follow the classical LASSO closely when using  $\hat{Z}$ . And when we look at the consistency :

	Classical LASSO	Explicative	Predictive
True $Z$	0	1000	830
$\hat{Z}$	0	340	<b>621</b>

Table 4: number of consistent model found ( $Y$  depending on  $X_1, X_2$  and only them) on 1000 tries

299 times on 1000 tries, the predictive model using  $\hat{Z}$  is better than classical LASSO in terms of MSE and consistent (classical LASSO is never consistent).

We also made the same experiment but with  $X_1, X_2$  (and consequently  $X_3$ ) following gaussian mixtures (to improve identifiability) randomly generated by our R package. True  $Z$  is now found 714 times on 1000 tries . So it confirms that non-gaussian models are easier to identify.

	Classical LASSO	Explicative	Predictive
True $Z$	1.571029	<b>1.569559</b>	<b>1.570801</b>
$\hat{Z}$	1.005402	1.465768	<b>1.005066</b>

Table 5: MSE observer on a validation sample (1000 individuals)

And when we look at the consistency :

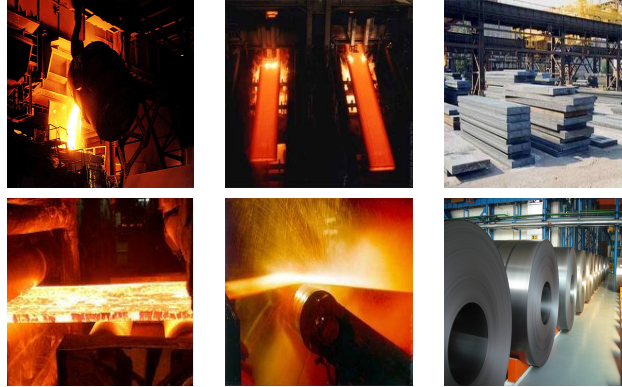
	Classical LASSO	Explicative	Predictive
True $Z$	0	1000	789
$\hat{Z}$	0	714	<b>608</b>

Table 6: number of consistent model found ( $Y$  depending on  $X_1, X_2$  and only them) on 1000 tries

299 times on 1000 tries, the predictive model using  $\hat{Z}$  is better than classical LASSO in terms of MSE and consistent (classical LASSO is never consistent).

## 6 Numerical results on real datasets

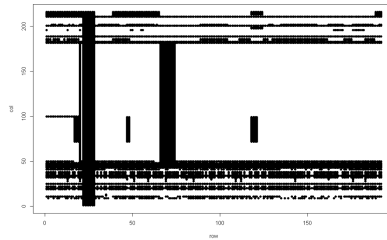
This work takes place in steel industry context, with quality oriented objective : to understand and prevent quality problems on finished product, knowing the whole process. In particular, we focus on regression problems.



Industrial context often means specific issues :

- Highly correlated parameters (parameters depends on the targeted product, physical models,...).
- Sometimes more variables than individuals.
- Missing values

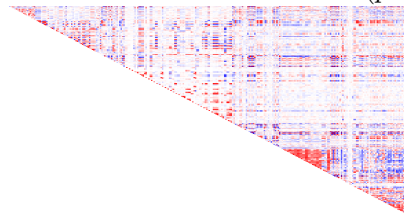
Figure 2: Missing values on a real industrial dataset



### 6.1 The dataset

- variables from the whole process
- The stakes : hundreds euros per ton (for information: Dunkerque site produces up to 7.5 millions tons a year)

Figure 3: Correlation matrix of the dataset ( $p = 293$ ,  $n = 3000$ )



Some observed correlations with physical meaning :

- Width and Weight : 0.905



- Temperature before and after a tool : 0.983
- Roughness of both faces : 0.919
- Mean & Max of a curve : 0.911

The method was tested on 205 variables without missing values.

## 6.2 Results

The algorithm gives a structure with 82 subregressions with a mean of 5 regressors. Some found subregressions with physical meaning :

- Mean.weight = f (Min.weight , Max.weight , Sigma.weight ) and other same-shaped subregressions.
- Width = f (Mean.flow , Mean.speed.CC)

True Physical model (not linear) :

- Width = flow / (speed \* thickness) (thickness is constant)

Some of the other subression represent physical models used to regulate the process and that were forgotten by the metallurgist we worked with. Found model has selected relevant variables (verified with metallurgist).

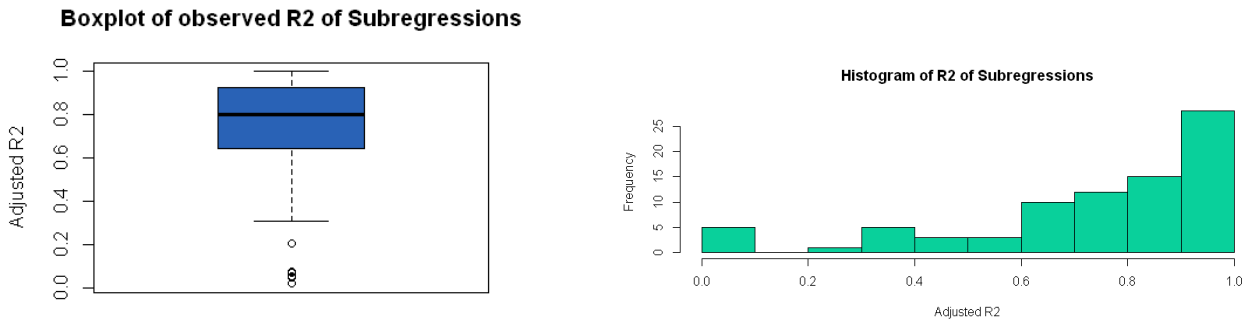


Figure 4: Adjusted R2 of found subregressions (industrial dataset)

We used Elastic Net[18] on this dataset for selection (get better results than LASSO). Here are the observed MSE on a  $n = 847$  validation sample. Predictive model (sequential elastic net base on estimated structure and using all the variables) is 5,82% better (Figure 6.2) than elastic net computed on the whole dataset.

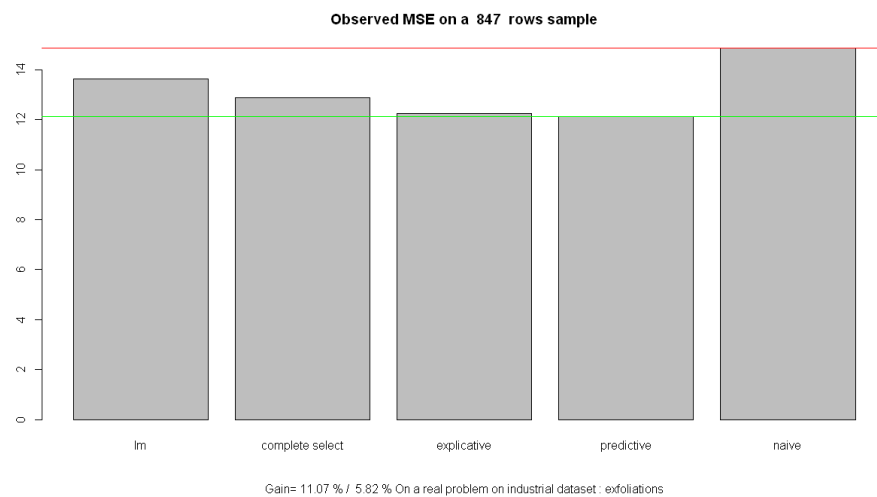


Figure 5: MSE comparison on industrial dataset

## 7 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 interpret 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 the method only works with numerical values. Further work is needed to face these two challenges.

## References

- [1] Takeshi Amemiya. *Advanced econometrics*. Harvard university press, 1985.
- [2] Norman Biggs. *Algebraic graph theory*. Cambridge University Press, 1993.
- [3] R. Davidson and J.G. MacKinnon. Estimation and inference in econometrics. *Oxford University Press Catalogue*, 1993.
- [4] J Diebolt and EHS Ip. A stochastic em algorithm for approximating the maximum likelihood estimate. *Markov chain Monte Carlo in practice*, 1996.
- [5] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
- [6] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
- [7] Charles Miller Grinstead and James Laurie Snell. *Introduction to probability*. American Mathematical Soc., 1997.
- [8] É. Lebarbier and T. Mary-Huard. Une introduction au critère bic: fondements théoriques et interprétation. *Journal de la SFdS*, 147(1):39–57, 2006.
- [9] D.W. Marquardt and R.D. Snee. Ridge regression in practice. *American Statistician*, pages 3–20, 1975.
- [10] R. Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51:406–413, 7 1955.
- [11] G. Saporta. *Probabilités, analyse des données et statistique*. Editions Technip, 2006.
- [12] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [13] N.H. Timm. *Applied multivariate analysis*. Springer Verlag, 2002.
- [14] Loic Yengo, Julien Jacques, Christophe Biernacki, et al. Variable clustering in high dimensional linear regression models. 2012.
- [15] 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.
- [16] Yongli Zhang and Xiaotong Shen. Model selection procedure for high-dimensional data. *Statistical Analysis and Data Mining*, 3(5):350–358, 2010.
- [17] Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, December 2006.
- [18] 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.