CorReg: Linear regression with highly correlated covariates

Clément THERY

June 10, 2014

To my sons, to my great friend Quentin Grimonprez, contributor of CorReg. I do not thank Vincent Kubicki, he does not play kazoo with me and said that my latex installation was rotten.

Contents

Abstract

Acknowledgments

Chapter 1

The industrial context

This work takes place in a steel industry context. The main objective is to be able to solve quality crisis when they occur. In such a case, a new type of unknown quality issue is observed and we have no idea of its origin. The defects, even generated at the beginning of the process, are often detected in its last part. The steel-making process includes several sub-process, each implying a whole manufactory. Thus we have many covariates and no a priori on the relevant ones. Moreover, the values of each covariates essentially depends on the characteristics of the final product, and many physical laws and tuning models are implied in the process. Therefore the covariates are highly correlated. We have several constraints :

- To be able to predict the defect and stop the process as early as possible to gain time (and money)
- To be able to understand the origin of the defect to try to optimize the process
- To be able to find parameters that can be changed because the objective is not only to understand but to correct the problematic part of the process.
- It also must be fast and automatic (without any a priori).

We will see in the state of the art that correlations are a real issue and that the number of variables increases the problem. The stakes are very high because of the high productivity of the steel plants but also because steel making is now well-known and optimized thus new defects only appears on innovative steels with high value. Any improvement on such crisis can have important impact on the market shares and when the customer is implied, each day won by the automation of the data mining process can lead to a gain of hundreds of thousands of euros, sometimes more. So we really need a kind of automatic method, able to manage the correlations without any a priori and giving an easily understandable and flexible model.

Chapter 2

State of the art

In the following we note classical norms: $\| \boldsymbol{\beta} \|_2^2 = \sum_{i=1}^{p} (\beta_i)^2$, $\| \boldsymbol{\beta} \|_1 = \sum_{i=1}^{p} |\beta_i|$ and $\| \boldsymbol{\beta} \|_\infty = \max(|\beta_1|, \ldots, |\beta_p|)$.

2.1 Ordinary least squares and associated problems

Linear regression is defined by this simple equation:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_Y \tag{2.1}$$

where $\boldsymbol{Y} \in \mathbf{R}^n$ is the response variable vector observed on n individuals.

2.2 Penalized models

2.2.1 Ridge regression

Ridge regression [Marquardt and Snee, 1975] proposes a biased estimator that can be written in terms of a parametric L_2 penalty:

$$\hat{\boldsymbol{\beta}} = \text{argmin} \left\{ \| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \|_2^2 \right\} \text{ subject to } \| \boldsymbol{\beta} \|_2^2 \leq \lambda \text{ with } \lambda > 0 \tag{2.2}$$

But this penalty is not guided by the correlations. It is the same for each covariates and will be too large for independent covariates and/or too small for correlated ones. So the efficiency of such a method is limited. Moreover, coefficients tend to 0 but don't reach 0 so it gives difficult interpretations for large values of p.

2.2.2 LASSO: Least Absolute Shrinkage and Selection Operator

[Tibshirani et al.,] [Tibshirani, 1996] [Efron et al., 2004] [Zhao and Yu, 2006][Zhang and Shen, 2010]The Least Absolute Shrinkage and Selection Operator (LASSO [Tibshirani, 1996]) consists in a shrinkage of the regression coefficients based on a λ parametric L_1 penalty to obtain zeros in $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = \text{argmin} \left\{ \| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \|_2^2 \right\} \text{ subject to } \| \boldsymbol{\beta} \|_1 \leq \lambda \text{ with } \lambda > 0 \tag{2.3}$$

The Least Angle Regression (LAR [Efron et al., 2004]) 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. But like the ridge regression, the penalty does not distinguish correlated and independent covariates so there is no guarantee to have less correlated covariates.

2.2.3 Adaptive LASSO and Random LASSO

[Zou, 2006][Wang et al., 2011]Some recent variants of the LASSO do exist for the choice of the penalization coefficient like the adaptive LASSO [Zou, 2006] or the random LASSO [Wang et al., 2011]. But LASSO also faces consistency problems [Zhao and Yu, 2006] when confronted with correlated covariates.

2.2.4 Elasticnet

[Zou and Hastie, 2005] Elastic net [Zou and Hastie, 2005] is a method developed to be a compromise between Ridge regression and the LASSO:

$$\hat{\boldsymbol{\beta}} = (1+\lambda_2)\,\text{argmin}\left\{\parallel \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \parallel_2^2\right\},\ \text{subject to } (1-\alpha) \parallel \boldsymbol{\beta} \parallel_1 + \alpha \parallel \boldsymbol{\beta} \parallel_2^2 \leq t \text{ for some } t \qquad (2.4)$$

where $\alpha = \frac{\lambda_2}{(\lambda_1+\lambda_2)}$. But it is based on the grouping effect so correlated covariates get similar coefficients and are selected together whereas LASSO will choose between one of them and will then obtain same predictions with a more parsimonious model. Once again, nothing specifically aims to reduce the correlations.

2.2.5 OSCAR: Octogonal Shrinkage and Clustering Algorithm for Regression

Like elasticnet, OSCAR [Bondell and Reich, 2008] uses combination of two norms for its penalty. Here the objective is to group covariates with the same effect (by a pairwise L_∞ norm) and give them exactly the same coefficient (reducing the dimension) with a simultaneous variable selection (implied by the L_1 norm).

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \parallel \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \parallel_2^2 \ \text{ subject to } \sum_{j=1}^{p} |\beta_j| + c\sum_{j<k} \max(|\beta_j|, |\beta_k|) \leq \lambda \qquad (2.5)$$

But OSCAR depends on two tuning parameters: c anf λ. For a fixed c the λ can be found by the LAR algorithm but c still has to be found "by hand" comparing final models for many values of c. Correlations are only implicitely taken into account and only pairwise. So it lacks of an efficient algorithm and need a supplementary study to interpret the groups found.

2.3 Modeling the parameters

2.3.1 CLERE: CLusterwise Effect REgression

[Yengo et al., 2012]The CLusterwise Effect REgression (CLERE [Yengo et al., 2012]) describes the β_j no longer as fixed effect parameters but as unobserved independant random variables with grouped β_j following a Gaussian Mixture distribution. The idea is to hope that the model have a small number of groups of covariates and that the mixture will have few enough components to have a number of parameters to estimate significantly lower than p. In such a case, it improves interpretability and ability to yeld reliable prediction with a smaller variance on $\hat{\boldsymbol{\beta}}$.

2.3.2 Spike and Slab

[Ishwaran and Rao, 2005]Spike and Slab variable selection [Ishwaran and Rao, 2005] 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.

2.4 Multiple Equations

2.4.1 SEM and Path Analysis

2.4.2 SUR: Seemingly Unrelated Regression

[Zellner, 1962]

2.4.3 SPRING: Structured selection of Primordial Relationships IN the General linear model

[Chiquet J. and S., 2013]

2.4.4 Selvarclust: Linear regression within covariates for clustering

[Maugis et al., 2009] The idea is to allow covariates to have different roles : (S, R, U, W). But:

- It is about clustering and not regression (not the same application field)
- No sub-regression allowed between relevant variables (in the True model)
- Using stepwise-like algorithm without protection against correlations [Raftery and Dean, 2006] even it is known to be often unstable [Miller, 2002]

We provide an specific MCMC algorithm with the ability to have redundant covariates in the true model.

Part I

CorReg : the concept

Chapter 3

Decorrelating covariates by a generative model

3.1 Generative model

3.2 Properties

3.2.1 general properties

3.2.2 Identifiability

3.3 About grouping effect

Chapter 4

Estimation of the Structure of subregression by MCMC

4.1 How to compare structures ?

4.1.1 Bayesian criterion for quality

4.1.2 Some indicators for proximity

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

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

4.2 Neighbourhood

4.2.1 Classical

4.2.2 Active relaxation of the constraints

4.3 The walk

4.4 Numerical results on simulated datasets

4.4.1 The datasets

Now we have defined the model and the way to obtain it, we can have a look on some numerical results to see if CorReg keeps its promises. The CorReg package has been tested on simulated datasets. Section 4.4.1 shows the results obtained in terms of $\hat{S}$. Sections **??** and 4.4.2 show the results obtained using only CorReg, or CorReg combined with other methods. Tables give both mean and standard deviation of the observed Mean Squared Errors (MSE) on a validation sample of 1000 individuals. For each simulation, $p = 40$, the R^2 of the main regression is 0.4, variables in $\boldsymbol{X}_f$ follow Gaussian mixture models of $\lambda = 5$ classes which means follow Poisson's law of parameter $\lambda = 5$ and which

standard deviation is λ. The β_j and the coefficients of the $\boldsymbol{\alpha}_j$ are generated according to the same Poisson law but with a random sign. $\forall j \in I_r, p_1^j = 2$ (sub-regressions of length 2) and we have $p_r = 16$ sub-regressions. The datasets were then scaled so that covariates X_r don't have a greater variance or mean. We used RMIXMOD to estimate the densities of each covariate. For each configuration, the MCMC walk was launched on 10 initial structures with a maximum of 1 000 steps each time. When $n < p$, a frequently used method is the Moore-Penrose generalized inverse [Katsikis and Pappas, 2008], thus OLS can obtain some results even with $n < p$. When using penalized estimators for selection, a last Ordinary Least Square step is added to improve estimation because penalisation is made to select variables but also shrinks remaining coefficients. This last step allows to keep the benefits of shrinkage (variable selection) without any impact on remaining coefficients (see [Zhang and Shen, 2010]) and is applied for both classical and marginal model. We compare different methods with and without CorReg as a pretreatment. All the results are provided by the CorReg package.

Results on $\hat{S}$

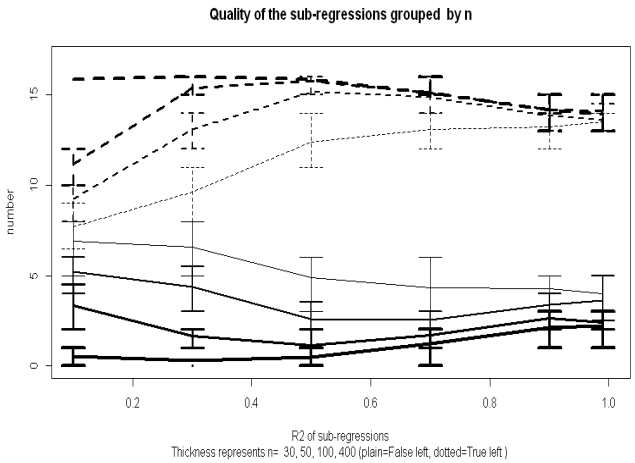

Figure 4.1: Quality of the subregressions found with classical BIC criterion

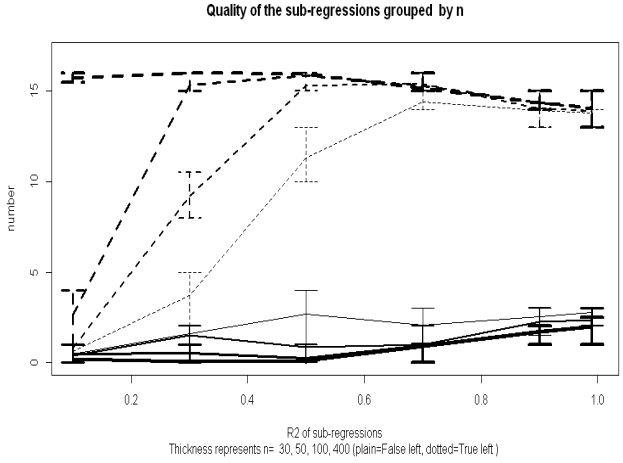

Figure 4.2: Quality of the subregressions found with our BIC_+ criterion

4.4.2 Results on prediction

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

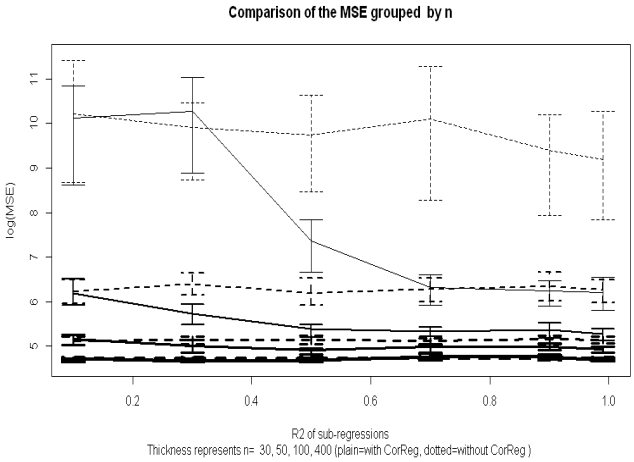

Figure 4.3: Comparison of the MSE between OLS and CorReg+OLS

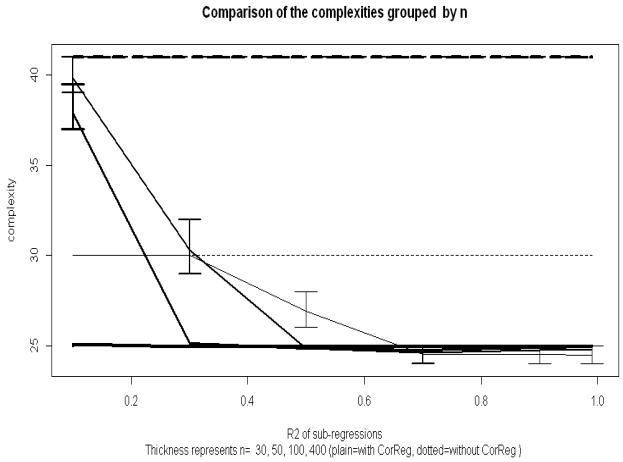

Figure 4.4: Comparison of the complexities between OLS and CorReg+OLS

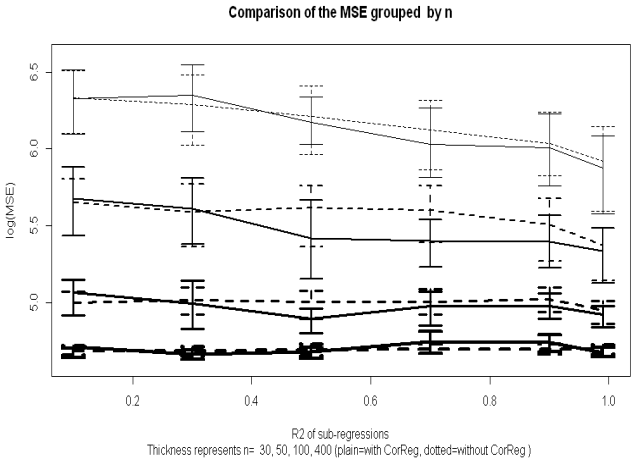

Figure 4.5: Comparison of the MSE between LASSO and CorReg+LASSO

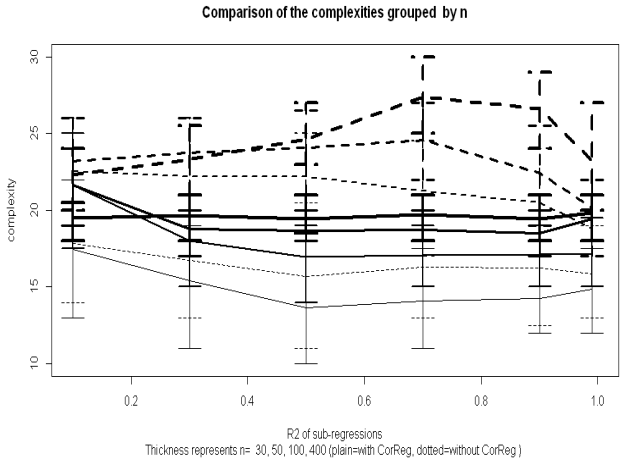

Figure 4.6: Comparison of the compexities between LASSO and CorReg+LASSO

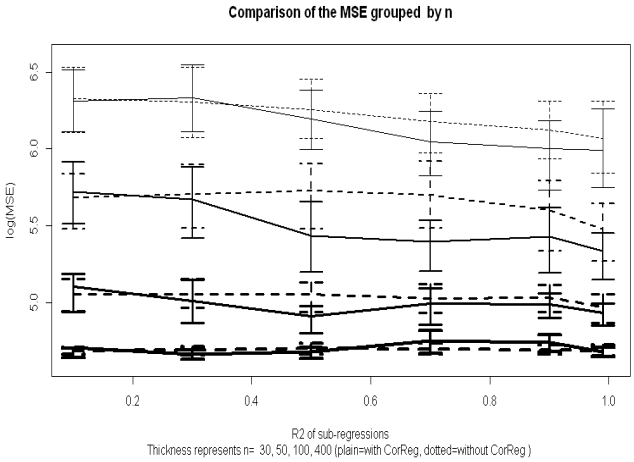

Figure 4.7: Comparison of the MSE between elasticnet and CorReg+elasticnet

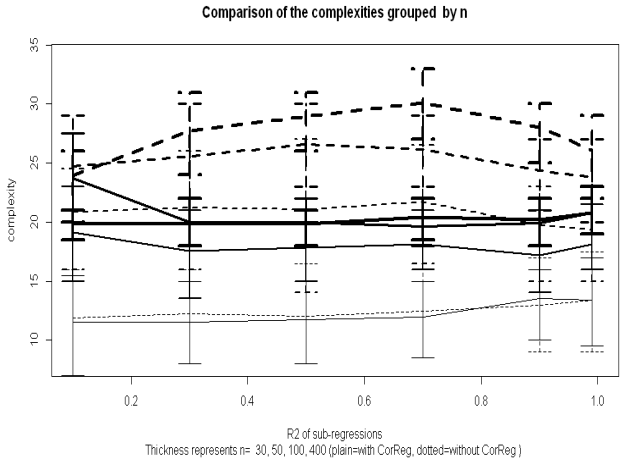

Figure 4.8: Comparison of the compexities between elasticnet and CorReg+elasticnet

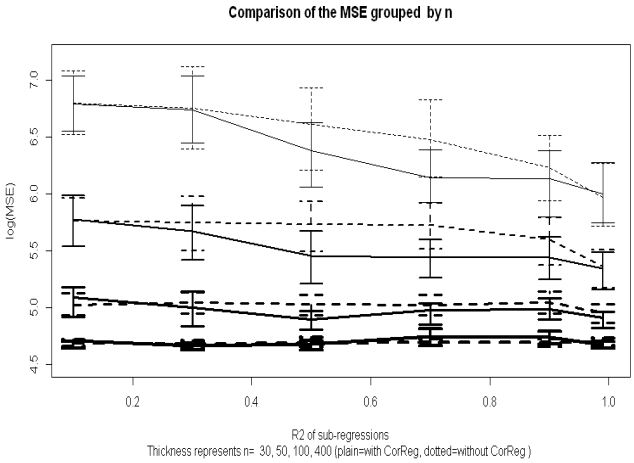

Figure 4.9: Comparison of the MSE between stepwise and CorReg+stepwise

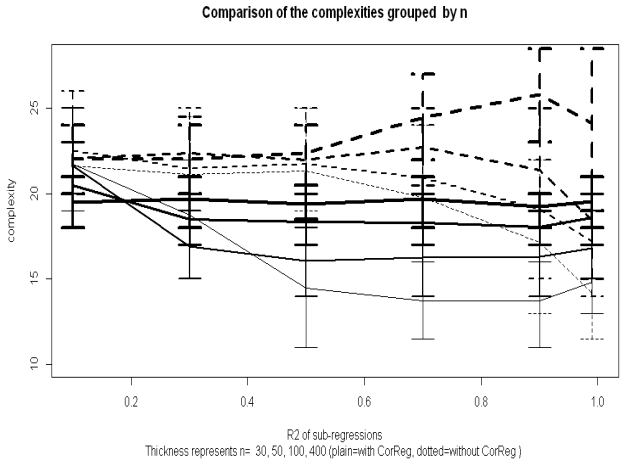

Figure 4.10: Comparison of the compexities between stepwise and CorReg+stepwise

Y depends on all variables in X

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

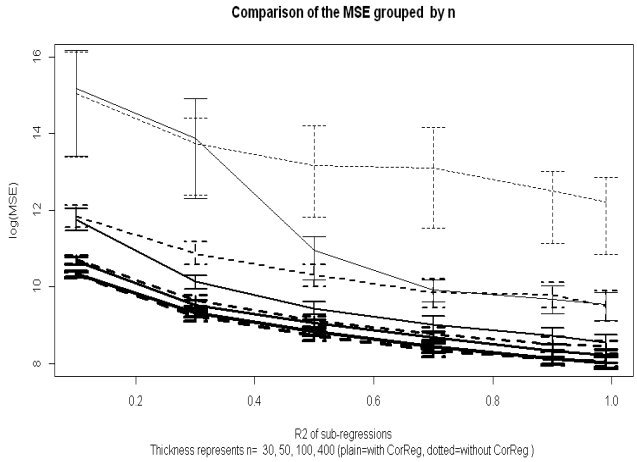

Figure 4.11: Comparison of the MSE between OLS and CorReg+OLS

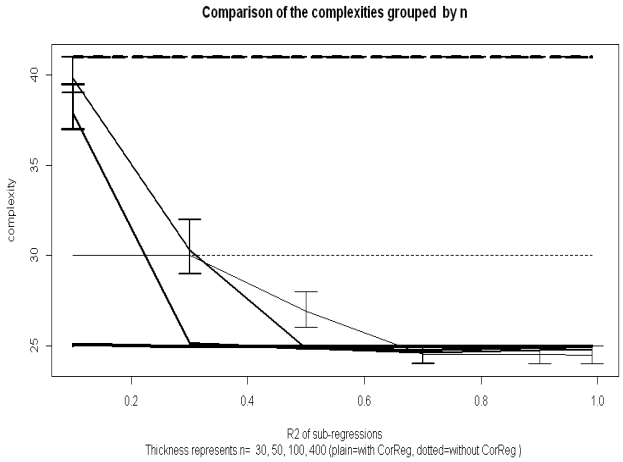

Figure 4.12: Comparison of the compexities between OLS and CorReg+OLS

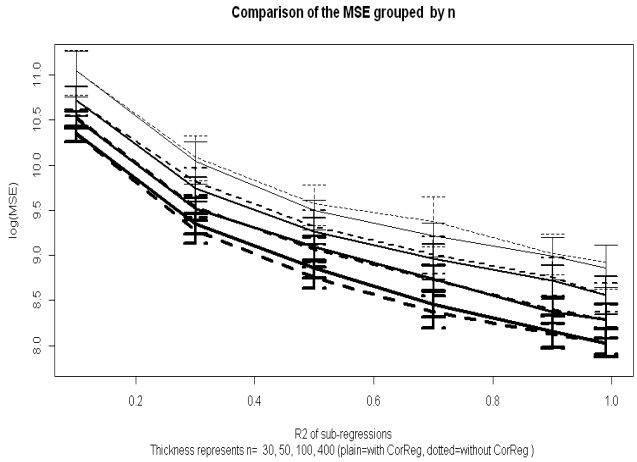

Figure 4.13: Comparison of the MSE between LASSO and CorReg+LASSO

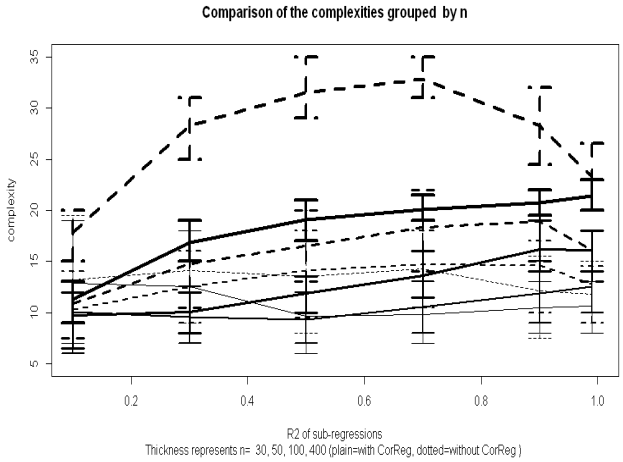

Figure 4.14: Comparison of the compexities between LASSO and CorReg+LASSO

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

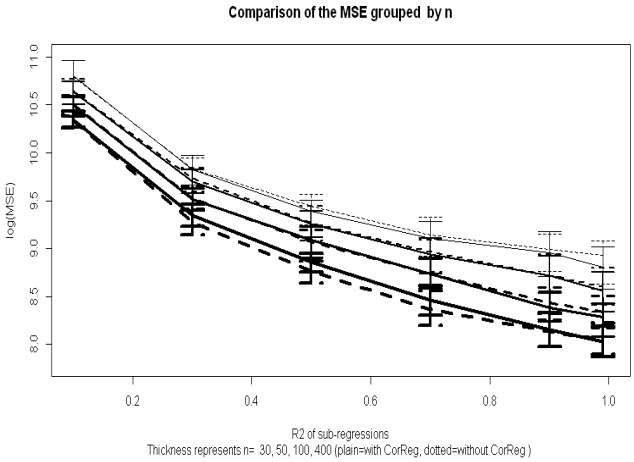

Figure 4.15: Comparison of the MSE between elasticnet and CorReg+elasticnet

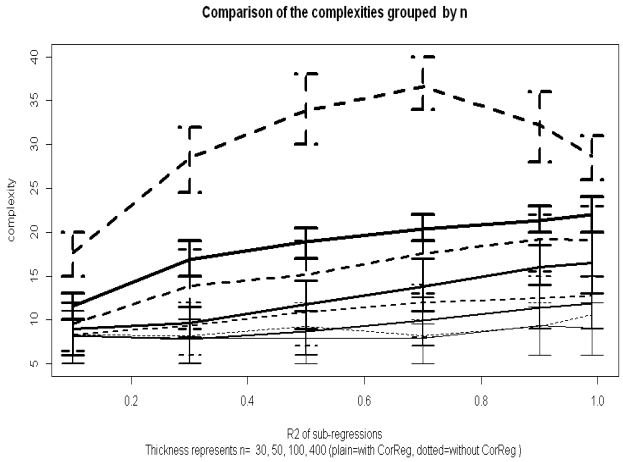

Figure 4.16: Comparison of the compexities between elasticnet and CorReg+elasticnet

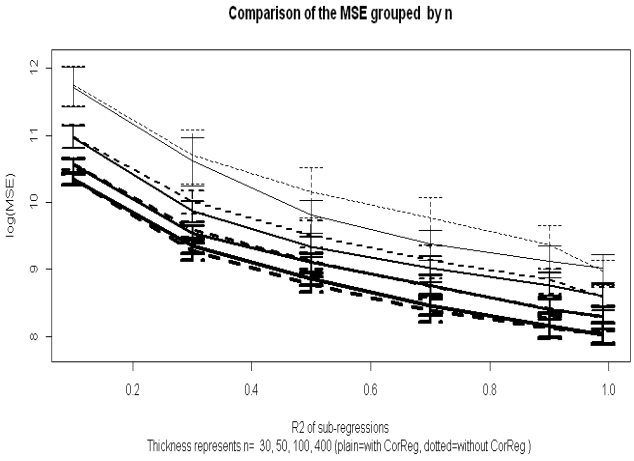

Figure 4.17: Comparison of the MSE between stepwise and CorReg+stepwise

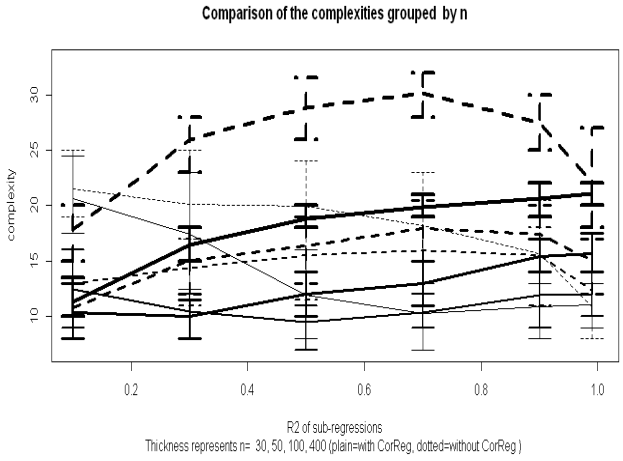

Figure 4.18: Comparison of the compexities between stepwise and CorReg+stepwise

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

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

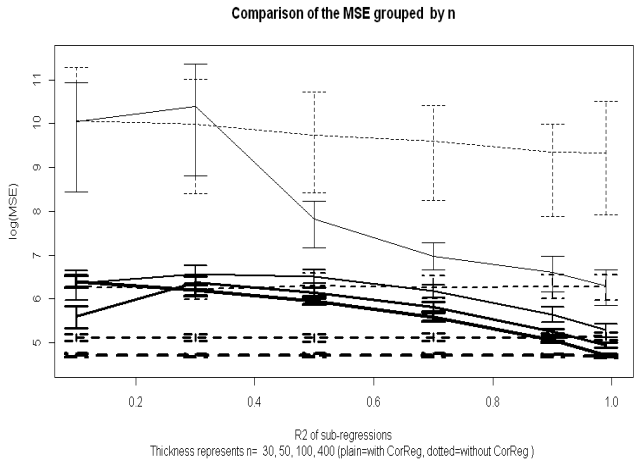

Figure 4.19: Comparison of the MSE between OLS and CorReg+OLS

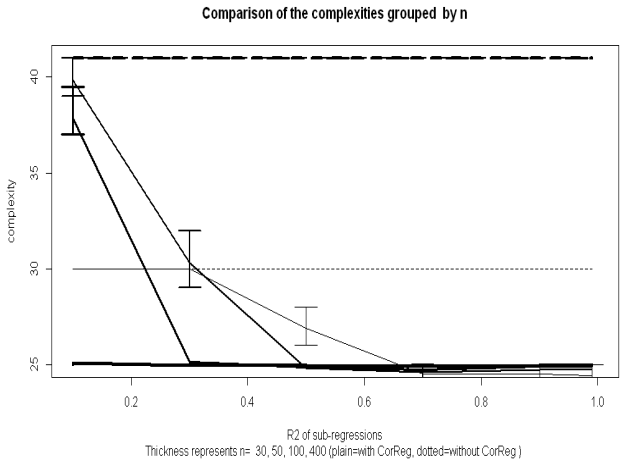

Figure 4.20: Comparison of the compexities between OLS and CorReg+OLS

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

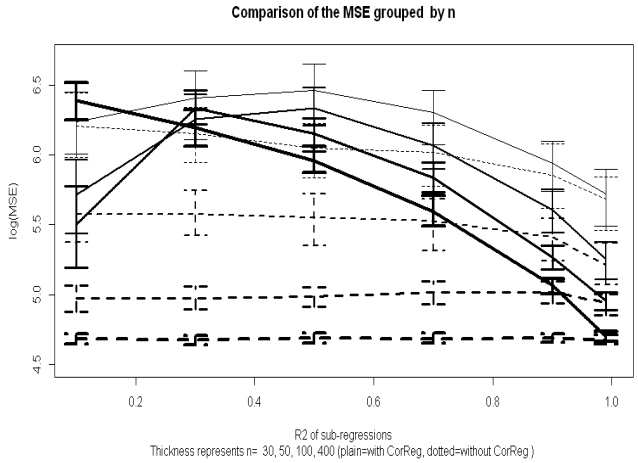

Figure 4.21: Comparison of the MSE between LASSO and CorReg+LASSO

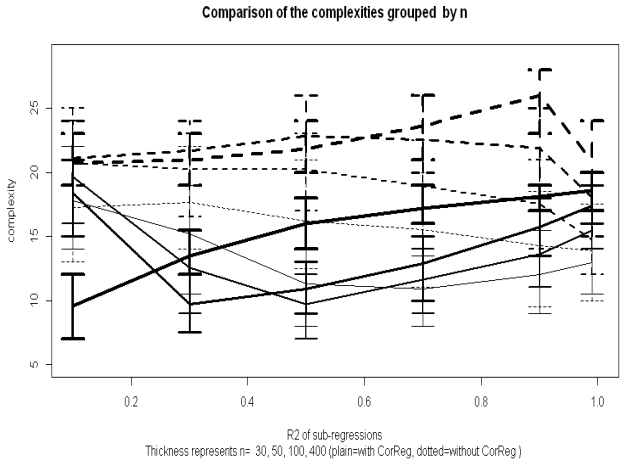

Figure 4.22: Comparison of the compexities between LASSO and CorReg+LASSO

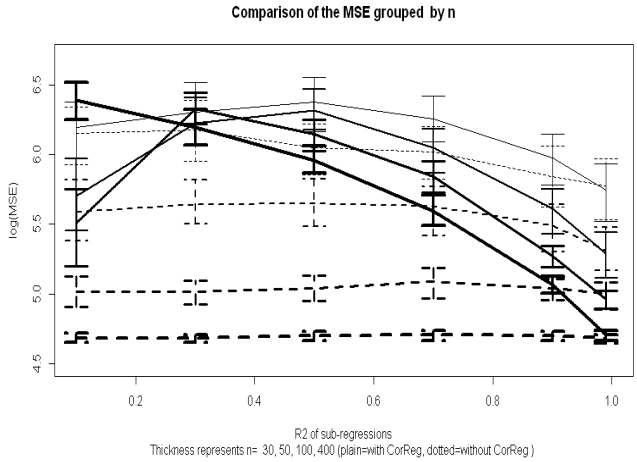

Figure 4.23: Comparison of the MSE between elasticnet and CorReg+elasticnet

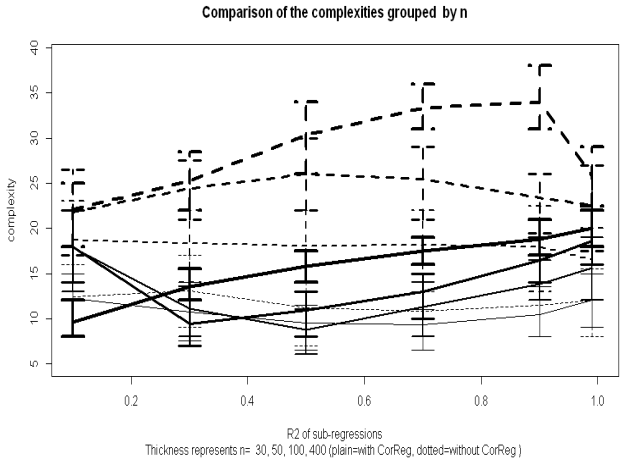

Figure 4.24: Comparison of the compexities between elasticnet and CorReg+elasticnet

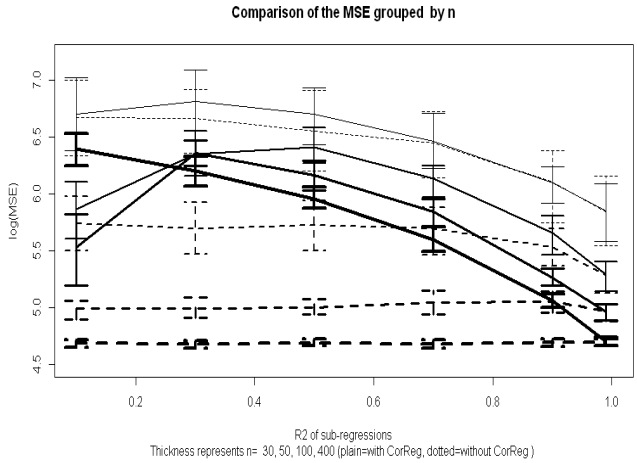

Figure 4.25: Comparison of the MSE between stepwise and CorReg+stepwise

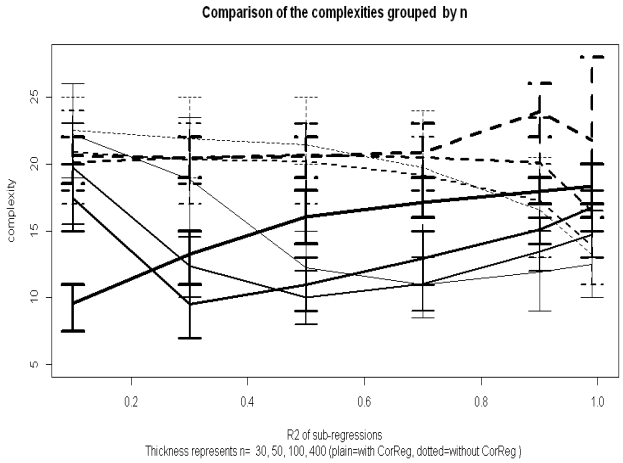

Figure 4.26: Comparison of the compexities between stepwise and CorReg+stepwise

Part II

Further usage of the structure

Chapter 5

Taking back the residuals

5.1 The model

5.2 Properties

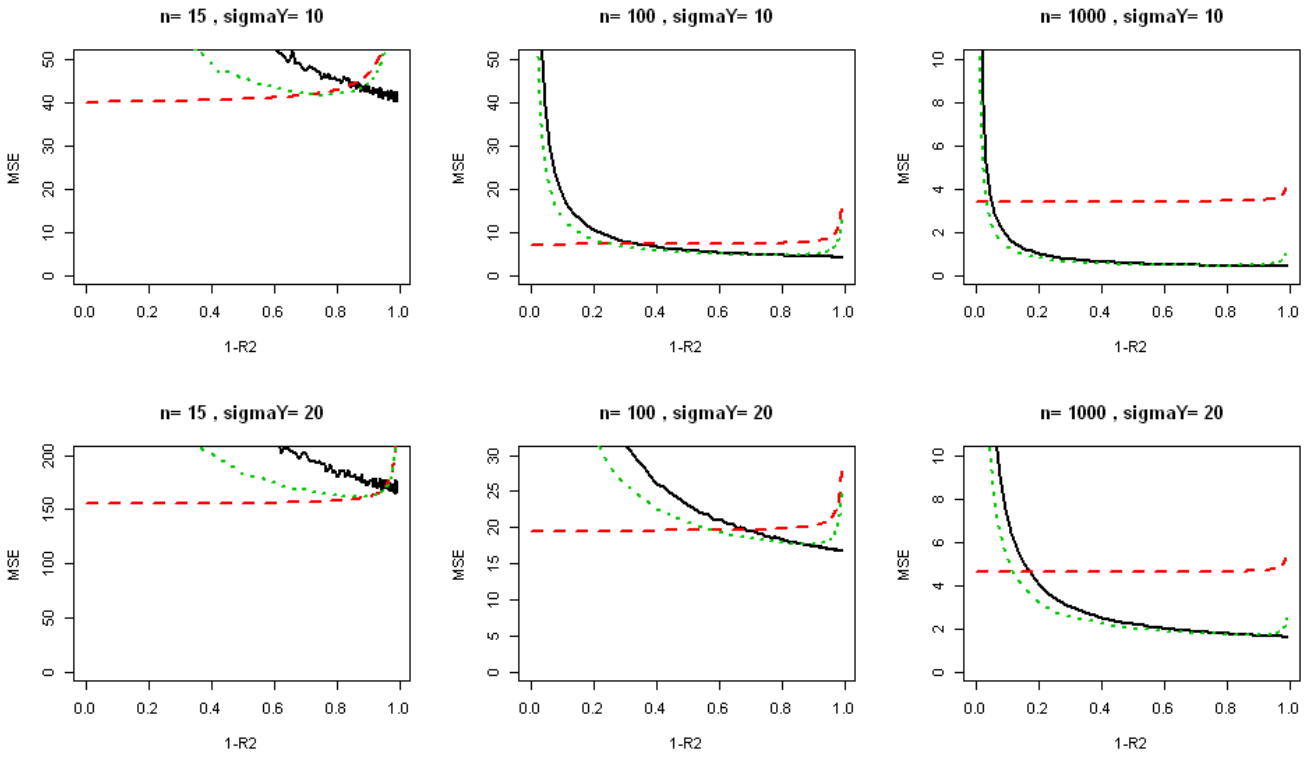

Figure 5.1: MSE of OLS (plain black) and CorReg marginal(red dashed) and CorReg full (green dotted) estimators for varying $(1 - R^2)$ of the sub-regression, n and σ_Y.

5.3 Consistency

5.3.1 Consistency Issues

Consistency issues of the LASSO are well known and Zhao [Zhao and Yu, 2006] 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 $\boldsymbol{X}_f, \boldsymbol{X}_r, \boldsymbol{\varepsilon}_Y, \boldsymbol{\varepsilon}_X i.i.d. \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ 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 structure but

a permuted one so we also look at the results obtained with the true S (but $\hat{B}$ is used) and with the structure found by the Markov chain after a few seconds.

True S is found 340 times on 1000 tries.

	Classical LASSO	CorReg Explicative	CorReg Predictive
True S	1.006479	**1.005468**	**1.006093**
$\hat{Z}$	**1.006479**	1.884175	1.006517

Table 5.1: MSE observer on a validation sample (1000 individuals)

We observe as we hoped that explicative model is better when using true S (coercing real zeros) and that explicative with $\hat{S}$ is penalized (coercing wrong coefficients to be zeros). But the main point is that the predictive model stay better than the classical one whith the true S and corrects enough the explicative model to follow the classical LASSO closely when using $\hat{S}$. And when we look at the consistency :

	Classical LASSO	Explicative	Predictive
True S	0	1000	830
$\hat{S}$	0	340	**621**

Table 5.2: 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 $\hat{S}$ 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 CorReg package for R. True S is now found 714 times on 1000 tries . So it confirms that non-gaussian models are easier to identify.

	Classical LASSO	Explicative	Predictive
True S	1.571029	**1.569559**	**1.570801**
$\hat{S}$	1.005402	1.465768	**1.005066**

Table 5.3: MSE observed on a validation sample (1000 individuals)

And when we look at the consistency :

	Classical LASSO	Explicative	Predictive
True S	0	1000	789
$\hat{S}$	0	714	**608**

Table 5.4: 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 $\hat{S}$ is better than classical LASSO in terms of MSE and consistent (classical LASSO is never consistent).

5.4 Numerical results

Chapter 6

Missing values

Real datasets often have missing values and it is a very recurrent issue in industry. We note $\boldsymbol{M}$ the $n \times p$ binary matrix indicating whereas a value is missing (1) or not (0) in $\boldsymbol{X}$. We note $\boldsymbol{X}_M$ the missing values and $\boldsymbol{X}_O$ the observed values. Θ stands for the parameters of the Gaussian mixture followed by $\boldsymbol{X}$. $\boldsymbol{\alpha}$ is the matrix of the sub-regression coefficients with $\alpha_{i,j}$ the coefficients associated to $\boldsymbol{X}^i$ in the sub-regression explaining $\boldsymbol{X}^j$ We have

$$g(\boldsymbol{X}|\Theta) = \int_{\boldsymbol{X}_M} f(\boldsymbol{X}|\Theta) d\boldsymbol{X} \tag{6.1}$$

6.1 State of the art

Many methods does exist to manage such problems [Little, 1992].

6.2 How to manage missing values in the MCMC ?

Here we suppose they are Missing Completely At Random (MCAR).

6.2.1 Decomposition of the integrated likelihood

In the MCMC we need to compute the likelihood of the dataset knowing the structure. When missing values occurs, we restrict the likelihood to the known values according to equation (6.1).

For the covariates in $\boldsymbol{X}_f$, we use the density estimated (*e.g.* a Gaussian Mixture model estimated by MIXMOD) or given as hypothesis. All individuals are supposed *iid* so $\forall(i,j)$ with $\boldsymbol{M}_{i,j} \neq 0$ and $j \notin I_r$:

$$g(x_{i,j}|\Theta) = f(x_{i,j}|\Theta) = \sum_{k=1}^{k_j} \pi_{j,k} \Phi(x_{i,j}|\mu_{j,k}, \Sigma_{j,k}) \tag{6.2}$$

with $k_j, \pi_{j,k}, \mu_{j,k}$ and $\Sigma_{j,k}$ estimated by Mixmod (for example).

Then we have

$$\begin{aligned} g(\boldsymbol{X}|\Theta) &= g(\boldsymbol{X}_r|\boldsymbol{X}_f, \Theta) g(\boldsymbol{X}_f|\Theta) && (6.3) \\ &= \prod_{i=1}^{n} \left[g(\boldsymbol{X}_i^{I_r}|\boldsymbol{X}_i^{I_f}, \Theta) \prod_{\substack{j \notin I_r \\ \boldsymbol{M}_{i,j}=0}} g(x_{i,j}|\Theta) \right] && (6.4) \\ &= \prod_{i=1}^{n} \left[g(\boldsymbol{X}_i^{I_r}|\boldsymbol{X}_i^{I_f}, \Theta) \prod_{\substack{j \notin I_r \\ \boldsymbol{M}_{i,j}=0}} \sum_{k=1}^{k_j} \pi_{j,k} \Phi(x_{i,j}|\mu_{j,k}, \Sigma_{j,k}) \right] && (6.5) \end{aligned}$$

reminding that covariates in $\boldsymbol{X}_f$ are orthogonal.

Residuals of the sub-regressions are orthogonal but missing values can make the residuals dependent. We have to decompose more precisely $g(\boldsymbol{X}_i^{I_r}|\boldsymbol{X}_i^{I_f},\Theta)$. To have a better view on the dependencies implied, we first write the marginal distributions.
$\forall(i,j)$ with $\boldsymbol{M}_{i,j}\neq 0$ and $j\in I_r$:

$$g(x_{i,j}|\boldsymbol{X}_i^{I_f},\Theta)=g(x_{i,j}|\boldsymbol{X}_i^{I_f^j},\Theta) = \sum_{k=1}^{k_{ij}}\pi_{ij,k}\Phi(x_{i,j}|\mu_{ij,k},\Sigma_{ij,k}) \text{ where} \tag{6.6}$$

$$\boldsymbol{\pi}_{ij} = \bigotimes_{\substack{l\in I_f^j\\ \boldsymbol{M}_{i,l}=1}}\boldsymbol{\pi}_l \text{ and } k_{ij}=|\pi_{ij}|, \tag{6.7}$$

$$\boldsymbol{\mu}_{ij} = \sum_{\substack{l\in I_f^j\\ \boldsymbol{M}_{i,l}=0}}\alpha_{l,j}x_{i,l}+\bigoplus_{\substack{l\in I_f^j\\ \boldsymbol{M}_{i,l}=1}}\alpha_{l,j}\boldsymbol{\mu}_l \tag{6.8}$$

$$\boldsymbol{\Sigma}_{ij} = \sigma_j^2+\bigoplus_{\substack{l\in I_f^j\\ \boldsymbol{M}_{i,l}=1}}\alpha_{i,l}^2\boldsymbol{\Sigma}_l \tag{6.9}$$

In first approximation we can suppose independence between the sub-regression:

$$\forall(j,j')\in I_r\times I_r, g(x_{i,j}|\boldsymbol{X}_i^{I_f},\Theta)\perp g(x_{i,j'}|\boldsymbol{X}_i^{I_f},\Theta) \tag{6.10}$$

then we have the complete expression of the likelihood with 6.5 and 6.6. Such approximation can be costless according to the position of the missing values (*e.g.* if they are all in $\boldsymbol{X}^{I_r}$). It is closer to the real model than the orthogonal hypothesis made by classical imputation by the mean. Moreover, sub-regressions are used only locally and errors don't cumulate whereas the true general decomposition combine many sub-regressions with cumulated noise of approximation. Thus, a general model would be better asymptotically but may not be efficient with finite dataset if the structure is complex. This first approximation is a good candidate to compare to the naive model (not taking into account the structure of sub-regression but making imputations by the mean for each covariate individually).

However, we write the real generalized expression for the log-likelihood. Let $\mathcal{I}_r$ be a permutation of I_r (arbitrary chosen, so the package will use identity). We define the general decomposition:

$$g(\boldsymbol{X}^{I_r}|\boldsymbol{X}^{I_f},\Theta) = \prod_{i=1}^{n}\left[g(x_{i,\mathcal{I}_r(p_r)}|\boldsymbol{X}_i^{I_f},\Theta)\prod_{j=1}^{p_r-1}g(x_{i,\mathcal{I}_r(j)}|\boldsymbol{X}_i^{\mathcal{I}_r(j+1)},\ldots,\boldsymbol{X}_i^{\mathcal{I}_r(p_r)},\boldsymbol{X}_i^{I_f},\Theta)\right] \tag{6.11}$$

where we don't know the expression of $g(x_{i,\mathcal{I}_r(j)}|\boldsymbol{X}_i^{\mathcal{I}_r(j+1)},\ldots,\boldsymbol{X}_i^{\mathcal{I}_r(p_r)},\boldsymbol{X}_i^{I_f},\Theta)$ so the previous approximation stands still.

6.2.2 Estimation of the coefficients in each regression

Estimating the $\boldsymbol{\alpha}_j$ with missing values is just estimating independent regressions with missing values. We have seen in equation (6.6) that we know the expression of this density for a given the $\boldsymbol{\alpha}_j$. So it's just about maximizing the likelihood of this density on the $\boldsymbol{\alpha}_j$. This can be done with an Expectation-Maximization (EM) algorithm [Dempster et al., 1977] or one of its extensions [McLachlan and Krishnan, 2007].

step E: (Θ stands for the parameters of the gaussian mixtures for the marginal distributions, estimated once by Mixmod):

$$\boldsymbol{X}^{(h)}=E[\boldsymbol{X}|\boldsymbol{X}_O,\boldsymbol{\alpha}^{(h)},\boldsymbol{\varepsilon},\Theta,S] \tag{6.12}$$

step M:

$$\boldsymbol{\alpha}^{(h+1)}=\text{argmax}_{\boldsymbol{\alpha}}(\mathcal{L}(\boldsymbol{X}^{(h)},\boldsymbol{\alpha},\boldsymbol{\varepsilon},\Theta,S)) \text{ by OLS} \tag{6.13}$$

But estimation of the $\boldsymbol{\alpha}_j$ is the most critical part of the MCMC in terms of computational time so it could be a bad idea to put there another iterative algorithm. Alternatives does exist :

- Because sub-regression are supposed to be parsimonious, we could imagine to estimate each column of $\boldsymbol{\alpha}$ with full sub-matrices of $\boldsymbol{X}_f$. When relying on too much missing values, $\hat{\boldsymbol{\alpha}}$ would be a bad candidate and then penalized directly by the likelihood (and it could be a good thing). Computational cost would be reduced significantly.
- To estimate the $\boldsymbol{\alpha}_j$ (and not for the global likelihood) we could use data imputation (by the mean) and then obtain a full matrix but still ignoring missing values when estimating the likelihood. Imputation only concerns the estimation of the sub-regression coefficients and because null coefficients in sub-regression are coerced at each step, imputation only concerns a few covariates each time.

$\forall j \in I_r$, estimation of $\boldsymbol{\alpha}^j$ only depends on individuals not missing in $\boldsymbol{X}^j$ (individuals are *iid*). So we work with a restriction of $\boldsymbol{X}$ for each $\boldsymbol{\alpha}^j$. Thus in this section, to simplify the notation, we will consider no missing values in $\boldsymbol{X}_r$ but in fact we work with restrictions.

The EM algorithm can be written here: we start with some $\Theta^{(0)} = (\boldsymbol{\alpha}, \boldsymbol{\varepsilon})$ initial value for Θ. The $\pi_{ij,k}$ are estimated once for each covariate (for example by Mixmod) and stay the same during the EM algorithm. Naive E step : estimation of

$$\boldsymbol{X}^{(h)} = E(\boldsymbol{X}|\boldsymbol{X}_{\bar{M}}, \Theta^{(h)}) \text{ so it simply is} \tag{6.14}$$

$\forall (i,j), \boldsymbol{M}_{i,j} = 1, j \neq I_r$,

$$x_{i,j}^{(h)} = E(x_{i,j}|\boldsymbol{X}_{\bar{M}}, \Theta^{(h)}) = \sum_{k=1}^{k_{ij}} \pi_{ij,k}\mu_{ij,k}^{(h)} \tag{6.15}$$

where, $\forall j \in I_f, k_{ij} = k_j, \pi_{ij,k} = \pi_{j,k}, \mu_{ij,k} = \mu_{j,k}$
M-step : we determine $\Theta^{(h+1)}$ as the solution of the equation

$$E(\boldsymbol{X}|\Theta) = \boldsymbol{X}^{(h)} \text{ done by OLS} \tag{6.16}$$

So the M step is just computing linear regressions on the filled dataset.

real E step : individuals are *iid* so we just look at the expression for one individual, and use it for all $\forall 1 \leq n \leq n, \forall j \notin I_r$, we note $\bar{\boldsymbol{X}}_{i,j} = (\boldsymbol{X}_{\bar{M}} \cap \boldsymbol{X}_i \setminus \boldsymbol{X}^j)$

$$\begin{aligned}
P(\boldsymbol{X}_{fi}^M, \boldsymbol{X}_{fi}^{\bar{M}}, \boldsymbol{X}_{ri}^{\bar{M}}|\Theta) &= P(\boldsymbol{X}_{fi}^M|\boldsymbol{X}_{fi}^{\bar{M}}, \boldsymbol{X}_{ri}^{\bar{M}}, \Theta)P(\boldsymbol{X}_{fi}^{\bar{M}}, \boldsymbol{X}_{ri}^{\bar{M}}|\Theta) && (6.17)\\
&= P(\boldsymbol{X}_{ri}^{\bar{M}}|\boldsymbol{X}_{fi}^M, \boldsymbol{X}_{fi}^{\bar{M}}, \Theta)P(\boldsymbol{X}_{fi}^M, \boldsymbol{X}_{fi}^{\bar{M}}|\Theta) && (6.18)\\
P(\boldsymbol{X}_{fi}^M|\boldsymbol{X}_{fi}^{\bar{M}}, \boldsymbol{X}_{ri}^{\bar{M}}, \Theta) &= \frac{P(\boldsymbol{X}_{ri}^M|\boldsymbol{X}_{fi}^M, \boldsymbol{X}_{fi}^{\bar{M}}, \Theta)P(\boldsymbol{X}_{fi}^M, \boldsymbol{X}_{fi}^{\bar{M}}|\Theta)}{P(\boldsymbol{X}_{fi}^{\bar{M}}, \boldsymbol{X}_{ri}^{\bar{M}}|\Theta)} && (6.19)\\
&= \frac{P(\boldsymbol{X}_{ri}^O|\boldsymbol{X}_{fi}^M, \boldsymbol{X}_{fi}^{\bar{M}}, \Theta)P(\boldsymbol{X}_{fi}^M|\Theta)P(\boldsymbol{X}_{fi}^{\bar{M}}|\Theta)}{P(\boldsymbol{X}_{ri}^{\bar{M}}|\boldsymbol{X}_{fi}^{\bar{M}}, \Theta)P(\boldsymbol{X}_{fi}^{\bar{M}}|\Theta)} && (6.20)\\
&= \frac{P(\boldsymbol{X}_{ri}^{\bar{M}}|\boldsymbol{X}_{fi}^M, \boldsymbol{X}_{fi}^{\bar{M}}, \Theta)P(\boldsymbol{X}_{fi}^M|\Theta)}{P(\boldsymbol{X}_{ri}^{\bar{M}}|\boldsymbol{X}_{fi}^{\bar{M}}, \Theta)} && (6.21)
\end{aligned}$$

No imputation for missing left. Imputations for missing right are just used to obtain $\hat{\boldsymbol{\alpha}}$ but not when computing the BIC or BIC_+.

6.2.3 Weighted penalty

Now we have defined the way to compute the likelihood, other questions remain : how to define the number of parameters in the structure ? How to take into account missingness (structures relying on highly missing covariates should be penalized) ? We have seen that for a same covariate X^j with $j \in I_r$, the number of parameters is not the same for each individual depending whether or not $M_{i,j} = 0$. But the penalty (for $\psi = BIC$) can't be added at the individual level (because $\log(1) = 0$ so it would be annihilated).

To penalize models that suppose dependencies based only on a few individuals, we propose to use the mean of the complexities obtained for a given covariate. Thus if a structure is only touched by one missing value the penalty will be smaller than another same shaped structure but with more missing values implied. Another way would be to use $\psi = RIC$ (see [Foster and George, 1994]) so the complexity is associated with $\log(p)$ and can be added individually. Another idea would be to make a compromise and penalize by $\frac{k_i \log(p)}{\log(n)}$.

6.3 Missing values in the main regression

6.3.1 explicative

The reduced model (explicative one) is just a linear regression without structure so estimating $\boldsymbol{\beta}$ is like estimating the $\boldsymbol{\alpha}_j$ and the same methods can be used. An EM algorithm would be prefered because this estimation is out of the MCMC, will be computed only one time and is the final objective where we want to minimize the error. The E step is the same as during the MCMC because we can take benefits from the structure. Where others will just fill $\boldsymbol{X}^j$ by their means, we can use when $j \in I_r$ the conditionnal mean $E(X^j|X_f, \boldsymbol{\alpha}^j)$ defined in 6.15.

6.3.2 predictive

If there are missing values in $\boldsymbol{X}^j \in \boldsymbol{X}_r$ a new possibility appears. Knowing S and $\boldsymbol{\alpha}_j$ we are able to try a conditional imputation based on the corresponding sub-regression, like every time someone use linear regression for prediction.

Chapter 7

CorReg: the package and its application in steel industry

7.1 CorReg package for R

CorReg is already downloadable on the CRAN under CeCILL Licensing. This package permits to generate datasets according to our generative model, to estimate the structure (C++ code) of regression within a given dataset and to estimate both explicative and predictive model with many regression tools (OLS,stepwise,LASSO,elasticnet,clere,spike and slab, adaptive lasso and every models in the lars package). So every simulation presented above can be done with CorReg. CorReg also provides tools to interpreat found structures and visualize the dataset (missing values and correlations). More informations can be found on the website www.correg.org which is dedicated to CorReg.

7.2 Application in steel industry

7.2.1 The dataset

7.2.2 Found Structure

7.2.3 Results

Chapter 8

Conclusion and perspectives

Bibliography

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

[Chiquet J. and S., 2013] Chiquet J., M.-H. T. and S., R. (2013). Multi-trait genomic selection via multivariate regression with structured regularization. *Proceedings of MLCB/NIPS'13 workshop*.

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

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

[Foster and George, 1994] Foster, D. and George, E. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975.

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

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

[Little, 1992] Little, R. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237.

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

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

[McLachlan and Krishnan, 2007] McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

[Miller, 2002] Miller, A. (2002). *Subset selection in regression*. CRC Press.

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

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

[Tibshirani et al.,] Tibshirani, R., Hoefling, G., Wang, P., and Witten, D. The lasso: some novel algorithms and applications.

[Wang et al., 2011] Wang, S., Nan, B., Rosset, S., and Zhu, J. (2011). Random lasso. *The annals of applied statistics*, 5(1):468.

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

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

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

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

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

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

Appendix A

Graphs and CorReg

A.1 Matricial notations

A.2 Properties

Appendix B

Mixture models

B.1 Linear combination

B.2 Industrial examples