CoMPASS: **Co**rrelation **M**odeling for **P**retreatment by **A**utomated **S**tructure **S**election

Clément THERY

June 30, 2014

To my sons

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 $\| \boldsymbol{\beta} \|_\infty = \max(|\beta_1|, \dots, |\beta_p|)$.

2.1 Ordinary least squares and associated problems

We note the linear regression model:

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

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

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{Y} \tag{2.2}$$

with variance matrix

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma_Y^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \tag{2.3}$$

and without any bias. Estimation of $\boldsymbol{\beta}$ requires the inversion of $\boldsymbol{X}'\boldsymbol{X}$ which will be ill-conditioned or even singular if some covariates depend linearly from each other. Conditionning of $\boldsymbol{X}'\boldsymbol{X}$ get worse based on two aspects: the dimension p (number of covariates) of the model (the more covariates you have the greater variance you get) and the correlations within the covariates: strongly correlated covariates give bad-conditioning and increase variance of the estimators . When correlations between covariates are strong, the matrix to invert is ill-conditioned and the variance increases, giving unstable and unusable estimator [Hoerl and Kennard, 1970]. Another problem is that matrix inversion requires $n \geq p$.

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:

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}\left\{\| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \|_2^2\right\} \text{ subject to } \| \boldsymbol{\beta} \|_2^2 \leq \lambda \text{ with } \lambda > 0 \tag{2.4}$$

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 $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin} \left\{ \parallel \boldsymbol{Y} - \boldsymbol{X\beta} \parallel_2^2 \right\} \text{ subject to } \parallel \boldsymbol{\beta} \parallel_1 \leq \lambda \text{ with } \lambda > 0 \tag{2.5}$$

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:

$$\hat{\boldsymbol{\beta}} = (1 + \lambda_2) \operatorname{argmin} \left\{ \parallel \boldsymbol{Y} - \boldsymbol{X\beta} \parallel_2^2 \right\}, \text{ subject to } (1 - \alpha) \parallel \boldsymbol{\beta} \parallel_1 + \alpha \parallel \boldsymbol{\beta} \parallel_2^2 \leq t \text{ for some } t \tag{2.6}$$

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).

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\beta} \parallel \boldsymbol{Y} - \boldsymbol{X\beta} \parallel_2^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| + c \sum_{j<k} \max(|\beta_j|, |\beta_k|) \leq \lambda \tag{2.7}$$

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

Pretreatment for correlations

Chapter 3

Decorrelating covariates by a generative model

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

3.1 Our proposal: modelisation of the correlations

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

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

$$\boldsymbol{X}_{r|\boldsymbol{X}_f,S} \text{ defined by } \forall \boldsymbol{X}^j \subset \boldsymbol{X}_r : \boldsymbol{X}^j_{|\boldsymbol{X}_f,S} = \boldsymbol{X}_f \boldsymbol{\alpha}_j + \boldsymbol{\varepsilon}_j \text{ with } \boldsymbol{\varepsilon}_j \sim \mathcal{N}(\boldsymbol{0}, \sigma_j^2 \boldsymbol{I}_n) \tag{3.1}$$

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

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

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

The structure obtained gives a system of linear regression that can be viewed as a recursive Simultaneous Equation Model (SEM)[Davidson and MacKinnon, 1993] [Timm, 2002]. Here we suppose the $\boldsymbol{\varepsilon}_j$ independent but in other cases SUR (Seemingly Unrelated Regression [Zellner, 1962]) takes into account correlations between residuals SUR (Seemingly Unrelated Regression [Zellner, 1962]) and could be used to estimate the $\boldsymbol{\alpha}_j$.

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

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

Now we know $P(\boldsymbol{X}_r|\boldsymbol{X}_f, S)$ by the structure of sub-regressions, we are able to define a marginal regression model $P(\boldsymbol{Y}|\boldsymbol{X}_f, S)$ based on the reduced set of independent covariates $\hat{\boldsymbol{\beta}}_f$ without significant information loss. We use the information of the correlations structure to rewrite the true model without bias in the marginal space defined by the independent covariates.
Using the partition $\boldsymbol{X} = [\boldsymbol{X}_f, \boldsymbol{X}_r]$ we can rewrite (2.1):

$$\boldsymbol{Y}_{|\boldsymbol{X}_f,\boldsymbol{X}_r,S} = \boldsymbol{X}_f\boldsymbol{\beta}_f + \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}_{\boldsymbol{Y}} \tag{3.2}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_f, \boldsymbol{\beta}_r) \in \mathcal{R}^p$ is the vector of the regression coefficients associated respectively to $\boldsymbol{X}_f$ and $\boldsymbol{I}_n$ the identity matrix. We note that (3.1) and (3.2) give also by simple integration on $\boldsymbol{X}_r$ a marginal regression model on $\boldsymbol{Y}$ *depending only on uncorrelated covariates* $\boldsymbol{X}_f$:

$$\begin{aligned} P(\boldsymbol{Y}|\boldsymbol{X}_f) &= \int_{\boldsymbol{X}_r} P(\boldsymbol{Y}|\boldsymbol{X}_r,\boldsymbol{X}_f)P(\boldsymbol{X}_r|\boldsymbol{X}_f)d\boldsymbol{X} & (3.3)\\ \boldsymbol{Y}_{|\boldsymbol{X}_f,S} &= \boldsymbol{X}_f(\boldsymbol{\beta}_f + \sum_{j\in I_r}\beta_j\boldsymbol{\alpha}_j) + \sum_{j\in I_r}\beta_j\boldsymbol{\varepsilon}_j + \boldsymbol{\varepsilon}_Y & (3.4)\\ &= \boldsymbol{X}_f\boldsymbol{\beta}_f^* + \boldsymbol{\varepsilon}_Y^* & (3.5)\end{aligned}$$

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

$$\boldsymbol{Y}_{|\boldsymbol{X},S} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}_Y^* \text{ where } \boldsymbol{\beta}^* = (\boldsymbol{\beta}_f^*, \boldsymbol{\beta}_r^*) \text{ and } \boldsymbol{\beta}_r^* = \boldsymbol{0} \tag{3.6}$$

for which OLS estimator of the coefficients may be biased.

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

3.3 Strategy of use: pre-treatment before classical estimation/selection methods

As a pre-treatment, the model allows usage of any method in a second time to estimate $\boldsymbol{\beta}_f^*$, even with variable selection methods like LASSO or a best subset algorithm like stepwise [Seber and Lee, 2012]. However, we always have $\boldsymbol{X}_r = \boldsymbol{0}$

After selection and estimation we will obtain a model with *two steps of variable selection*: the decorrelation step by marginalization(coerced selection associated to redundant information defined in S) and the classical selection step, with different meanings for obtained zeros in $\hat{\boldsymbol{\beta}}_f^*$ (irrelevant covariates) and for $\hat{\boldsymbol{\beta}}_r^* = 0$ (redundant information). Thus we are able to distinguish the reasons of selection and consistency issues don't mean interpretation issues any more. So we dodge the drawbacks of both grouping effect and variable selection.

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

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

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

3.4 Illustration of the tradeoff conveyed by the pre-treatment

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

For the marginal regression model defined in (3.5) we have the OLS unbiased estimator of $\boldsymbol{\beta}^*$:

$$\hat{\boldsymbol{\beta}}^*_f = (\boldsymbol{X}'_f\boldsymbol{X}_f)^{-1}\boldsymbol{X}'_f\boldsymbol{Y} \text{ and } \hat{\boldsymbol{\beta}}^*_r = \boldsymbol{0} \tag{3.7}$$

We see in (3.4) that it gives an unbiased estimation of $\boldsymbol{Y}$ and $\boldsymbol{\beta}^*$ but in terms of $\boldsymbol{\beta}$ this estimator is biased:

$$\mathrm{E}[\hat{\boldsymbol{\beta}}^*_f|\boldsymbol{X}_f] = \boldsymbol{\beta}_f + \sum_{j\in I_r}\beta_j\boldsymbol{\alpha}_j \text{ and } \mathrm{E}[\hat{\boldsymbol{\beta}}^*_r|\boldsymbol{X}_f] = \boldsymbol{0} \tag{3.8}$$

with variance:

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}^*_f|\boldsymbol{X}_f] = (\sigma_Y^2 + \sum_{j\in I_r}\sigma_j^2\beta_j^2)(\boldsymbol{X}'_f\boldsymbol{X}_f)^{-1} \text{ and } \mathrm{Var}[\hat{\boldsymbol{\beta}}^*_r|\boldsymbol{X}_f] = \boldsymbol{0} \tag{3.9}$$

We see that the variance is reduced compared to OLS described in equation (2.3)(no correlations and smaller matrix give better conditioning) for small values of σ_j *i.e.* strong correlations. So we play on the bias-variance tradeoff, reducing the variance by adding a bias.

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

$$\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= \parallel \mathrm{Bias} \parallel_2^2 + \mathrm{Tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}})) && (3.10)\\
\mathrm{MSE}(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{X}) &= 0 + \sigma_Y^2\,\mathrm{Tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}) && (3.11)\\
\mathrm{MSE}(\hat{\boldsymbol{\beta}}^*_{OLS}|\boldsymbol{X}) &= \parallel \sum_{j\in I_r}\beta_j\boldsymbol{\alpha}_j \parallel_2^2 + \parallel \boldsymbol{\beta}_r \parallel_2^2 + (\sigma_Y^2 + \sum_{j\in I_2}\sigma_j^2\beta_j^2)\,\mathrm{Tr}((\boldsymbol{X}'_f\boldsymbol{X}_f)^{-1}) && (3.12)
\end{aligned}$$

To better illustrate the bias-variance tradeoff, we look at the running example. We observe the theoretical Mean Squared Error (MSE) of the estimator of both OLS and CorReg's marginal model for several values of σ_3 (strength of the sub-regression) and n. Figure 3.4 shows the theoretical MSE evolution with the strength of the sub-regression:

$$1 - \mathcal{R}^2 = \frac{\mathrm{Var}(\boldsymbol{\varepsilon})_{\boldsymbol{3}}}{\mathrm{Var}(\boldsymbol{x}_3)} = \frac{\sigma_3^2}{\sigma_3^2 + 2} \tag{3.13}$$

It is clear in Figure 3.4 that the marginal model is more robust than OLS on $\boldsymbol{X}$. And when sub-regression get weaker ($1-\mathcal{R}^2$ tends to 1) it remains stable until extreme values (sub-regression nearly fully explained by the noise). We also see that the error implied by strong correlations shrinks with the rise of n. We see that σ_Y multiplies $\mathrm{Tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}})) = \mathrm{Tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}}_f)) + \mathrm{Tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}}_r))$ for both models but for the marginal model $\mathrm{Tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}}_r)) = 0$. Thus, when σ_Y^2 rises it increases the advantage of CorReg versus OLS. It illustrates the importance of dimension reduction when the model has a strong noise (very usual case on real datasets where true model is not even exactly linear). Further results are provided in sections 4.5 and 4.6.

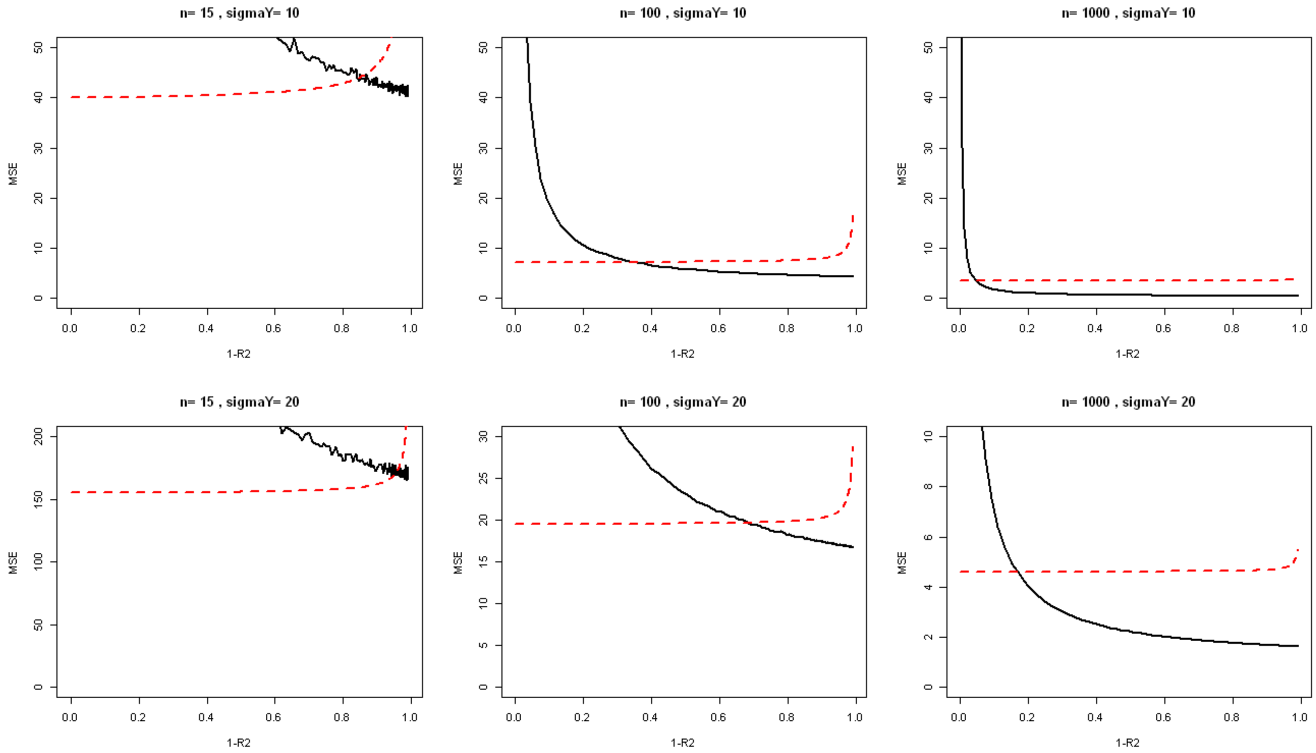


Figure 3.1: MSE of OLS (plain) and CorReg (dotted) estimators for varying $(1 - R^2)$ of the sub-regression, n and σ_Y.

Chapter 4

Estimation of the Structure of subregression by MCMC

4.1 Sub-regressions model selection

Structural equations models like SEM are often used in social sciences and economy where a structure is supposed "by hand" but here we want to find it automatically. Graphical LASSO [Friedman et al., 2008] offers a method to obtain a structure based on the precision matrix (inverse of the variance-covariance matrix), setting some coefficients of the precision matrix to zero. But the resulting matrix is symmetric and we need an oriented structure for S to avoid cycles.

Cross-validation is very time-consuming and thus not friendly with combinatory problematics. Moreover, we need a criterion compatible with structures of different sizes (varying p_r) and not related with $\boldsymbol{Y}$ because the structure is inherent to $\boldsymbol{X}$ only. Thus it must be a global criterion. Because it is about model selection and we are able to provide a full generative model (section 4.1.1), we decide to follow a Bayesian approach ([Raftery, 1995], [Andrieu and Doucet, 1999],[Chipman et al., 2001]).

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

$$P(S|\boldsymbol{X}) \quad \propto \quad P(\boldsymbol{X}|S)P(S) = P(\boldsymbol{X}^{I_r}\boldsymbol{X}^{I_f},S)P(\boldsymbol{X}^{I_f}|S)P(S) \tag{4.1}$$

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

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

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

$$\forall \boldsymbol{X}^j \notin \boldsymbol{X}^{I_r} : \boldsymbol{X}^j_{|S} \sim f(\boldsymbol{\theta}_j) = \mathcal{GM}(\boldsymbol{\pi}_j;\boldsymbol{\mu}_j;\boldsymbol{\sigma}^2_j) \text{ with } \boldsymbol{\pi}_j,\boldsymbol{\mu}_j,\boldsymbol{\sigma}^2_j \text{ vectors of size } K_j. \tag{4.2}$$

The great flexibility [McLachlan and Peel, 2004] of such models makes our model more robust. Gaussian case is just a special case ($K_j = 1$) of Gaussian mixture so it is included in our hypothesis but identifiability of S requires to have Gaussian mixtures with at least two distinct components in each sub-regression (derived from the identifiability of the SR model in [Maugis et al., 2009], more details in the Appendices 4.1.2).

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

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

4.1.2 Identifiability of the structure

The model presented above relies on a discrete structure S between the covariates. But to find it we need identifiability property to insure the MCMC will asymptotically find the true model.

Identifiability of the structure is asked in following terms: Is it possible to find another structure $\tilde{S}$ of linear regression between the covariates leading to the same joint distribution and marginal distributions?

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

Our full generative model is a p-sized Gaussian mixture model of K distinct components and can be seen as a **SR** model defined by Maugis [Maugis et al., 2009]. In this section, S will denote the set of variable as in the paper from Maugis and we call Gaussian mixtures the Gaussian mixtures with at least two distinct components. The equivalence with Maugis's model is defined by: $\boldsymbol{X}^{I_r} = \boldsymbol{y}^{S^c}$ and $\boldsymbol{X}^{I_f} = \boldsymbol{y}^R$. We have supposed independence between variables in $\boldsymbol{X}^{I_f}$ so the identifiability theorem from Maugis tells that our model is identifiable if variables in $\boldsymbol{X}^{I_f}$ are Gaussian mixtures (what we supposed in section 4.1.1).

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

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

$$\begin{aligned}
\boldsymbol{X}_{r|\boldsymbol{X}^G,\boldsymbol{X}^{G^c}} &= \boldsymbol{X}^G\boldsymbol{\alpha}_G + \boldsymbol{X}^{G^c}\boldsymbol{\alpha}_{G^c} + \boldsymbol{\varepsilon} && (4.3)\\
\boldsymbol{X}_{r|\boldsymbol{X}^{G^c}} &= \boldsymbol{X}^{G^c}\boldsymbol{\alpha}_{G^c} + \tilde{\boldsymbol{\varepsilon}} \text{ is identifiable where} && (4.4)\\
\tilde{\boldsymbol{\varepsilon}}_j &= \boldsymbol{X}^G\boldsymbol{\alpha}_j^G + \boldsymbol{\varepsilon}_j \text{ is Gaussian.} && (4.5)
\end{aligned}$$

So a sufficient condition for identifiability is to have at least one Gaussian mixture in each sub-regression. It implies then that: $\forall j \in I_r, \boldsymbol{X}^j \subset \boldsymbol{X}^G$ and $\exists i \in I_f^j, \boldsymbol{X}^i \subset \boldsymbol{X}^G$.

4.2 How to compare structures ?

4.2.1 Bayesian criterion for quality

Our full generative generative model allows us to compare structures with criterions like the Bayesian Information Criterion (BIC) which penalize the log-likelihood of the joint law on $\boldsymbol{X}$ according to the complexity of the structure [Lebarbier and Mary-Huard, 2006].
We can also imagine to use other criterions, like the RIC (Risk Inflation Criterion [Foster and George, 1994]) that choose a penalty in $\log p$ instead of $\log n$ and thus gives more parsimonious models when p is larger than n (high dimension) or any other criterion [George and McCulloch, 1993] thought to be better in a given context. In the followings we use the BIC.

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

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

$$-2\log P(\boldsymbol{X}|S) \approx BIC = -2\mathcal{L}(\boldsymbol{X}, S, \boldsymbol{\Theta}) + |\boldsymbol{\Theta}|\log(n) \qquad (4.6)$$

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

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

$$BIC_+(X|S) \quad = \quad BIC(X|S) - \ln(P(S)) \tag{4.7}$$

It increases penalty on complexity for $p_r \leq \frac{p}{2}$ and $p_f^j \leq \frac{p}{2}$. Hence this constraint on $\hat{p}_r$ and $\hat{p}_f^j$ is given in the research algorithm when the Hierarchical Uniform hypothesis is made instead of Uniform one in numerical experiments (section 4.5 and 4.6). BIC_+ does not change BIC but only $P(S)$ so the properties of BIC_+ are the same as classical BIC but we obtain better results when the constraints on the complexity are verified.

4.2.3 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 (4.1))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 : $\Delta compl = \sum_{j \in p_r} p_f^j - \sum_{j \in \hat{p}_r} \hat{p}_f^j$

4.3 Neighbourhood

4.3.1 Classical

4.3.2 Active relaxation of the constraints

4.4 The walk

4.5 Numerical results on simulated datasets

4.5.1 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 4.5.2 shows the results obtained in terms of $\hat{S}$. Sections **??** and 4.5.3 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.

4.5.2 Results on $\hat{S}$

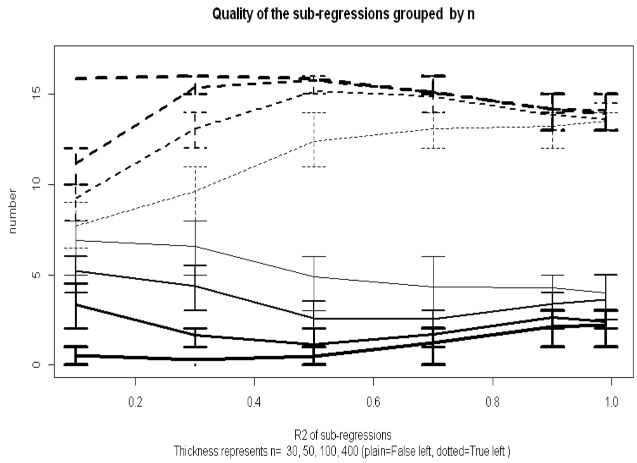


Figure 4.1: Quality of the subregressions found with classical BIC criterion

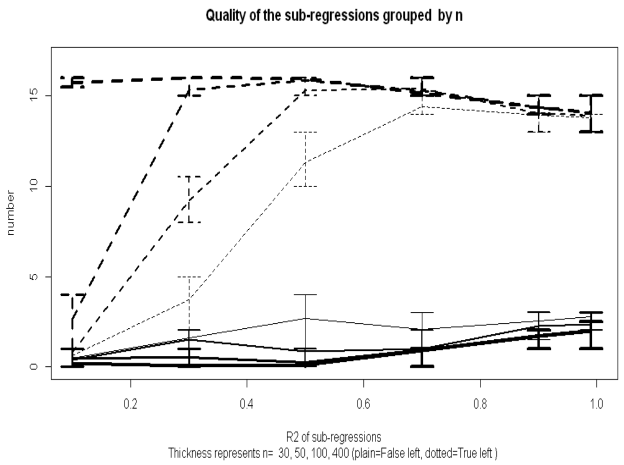
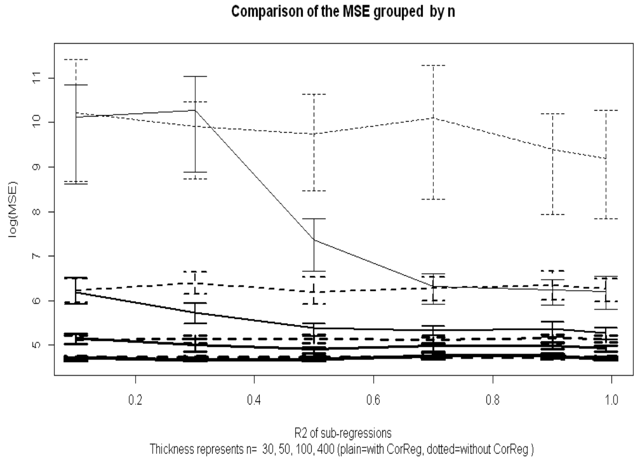


Figure 4.2: Quality of the subregressions found with our BIC_+ criterion

4.5.3 Results on prediction

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



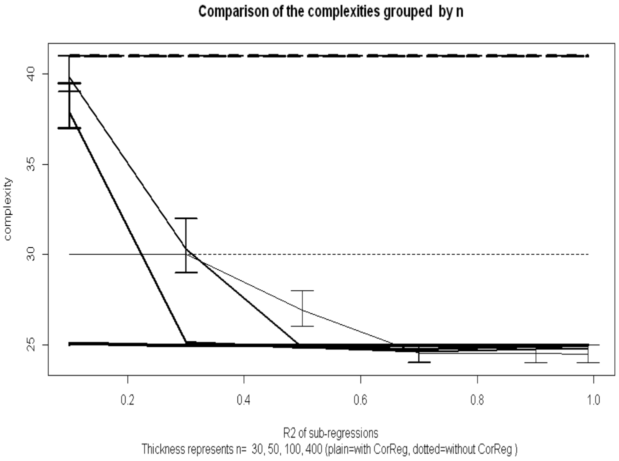
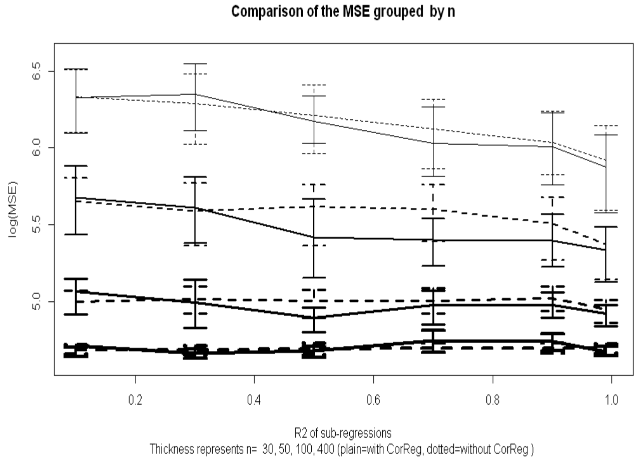


Figure 4.3: Comparison of the MSE between OLS and CorReg+OLS

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



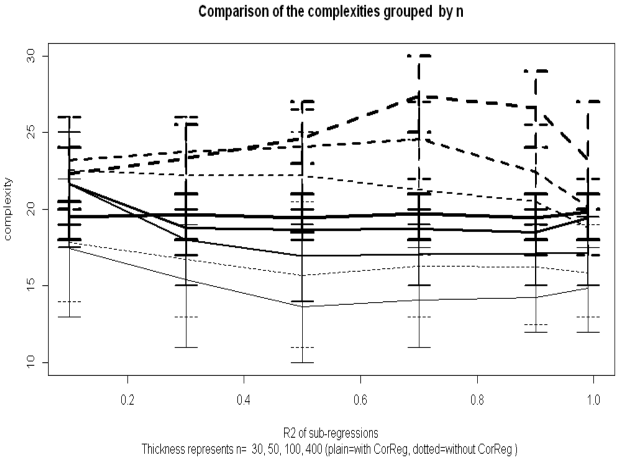


Figure 4.5: Comparison of the MSE between LASSO and CorReg+LASSO

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

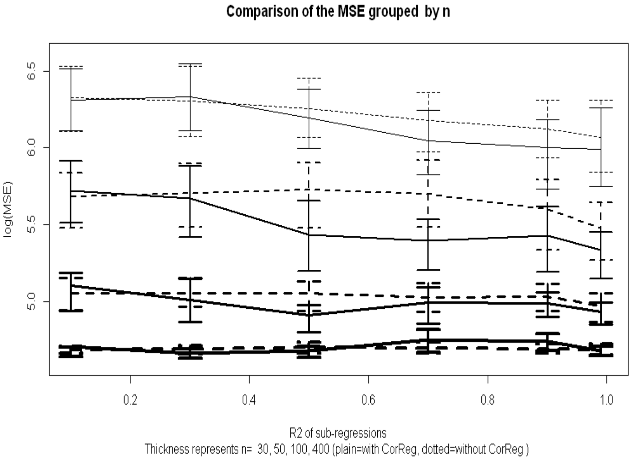


Figure 4.7: Comparison of the MSE between elasticnet and CorReg+elasticnet

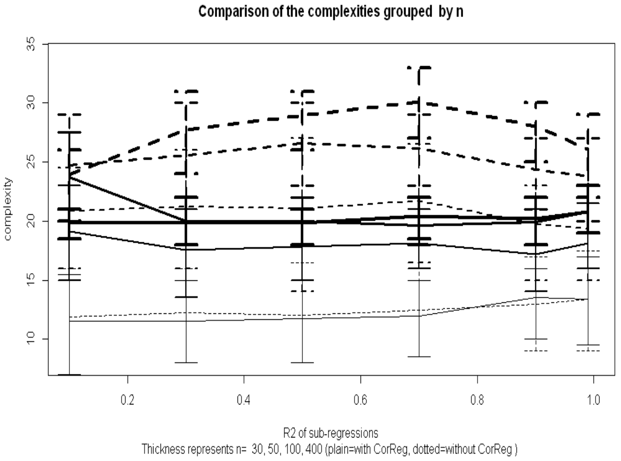


Figure 4.8: Comparison of the compexities between elasticnet and CorReg+elasticnet

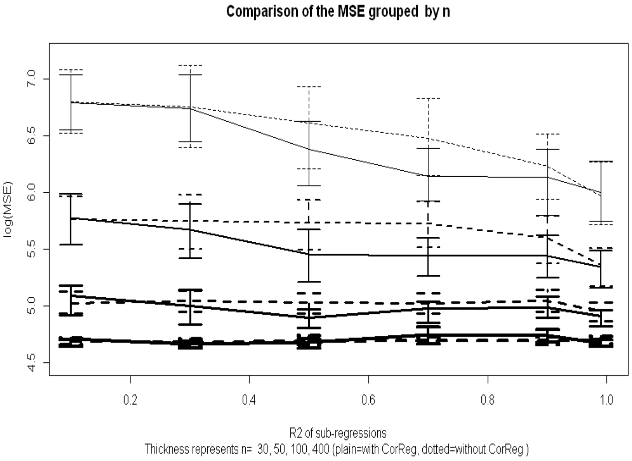


Figure 4.9: Comparison of the MSE between stepwise and CorReg+stepwise

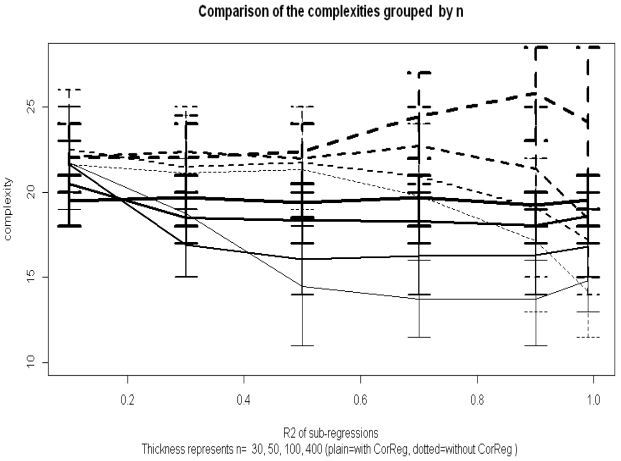


Figure 4.10: Comparison of the compexities between stepwise and CorReg+stepwise

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).

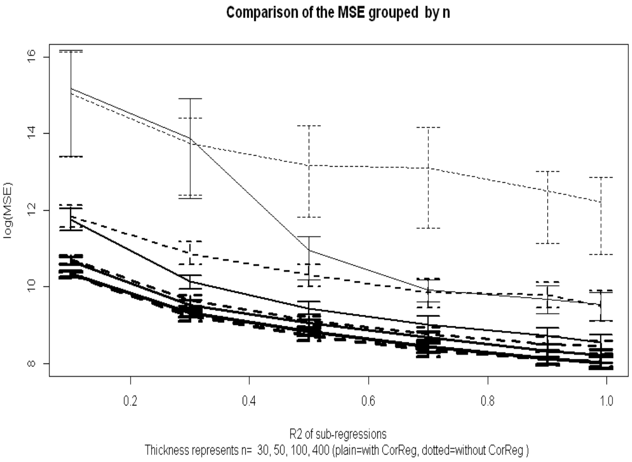


Figure 4.11: Comparison of the MSE between OLS and CorReg+OLS

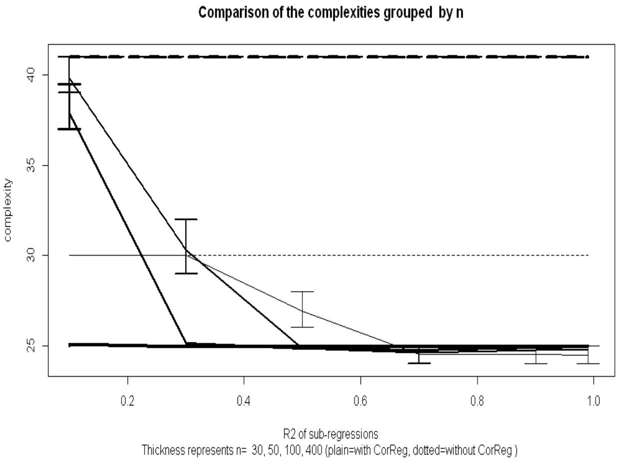


Figure 4.12: Comparison of the compexities between OLS and CorReg+OLS

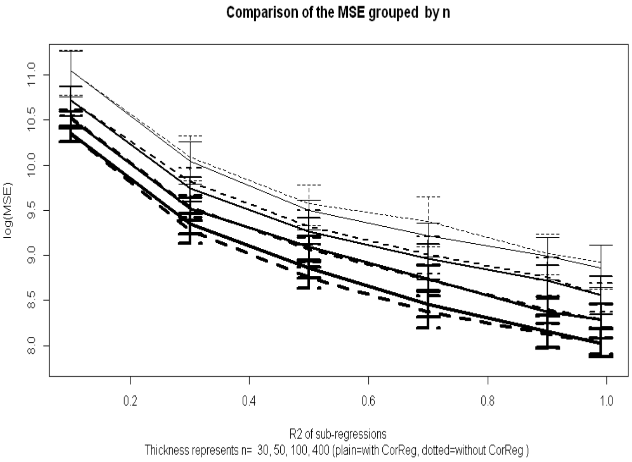


Figure 4.13: Comparison of the MSE between LASSO and CorReg+LASSO

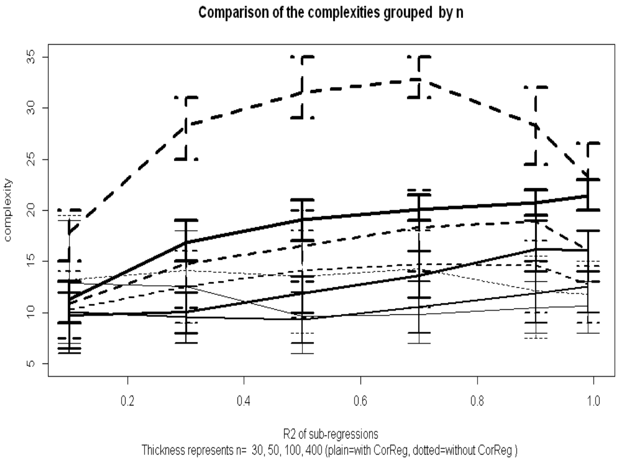


Figure 4.14: Comparison of the compexities between LASSO and CorReg+LASSO

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).

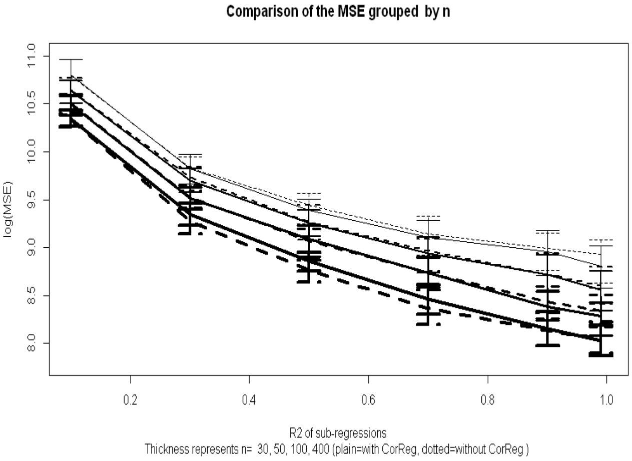


Figure 4.15: Comparison of the MSE between elasticnet and CorReg+elasticnet

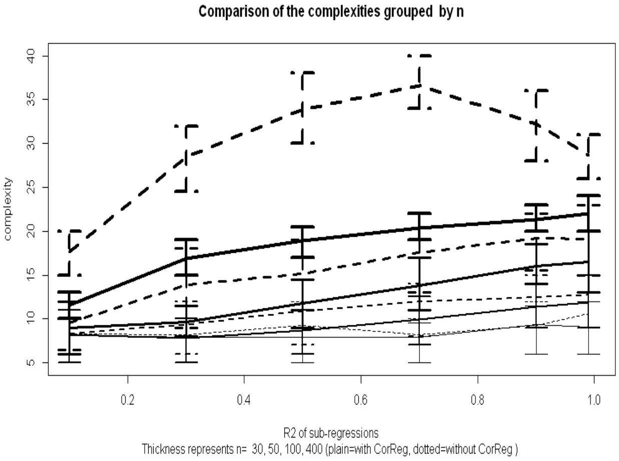


Figure 4.16: Comparison of the compexities beween elasticnet and CorReg+elasticnet

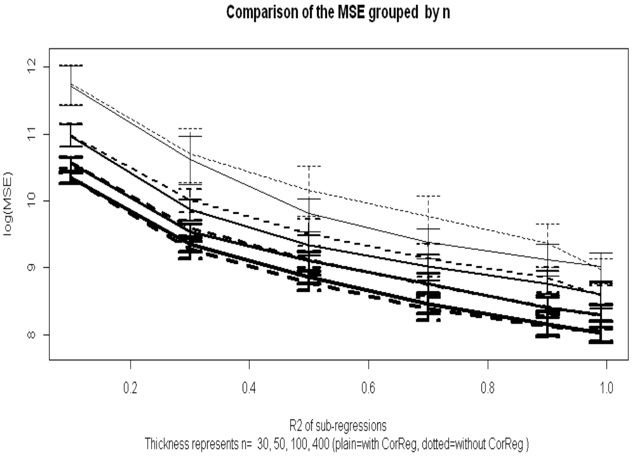


Figure 4.17: Comparison of the MSE between stepwise and CorReg+stepwise

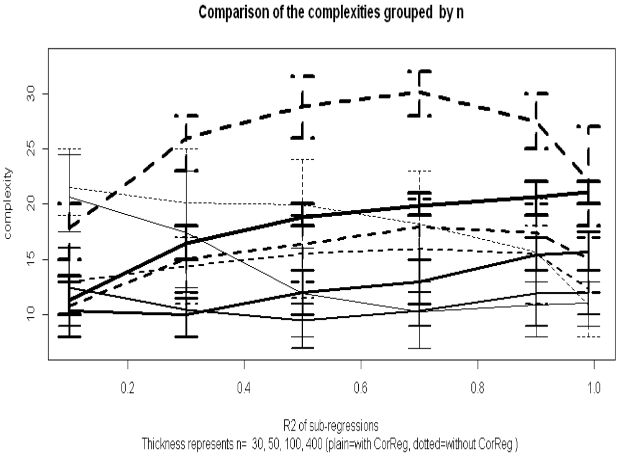


Figure 4.18: Comparison of the compexities beween stepwise and CorReg+stepwise

$\boldsymbol{Y}$ depends only on covariates in $\boldsymbol{X}_r$ (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.

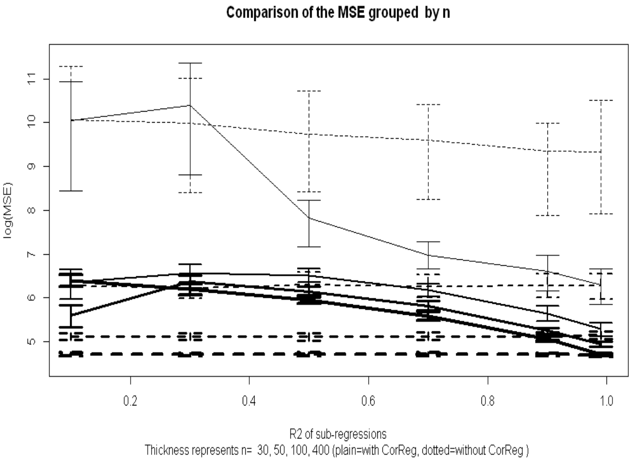


Figure 4.19: Comparison of the MSE between OLS and CorReg+OLS

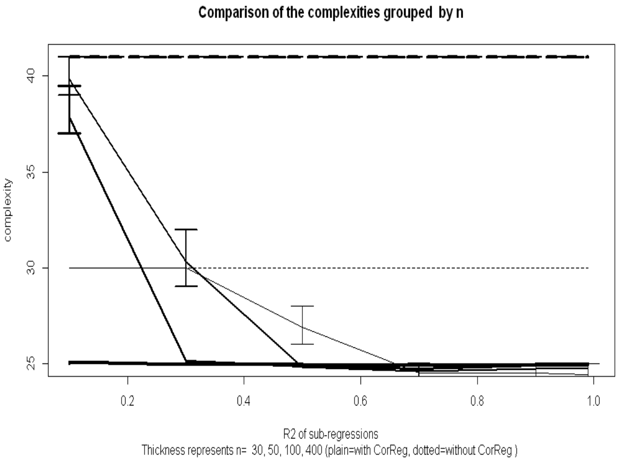


Figure 4.20: Comparison of the compexities between OLS and CorReg+OLS

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

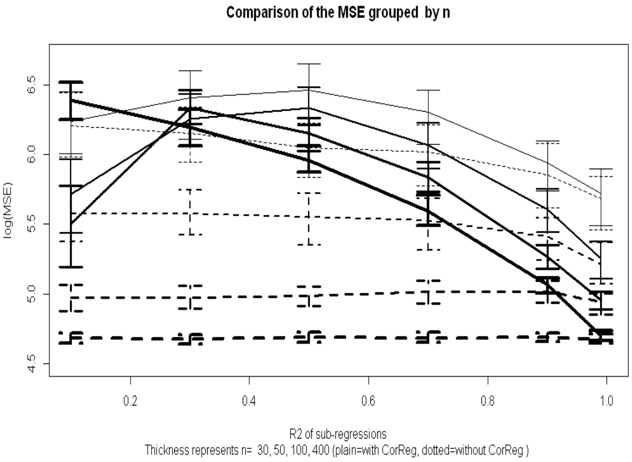


Figure 4.21: Comparison of the MSE between LASSO and CorReg+LASSO

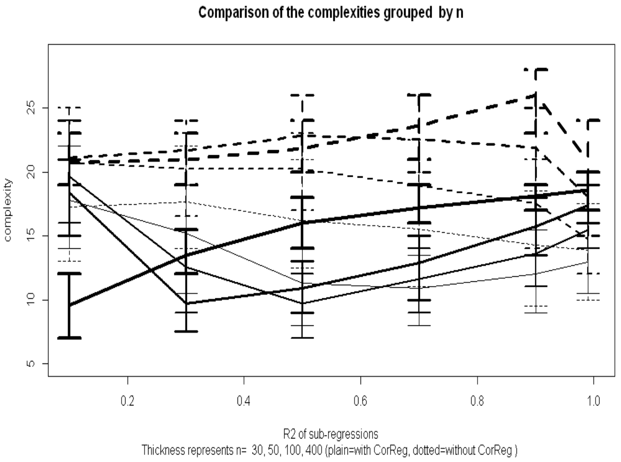


Figure 4.22: Comparison of the compexities between LASSO and CorReg+LASSO

4.6 Numerical results on real datasets

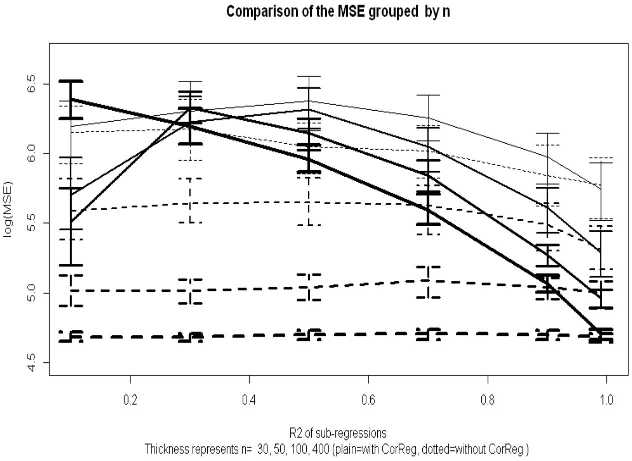


Figure 4.23: Comparison of the MSE between elasticnet and CorReg+elasticnet

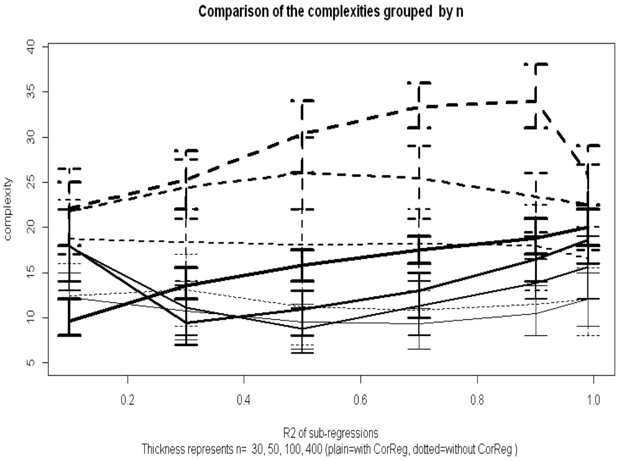


Figure 4.24: Comparison of the compexities betweeen elasticnet and CorReg+elasticnet

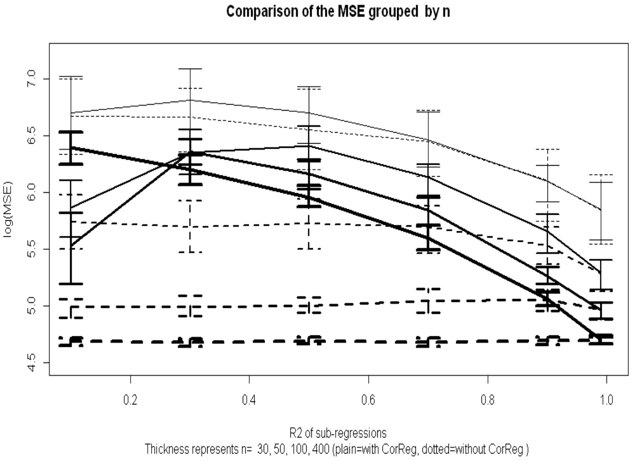


Figure 4.25: Comparison of the MSE between stepwise and CorReg+stepwise

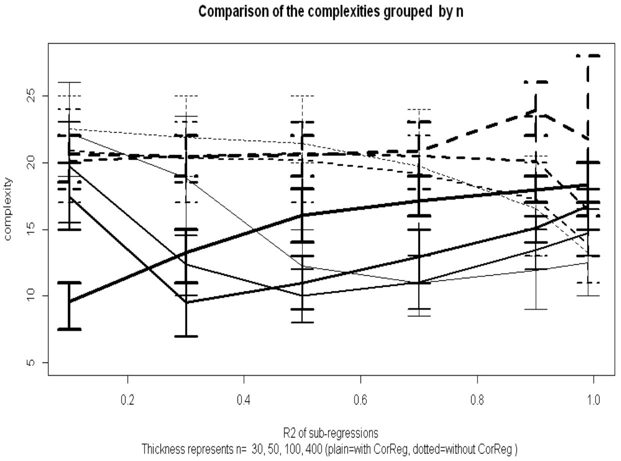


Figure 4.26: Comparison of the compexities betweeen stepwise and CorReg+stepwise

Part II

Further usage of the structure

Chapter 5

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. $\Theta = \{\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X\}$ 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 $\boldsymbol{X}^j$.
Here we suppose that missing values are Missing Completely At Random (MCAR). Many methods does exist to manage such problems [Little, 1992] but they make approximation , add noise (imputation methods) or delete information (cutting methods).

5.1 Some results on missing values and Gaussian mixtures

5.1.1 Decomposition of the integrated likelihood

We start with the complete likelihood of $\boldsymbol{X}$

$$\begin{aligned}
L(\boldsymbol{\alpha},\Theta,S;\boldsymbol{X}) &= \prod_{i=1}^n f(\boldsymbol{X}_i) = \prod_{i=1}^n \left[f(\boldsymbol{X}_i^{I_r}|\boldsymbol{X}_i^{I_f};\boldsymbol{\alpha},\Theta,S) f(\boldsymbol{X}_i^{I_f};\boldsymbol{\alpha},\Theta,S)\right] && (5.1)\\
&= \prod_{i=1}^n \left[\prod_{j\in I_r} f(x_{i,j}|\boldsymbol{X}_i^{I_f};\boldsymbol{\alpha},\Theta,S) \prod_{j\notin I_r} f(x_{i,j};\boldsymbol{\alpha},\Theta,S)\right] && (5.2)\\
&= \prod_{i=1}^n \left[\prod_{j\in I_r} f(x_{i,j}|\boldsymbol{X}_i^{I_f^j};\boldsymbol{\alpha},\Theta,S) \prod_{j\notin I_r} f(x_{i,j};\boldsymbol{\alpha},\Theta,S)\right] && (5.3)\\
\mathcal{L}(\boldsymbol{\alpha},\Theta,S;\boldsymbol{X}) &= \sum_{i=1}^n \left[\sum_{j\in I_r} \log\left(f(x_{i,j}|\boldsymbol{X}_i^{I_f^j};\boldsymbol{\alpha},\Theta,S)\right) + \sum_{j\notin I_r} \log\left(f(x_{i,j};\boldsymbol{\alpha},\Theta,S)\right)\right] && (5.4)
\end{aligned}$$

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 by integration on $\boldsymbol{X}_M$.

We know that $\boldsymbol{X}$ is a Gaussian mixture (*iid* individuals, vectors of orthogonal Gaussian mixtures $\boldsymbol{X}^{I_f}$ and linear combinations of these Gaussian mixtures and some Gaussian for $\boldsymbol{X}^{I_r}$) with K the number of its components.

$$\begin{aligned}
L(\boldsymbol{\alpha},\Theta,S;\boldsymbol{X}_0) &= \int_{\boldsymbol{X}_M} L(\boldsymbol{\alpha},\Theta,S;\boldsymbol{X})d\boldsymbol{X} = \int_{\boldsymbol{X}_M}\sum_{k=1}^K \pi_k\phi_k(\boldsymbol{X};\boldsymbol{\alpha},\Theta,S)d\boldsymbol{X} && (5.5)\\
&= \sum_{k=1}^K \pi_k \int_{\boldsymbol{X}_M}\phi_k(\boldsymbol{X};\boldsymbol{\alpha},\Theta,S)d\boldsymbol{X} = \sum_{k=1}^K \pi_k \int_{\boldsymbol{X}_M}\prod_{i=1}^n \phi_k(\boldsymbol{X}_i;\boldsymbol{\alpha},\Theta,S)d\boldsymbol{X} && (5.6)\\
&= \sum_{k=1}^K \pi_k \prod_{i=1}^n \int_{\boldsymbol{X}_{i,M}}\phi_k(\boldsymbol{X}_i;\boldsymbol{\alpha},\Theta,S)d\boldsymbol{X}_i = \sum_{k=1}^K \pi_k \prod_{i=1}^n \phi_k(\boldsymbol{X}_{i,O};\boldsymbol{\alpha},\Theta,S) && (5.7)\\
&= \sum_{k=1}^K \pi_k\phi_k(\boldsymbol{X}_O;\boldsymbol{\alpha},\Theta,S) = f(\boldsymbol{X}_O,\boldsymbol{\alpha},\Theta,S) && (5.8)
\end{aligned}$$

To compute this likelihood, we will use the decomposition

$$\begin{aligned}
L(\boldsymbol{\alpha},\Theta,S;\boldsymbol{X}_0) &= f(\boldsymbol{X}_O;\boldsymbol{\alpha},\Theta,S) = \prod_{i=1}^n f(\boldsymbol{X}^{I_r}_{i,O}|\boldsymbol{X}^{I_f}_{i,O};\boldsymbol{\alpha},\Theta,S) f(\boldsymbol{X}^{I_f}_{i,O};\boldsymbol{\alpha},\Theta,S) && (5.9)\\
&= \prod_{i=1}^n f(\boldsymbol{X}^{I_r}_{i,O}|\boldsymbol{X}^{I_f}_{i,O};\boldsymbol{\alpha},\Theta,S) \prod_{\substack{j\in I_f\\ M_{i,j}=0}} f(x_{i,j};\boldsymbol{\alpha},\Theta,S) && (5.10)
\end{aligned}$$

with $\forall(i,j)$ with $\boldsymbol{M}_{i,j}=0$ and $j\notin I_r$:

$$f(x_{i,j};\boldsymbol{\alpha},\Theta,S) = \sum_{k=1}^{K_j}\pi_{j,k}\Phi_k(x_{i,j};\mu_{j,k},\Sigma_{j,k}) \quad (5.11)$$

with $K_j,\pi_{j,k},\mu_{j,k},\Sigma_{j,k}$ and the likelihood estimated by Mixmod (for example) once before the MCMC starts.
And, $\forall(i,j)$ with $\boldsymbol{M}_{i,j}=0$ and $j\in I_r$:

$$\begin{aligned}
f(x_{i,j}|\boldsymbol{X}^{I_f^j}_{i,O};\boldsymbol{\alpha},\Theta,S) &= \sum_{k=1}^{K_{ij}}\pi_{ij,k}\Phi(x_{i,j};\mu_{ij,k},\Sigma_{ij,k}) \text{ where} && (5.12)\\
\boldsymbol{\pi}_{ij} &= \bigotimes_{\substack{l\in I_f^j\\ \boldsymbol{M}_{i,l}=1}} \boldsymbol{\pi}_l \text{ and } K_{ij}=|\boldsymbol{\pi}_{ij}|, && (5.13)\\
\boldsymbol{\mu}_{ij} &= \sum_{\substack{l\in I_f^j\\ \boldsymbol{M}_{i,l}=0}} \alpha_{l,j}x_{i,l} + \bigoplus_{\substack{l\in I_f^j\\ \boldsymbol{M}_{i,l}=1}} \alpha_{l,j}\boldsymbol{\mu}_l && (5.14)\\
\boldsymbol{\Sigma}_{ij} &= \sigma_j^2 + \bigoplus_{\substack{l\in I_f^j\\ \boldsymbol{M}_{i,l}=1}} \alpha_{i,l}^2\boldsymbol{\Sigma}_l && (5.15)
\end{aligned}$$

This could be easily used for imputation of the missing values in $\boldsymbol{X}^{I_r}$ knowing the parameters $\boldsymbol{\alpha},\Theta$ and S. We note that we obtain a Gaussian when there is no missing value in I_f^j. But we see that $f(\boldsymbol{X}^{I_r}_{i,O}|\boldsymbol{X}^{I_f}_{i,O};\boldsymbol{\alpha},\Theta,S)$ is not the product of the $f(x_{i,j}|\boldsymbol{X}^{I_f^j}_{i,O};\boldsymbol{\alpha},\Theta,S)$ if a same missing value occurs in distinct sub-regressions. Thus if every sub-regression are distinct connex component then we can use (5.12) and we have

$$L(\boldsymbol{\alpha},\Theta,S;\boldsymbol{X}_0) = \prod_{i=1}^n \prod_{\substack{j\in I_r\\ M_{i,j}=0}} f(x_{i,j}|\boldsymbol{X}^{I_f^j}_{i,O};\boldsymbol{\alpha},\Theta,S) \prod_{\substack{j\in I_f\\ M_{i,j}=0}} f(x_{i,j};\boldsymbol{\alpha},\Theta,S) \quad (5.16)$$

But for the general case we need to manage the dependencies implied by missing values in common covariates in the I_f^j. We note $f(\boldsymbol{X})=\sum_{k=1}^K \pi_k\mathcal{N}(\boldsymbol{\mu}_{X,k};\boldsymbol{\Sigma}_{X,k})$.

$$\begin{aligned}
L(\boldsymbol{\alpha},\Theta,S;\boldsymbol{X}_0) &= \prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k\Phi_k(\boldsymbol{X}_{i,O};\boldsymbol{\alpha},\Theta,S) && (5.17)\\
&= \prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k\Phi_k(\boldsymbol{X}_{i,O}^{I_r}|\boldsymbol{X}_{i,O}^{I_f};\boldsymbol{\alpha},\Theta,S)\Phi_k(\boldsymbol{X}_{i,O}^{I_f};\boldsymbol{\alpha},\Theta,S) && (5.18)\\
&= \prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k\Phi_k(\boldsymbol{X}_{i,O}^{I_r}|\boldsymbol{X}_{i,O}^{I_f};\boldsymbol{\alpha},\Theta,S)\prod_{\substack{j\in I_f\\ M_{i,j}=0}}\Phi_k(x_{i,j};\mu_{j,k},\Sigma_{j,k}) && (5.19)
\end{aligned}$$

Where

$$\begin{aligned}
\boldsymbol{\pi} &= \bigotimes_{j\in I_f}\boldsymbol{\pi}_j \text{ (Kronecker product)} && (5.20)\\
K &= |\boldsymbol{\pi}| && (5.21)\\
\boldsymbol{\mu}_{X^{I_f}} &= \prod_{j\in I_f}\boldsymbol{\mu}_j \text{ (Cartesian product)} && (5.22)\\
\boldsymbol{\sigma}_X &= \prod_{j\in I_f}\boldsymbol{\sigma}_j \text{ (Cartesian product)} && (5.23)
\end{aligned}$$

with $\boldsymbol{\pi}_j,\mu_{j,k},\Sigma_{j,k}$ are estimated once before the MCMC starts (by Mixmod for example).

$\forall 1\le i\le n,\forall 1\le k\le K$ we have

$$\begin{aligned}
\Phi_k(\boldsymbol{X}_{i,O}^{I_r}|\boldsymbol{X}_{i,O}^{I_f};\boldsymbol{\alpha},\Theta,S) &= \Phi_k(\boldsymbol{X}_{i,O}^{I_r}|\boldsymbol{X}_{i,O}^{I_f};\boldsymbol{\mu}_{\boldsymbol{X}_{i,O}^{I_r}|\boldsymbol{X}_{i,O}^{I_f},k},\boldsymbol{\Sigma}_{\boldsymbol{X}_{i,O}^{I_r}|\boldsymbol{X}_{i,O}^{I_f},k}) && (5.24)\\
P(\boldsymbol{X}_{i,O}^{I_r}|\boldsymbol{X}_{i,O}^{I_f};\boldsymbol{\alpha},\Theta,S) &= \Phi_k(\boldsymbol{X}_{i,O}^{I_r};\boldsymbol{\mu}_{\boldsymbol{X}_{i,O}^{I_r}|\boldsymbol{X}_{i,O}^{I_f},k},\boldsymbol{\Sigma}_{\boldsymbol{X}_{i,O}^{I_r}|\boldsymbol{X}_{i,O}^{I_f},k}) && (5.25)\\
\boldsymbol{\mu}_{\boldsymbol{X}_{i,O}^{I_r}|\boldsymbol{X}_{i,O}^{I_f},k} &= \boldsymbol{\mu}_{\boldsymbol{X}_{i,O}^{I_r},k}+\boldsymbol{\Sigma}_{X_{i,O}^{I_r},X_{i,O}^{I_f},k}(\boldsymbol{\Sigma}_{X_{i,O}^{I_f},X_{i,O}^{I_f},k})^{-1}({}^t\boldsymbol{X}_{i,O}^{I_f}-\boldsymbol{\mu}_{X_{i,O}^{I_f},k}) && (5.26)\\
\boldsymbol{\Sigma}_{\boldsymbol{X}_{i,O}^{I_r}|\boldsymbol{X}_{i,O}^{I_f},k} &= \boldsymbol{\Sigma}_{X_{i,O}^{I_r},X_{i,O}^{I_r},k}-\boldsymbol{\Sigma}_{X_{i,O}^{I_r},X_{i,O}^{I_f},k}(\boldsymbol{\Sigma}_{X_{i,O}^{I_f},X_{i,O}^{I_f},k})^{-1}\boldsymbol{\Sigma}_{X_{i,O}^{I_f},X_{i,O}^{I_r},k} && (5.27)\\
\forall j\in I_r:\quad \boldsymbol{\mu}_{X_{i,O}^j} &= \sum_{l\in I_f^j}\alpha_{l,j}\mu_{l,k} && (5.28)
\end{aligned}$$

$\forall j\in I_r$ with $M_{i,j}=0$

$$\text{var}_k(x_{i,j})=\sigma_j^2+\sum_{l\in I_f^j}\alpha_{l,j}^2\sigma_{X^l,k}^2 \qquad (5.29)$$

$\forall j\notin I_r$ with $M_{i,j}=0$

$$\text{var}_k(x_{i,j})=\sigma_{X^j,k}^2 \qquad (5.30)$$

$\forall j_1\in I_r,j_2\in I_r,I_f^{j_1}\cap I_f^{j_2}\neq\emptyset$ with $M_{i,j_1}=M_{i,j_2}=0$

$$\text{cov}_k(x_{i,j_1},x_{i,j_2})=\sum_{l\in I_f^{j_1}\cap I_f^{j_2}}\alpha_{l,j_1}\alpha_{l,j_2}\,\text{var}_k(x_{i,l})=\sum_{l\in I_f^{j_1}\cap I_f^{j_2}}\alpha_{l,j_1}\alpha_{l,j_2}\sigma_{X^l,k}^2 \qquad (5.31)$$

$\forall j_1\in I_r,j_2\in I_r,I_f^{j_1}\cap I_f^{j_2}=\emptyset$ with $M_{i,j_1}=M_{i,j_2}=0$

$$\text{cov}_k(x_{i,j_1},x_{i,j_2})=0 \qquad (5.32)$$

$\forall j_1\in I_f,j_2\in I_f$ with $M_{i,j_1}=M_{i,j_2}=0$

$$\text{cov}_k(x_{i,j_1},x_{i,j_2})=0 \qquad (5.33)$$

$\forall j_1\in I_r,j_2\in I_f^{j_1}$ with $M_{i,j_1}=M_{i,j_2}=0$

$$\text{cov}_k(x_{i,j_1},x_{i,j_2})=\alpha_{j_2,j_1}\sigma_{X^{j_2},k}^2 \qquad (5.34)$$

$\forall j_1 \in I_r, j_2 \notin I_f^{j_1} \cup I_r$ with $M_{i,j_1} = M_{i,j_2} = 0$

$$\mathrm{cov}_k(x_{i,j_1}, x_{i,j_2}) = 0 \tag{5.35}$$

We see that the 0 in the variance-covariance matrix does not depend on the component k so the structure of sparsity of $\boldsymbol{\Sigma}$ can be stored and used back in each iteration for a given structure S to reduce computing time.

5.1.2 Likelihood computation optimized

The main problem with the likelihood in its global form (5.19) is that the number of components explodes so we can't use it in practice. But in many case, it can be simplified. First, we can look if there are missing values shared by several sub-regression. We just need to compute the row-sums of the adjacency matrix or to search for redundancy in I_f and then if there is no redundancy or if $\forall j$ redundant we have $\sum_{i=1}^n M_{i,j} = 0$ then we can use the simplified form of the likelihood given in (5.16). For faster computation we can stock the vector of covariates that have missing values. So the true value of the likelihood can be computed in most of cases but in the MCMC, it remains the possibility to have a structure with explosive likelihood expression when combined with the missing values. Then we propose in the package to use the simplified form of the likelihood, that can be seen as an approximation of the likelihood. Numerical results on simulated datasets will show if this approximation is effective.

5.2 SEM

The likelihood depends on $\boldsymbol{\alpha}$ which was formerly estimated by OLS when there was no missing values. Here we use a Stochastic Expectation Maximization (SEM) algorithm [Celeux and Diebolt, 1986] to estimate $\boldsymbol{\alpha}$ because missing values do not allow to use OLS and the log-likelihood (5.4) is not linear so a simple Expectation-Maximization (EM) would be difficult to compute.

initialization: We start with imputation by the mean for each missing value (done only once for the MCMC). $\boldsymbol{\alpha}^{(0)}$ can be initialized by cutting method (sparse structure) or using imputed values in $\boldsymbol{X}$. At iteration h,

SE step: We generate the missing values according to $P(\boldsymbol{X}_M|\boldsymbol{X}_O;\alpha^{(h)},\Theta,S)$, that is stochastic imputation.

M step: We estimate

$$\boldsymbol{\alpha}^{(h+1)} = \mathrm{argmax}_{\boldsymbol{\alpha}}\, E\left[\mathcal{L}(\boldsymbol{X}|\boldsymbol{\alpha},S,\Theta)\right] \tag{5.36}$$

and we can use the same method as the one for classical case without missing values (OLS, SUR, *etc.*). We continue until convergence ($\parallel \boldsymbol{\alpha}^{(h+1)} - \boldsymbol{\alpha}^{(h)} \parallel < tol$ where tol is the tolerance). Then we make m iterations and take $\hat{\boldsymbol{\alpha}}$ as the mean of these m last iterations.

5.2.1 Stochastic imputation by Gibbs sampling

We use a Gibbs sampling method to generate the missing values at the SE step. $\boldsymbol{X}$ follows a multivariate Gaussian mixture with K component and we note Z the set of the $Z_{i,j}$ indicating the component from which $x_{i,j}$ is generated.

Initialisation: all the $z_{i,j}$ are set to the first component (such an initialisation does not depend on K) and $\boldsymbol{X}_M$ are imputed by the marginal means.

Iteration: At each iteration of the Gibbs sampler:
$\forall x_{i,j} \in \boldsymbol{X}_M^{I_r}$: $x_{i,j}$ is generated according to

$$P(x_{i,j}|\boldsymbol{X}_{i,O}, \boldsymbol{X}_{i,\bar{M}_{i,j}}, Z; \boldsymbol{\alpha}^{(h)}, \Theta, S) = P(x_{i,j}|\boldsymbol{X}_{i,O}, \boldsymbol{X}_{i,\bar{M}_{i,j}}; \alpha^{(h)}, \Theta, S) \quad (5.37)$$

$$= P(x_{i,j}|\boldsymbol{X}_i^{I_f^j}; \boldsymbol{\alpha}^{(h)}, \Theta, S) = \mathcal{N}(\boldsymbol{X}_i^{I_f^j} \boldsymbol{\alpha}_{I_f^j,j}^{(h)}; \sigma_j^2) \quad (5.38)$$

We have $P(\boldsymbol{X}|Z) = \mathcal{N}(\boldsymbol{\mu}_{|Z}, \boldsymbol{\Sigma}_{|Z})$.
$\forall x_{i,j} \in \boldsymbol{X}_M^{I_f}$: $x_{i,j}$ is generated according to

$$P(x_{i,j}|\boldsymbol{X}_{i,O}, \boldsymbol{X}_{i,\bar{M}_{i,j}}, Z; \boldsymbol{\alpha}^{(h)}, \Theta, S) = P(x_{i,j}|\boldsymbol{X}_{i,\bar{j}}, Z_i; \boldsymbol{\alpha}^{(h)}, \Theta, S) \quad (5.39)$$

$$= \mathcal{N}(\mu_{j|Z_{i,j}} + \boldsymbol{\Sigma}_{j,X_{i\bar{j}}|Z_i} \boldsymbol{\Sigma}^{-1}_{X_{i\bar{j}},X_{i\bar{j}}|Z_i} (X_{i\bar{j}} - \boldsymbol{\mu}_{X_{i\bar{j}}|Z_i}); \sigma^2_{j|Z_{i,j}} - \boldsymbol{\Sigma}_{j,X_{i\bar{j}}|Z_i} \boldsymbol{\Sigma}^{-1}_{X_{i\bar{j}},X_{i\bar{j}}|Z_i} \boldsymbol{\Sigma}'_{j,X_{i\bar{j}}|Z_i}) \quad (5.40)$$

Where all the values needed here were described above for the likelihood computation.

Then, $\forall 1 \leq i \leq n, \forall j \in I_f$ we draw new values for $Z_{i,j}$ according to

$$P(Z_{i,j}|\boldsymbol{X}, Z_{\bar{i,j}}; \Theta, \boldsymbol{\alpha}, S) = P(Z_{i,j}|\boldsymbol{X}_i, Z_{i,\bar{j}}; \Theta, \boldsymbol{\alpha}, S) = \mathcal{M}(t_{i,j,1}, \dots, t_{i,j,K_j}) \quad (5.41)$$

$$\text{where } t_{i,j,k} = \frac{\pi_{j,k}\Phi(x_{i,j}; \mu_{j,k}, \sigma^2_{j,k})}{\sum_{l=1}^{K_j} \pi_{j,l}\Phi(x_{i,j}; \mu_{j,l}, \sigma^2_{j,l})} \quad (5.42)$$

We see that $Z_{i,j}$ are not used if there is no missing values in $\boldsymbol{X}_i$ and others are not all needed so we can also optimize computation time by computing only the $Z_{i,j}$ that are needed in the Gibbs. For the last iteration of the Gibbs, in the last iteration of the SEM, we do not need to draw Z.

Instead of using long chain for each Gibbs, we can use small chains because SEM iteration will simulate longer chains so it remains efficient with a smaller computation cost.

Computation cost will be the main purpose here because we need an iterative algorithm (Gibbs sampler) at each iteration of another iterative algorithm (SEM) for each candidate of the MCMC. So alternative method should be preferred for large datasets with many missing values and only a small amount of time.

Because K can be very large we search a way to compute the likelihood. We can use a Gibbs algorithm to estimate the likelihood:

$$P(\boldsymbol{X}_O; \Theta, S, \boldsymbol{\alpha}) = \sum_{Z \in \mathcal{Z}} \int_{\boldsymbol{X}_M} \frac{P(\boldsymbol{X}_M, Z, \boldsymbol{X}_O; \Theta, \boldsymbol{\alpha}, S)}{P(\boldsymbol{X}_M, Z|\boldsymbol{X}_O; \Theta, \boldsymbol{\alpha}, S)} P(\boldsymbol{X}_M, Z|\boldsymbol{X}_O; \Theta, \boldsymbol{\alpha}, S) dX \quad (5.43)$$

$$\approx \frac{1}{Q} \sum_{q=1}^{Q} \frac{P(\boldsymbol{X}_M^{(q)}, \boldsymbol{X}_O, Z^{(q)}; \Theta, \boldsymbol{\alpha}, S)}{P(\boldsymbol{X}_M^{(q)}|\boldsymbol{X}_O, Z^{(q)}; \Theta, \boldsymbol{\alpha}, S)} \text{ by the law of large numbers} \quad (5.44)$$

where Q is the number of iterations of the Gibbs sampler. But to be faster, we use the previous Gibbs algorithm with:

$$P(\boldsymbol{X}_O; \Theta, S, \boldsymbol{\alpha}) = \frac{1}{Q} \sum_{q=1}^{Q} \frac{P(\boldsymbol{X}_M^{(q)}, \boldsymbol{X}_O, Z^{(q)}; \Theta, \boldsymbol{\alpha}^{(q)}, S)}{P(\boldsymbol{X}_M^{(q)}|\boldsymbol{X}_O, Z^{(q)}; \Theta, \boldsymbol{\alpha}^{(q)}, S)} \quad (5.45)$$

5.2.2 Alternative E step

If we can't (or don't want to) compute the SE step described above, then we can use alternative imputation step for missing data based on $\boldsymbol{\alpha}$ (and keep the alternate optimisation to find the best $\boldsymbol{\alpha}$).

$\forall x_{i,j} \in \boldsymbol{X}_M$ we have:
if $j \in I_r$, Equation(5.12) gives:

$$E[x_{i,j}|\boldsymbol{\alpha}^{(h)}, \Theta, \boldsymbol{X}_O, S] = E[\sum_{k=1}^{k_{ij}} \pi_{ij,k}\Phi(x_{i,j}|\mu_{ij,k}, \Sigma_{ij,k})|\boldsymbol{\alpha}^{(h)}, \Theta, \boldsymbol{X}_O, S] \quad (5.46)$$

Let $r_{i,j} = \{l \in I_r | \boldsymbol{\alpha}_{j,l} \neq 0, \boldsymbol{M}_{i,j} = 0\}$ the set of observed covariates for individual i that are explained by $x_{i,j}$ according to S.
If $j \notin I_r$ we can do:

$$\begin{aligned} E[x_{i,j}|\boldsymbol{\alpha}^{(h)}, \Theta, \boldsymbol{X}_O, S] &= \frac{1}{|r_{i,j}|} \sum_{k \in r_{i,j}} E_{|\boldsymbol{\alpha}^{(h)},\Theta,\boldsymbol{X}_O,S} \left[\frac{1}{\alpha_{j,k}} \left(x_{i,k} - \varepsilon_k(i) - \sum_{l \in I_f^k} x_{i,l}\alpha_{l,k} \right) \right] \quad (5.47) \\ &= \frac{1}{|r_{i,j}|} \sum_{k \in r_{i,j}} E_{|\boldsymbol{\alpha}^{(h)},\Theta,\boldsymbol{X}_O,S} \left[\frac{1}{\alpha_{j,k}} \left(x_{i,k} - \sum_{l \in I_f^k} x_{i,l}\alpha_{l,k} \right) \right] \quad (5.48) \end{aligned}$$

that is the mean of the expectations of the inverse sub-regressions implying x_i, j with value in $\boldsymbol{X}_i^{I_r}$ not missing.

Another way is to only use the structure for $\boldsymbol{X}^{I_r}$ and use the distribution given by Mixmod for $\boldsymbol{X}^{I_f}$ along the MCMC. The full SEM would then be used only once with the final structure to make imputation in $\boldsymbol{X}$ before using variable selection methods like the LASSO.

5.2.3 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.

$$k_j = \frac{1}{n} \sum_{i=1}^{n} k_{i,j} \quad (5.49)$$

where $k_{i,j}$ is the number of parameter to estimate in $P(x_{i,j}|\boldsymbol{X}_i \setminus \boldsymbol{X}_i^j)$.

$$\begin{aligned} -2 \log P(\boldsymbol{X}|S) &\approx BIC = -2\mathcal{L}(\boldsymbol{X}, S, \boldsymbol{\Theta}) + |\boldsymbol{\Theta}| \log(n) \quad (5.50) \\ &= -2\mathcal{L}(\boldsymbol{X}, S, \boldsymbol{\Theta}) + (\sum_{j=1}^{p} k_j) \log(n) \quad (5.51) \end{aligned}$$

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 $\frac{k_i \log(p)}{\log(n)}$.

5.2.4 new criterion ?

In fact we have the same number of parameters to estimate with or without missing values but the problem is that some of the parameters are estimated based only on a portion of the individuals so each parameter has a different weight. The penalty in $k \log(n)$ stands for k parameters each $n - estimated$

5.3 Missing values in the main regression

The easier way would be to draw missing values with the SEM described above and then use classical methods on the completed dataset, with the possibility to repeat this procedure a few times and then take the mean. We should for example try multiple draw and LASSO for variable selection like variable selection by random forest. One great advantage of multiple drow procedures is that it gives an idea of the precision of the imputations with the variance of these imputed values among the multiple draws. So we know whether it is reliable or not.

But another way would be to consider classical estimation methods as likelihood optimizer and then adapt them to the integrated likelihood of our model. Thus we can imagine to use LASSO without imputation. But the choice of the penalty using the LAR algorithm need also to adapt the LAR that is based on correlations that are computed on vectors with distinct number of individuals (due to missing values). So it requires a bit more reflexion but could be a good perspective for our method.

5.4 Numerical results

5.4.1 Finding the structure

5.4.2 Efficiency for main regression

5.5 Missing values in real life

One advantage of our regression model is that it does not depend on the response variable $\boldsymbol{Y}$ so the structure can be learnt independently. Thus we can imagine to obtain big samples to learn the structure without being annoyed by the missing values. Then when a response variable is chosen,

Chapter 6

Taking back the residuals

We have seen that eviction of redundant covariates improves the results by a good trade-off between dimension reduction and better conditioning versus keeping all the information. But The fact is that we lost some information and we want to get it back.

6.1 The model

After the estimation of the marginal model, we know both $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}^*$.

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}^{I_r}\boldsymbol{\beta}_{I_r} + \boldsymbol{X}^{I_f}\boldsymbol{\beta}_{I_f} + \boldsymbol{\varepsilon}_Y & (6.1)\\ \boldsymbol{X}^{I_r} &= \boldsymbol{X}^{I_f}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} & (6.2)\\ \boldsymbol{Y} &= \boldsymbol{X}^{I_r}\underbrace{(\boldsymbol{\beta}_{I_r} + \boldsymbol{\alpha}\boldsymbol{\beta}_{I_f})}_{\boldsymbol{\beta}^*} + \boldsymbol{\varepsilon}\boldsymbol{\beta}_{I_r} + \boldsymbol{\varepsilon}_Y & (6.3)\\ \boldsymbol{Y} - \boldsymbol{X}^{I_r}\boldsymbol{\beta}^* &= \boldsymbol{\varepsilon}\boldsymbol{\beta}_{I_r} + \boldsymbol{\varepsilon}_Y & (6.4)\\ \boldsymbol{\varepsilon} &= \boldsymbol{X}^{I_r} - \boldsymbol{X}^{I_f}\boldsymbol{\alpha} & (6.5) \end{aligned}$$

So we introduce a plug-in model

$$\begin{aligned} \underbrace{\boldsymbol{Y} - \boldsymbol{X}^{I_r}\hat{\boldsymbol{\beta}}^*}_{\tilde{\boldsymbol{Y}}} &= \underbrace{(\boldsymbol{X}^{I_r} - \boldsymbol{X}^{I_f}\hat{\boldsymbol{\alpha}})}_{\tilde{\boldsymbol{X}}}\boldsymbol{\beta}_{I_r} + \boldsymbol{\varepsilon}_Y & (6.6)\\ & & (6.7) \end{aligned}$$

That allows us to estimate $\boldsymbol{\beta}_{I_r}$ with a classical linear model based on previous estimations of $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}$. Then we have a model with a smaller noise

$$\boldsymbol{Y} = \boldsymbol{X}^{I_r}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}\hat{\boldsymbol{\beta}}_{I_r} + \boldsymbol{\varepsilon}_Y \qquad (6.8)$$

6.2 Interpretation and latent variables

$\hat{\boldsymbol{\beta}}_{I_r}$ can be interpreted as the proper effect of $\boldsymbol{X}^{I_r}$ on $\boldsymbol{Y}$ in that it is the effect of the part of $\boldsymbol{X}^{I_r}$ that is independent from other covariates. Then if $\boldsymbol{X}^{I_r}$ is correlated to $\boldsymbol{Y}$ only through its correlation with $\boldsymbol{X}^{I_f}$ this sequential estimation will point it out and give a parsimonious model ($\hat{\boldsymbol{\beta}}_{I_r} = 0$) but the real stake is greater. We can see $\boldsymbol{\varepsilon}$ as a latent variable instead of the noise of a sub-regression. But this latent variable is known to be independent of $\boldsymbol{X}^{I_f}$ and dependent of $\boldsymbol{X}^{I_r}$ so we can appreciate its meaning and we also know its value by $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{X}^{I_r} - \boldsymbol{X}^{I_f}\hat{\boldsymbol{\alpha}}$. Thus, it reveals latent variables.

6.3 Consistency

6.3.1 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.

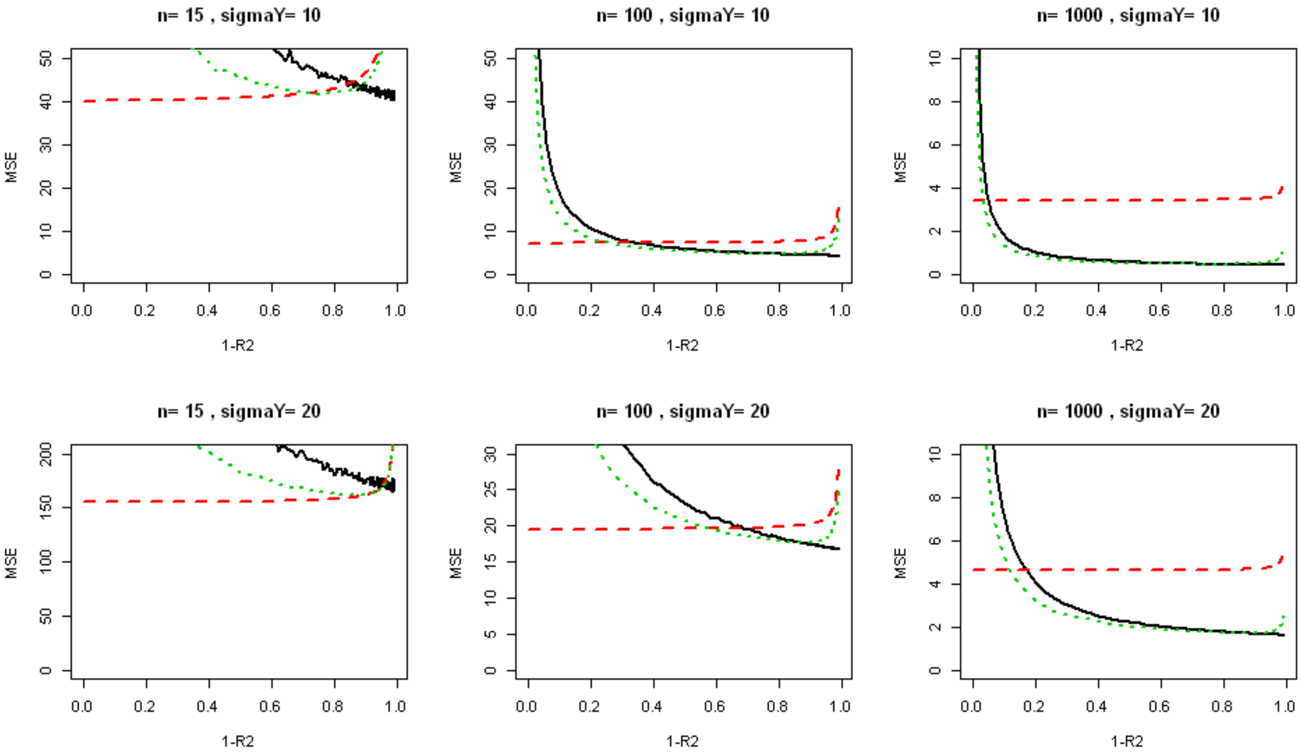


Figure 6.1: MSE of OLS (plain black) and CorReg marginal(red dashed) and CorReg full (green dotted) estimators for varying $(1 - R^2)$ of the sub-regression, n and σ_Y.

Here $p = 3$ and $n = 1000$.We define $\boldsymbol{X}_f, \boldsymbol{X}_r, \varepsilon_Y, \varepsilon_X i.i.d. \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ 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 6.1: 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 6.2: 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 6.3: 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 6.4: 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).

6.4 Numerical results

Chapter 7

CorReg: the R package

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).

Chapter 8

Conclusion and perspectives

8.1 Conclusion

Our model is easy to understand and to use. Usage of linear regression to model the correlations definitely separates us from ”black boxes” so users are confident in what they do. The well-known and trivial sub-regression found comfort users in that if a structure does exist, CoMPASS will find it so when a new sub-regression, or a new main regression is given they are more likely to look further and try it. The automated aspect shows the power of statistics without a priori so users begin to understand that statistics are not only descriptive or predictive but based on *a priori* models. This method has a positive impact on the way users looks at the statistics. It is good to see that sequential methods (predictive model) and automation can produce good results. Probabilistic models are efficient even without human expertise and let the experts improve the results by adding their expertise in the model (coercing some sub-regression for example).

8.2 Perspective

8.2.1 Non-linear regression

Polynomial regression, logistic regression, *etc.* could be improved by a method like this.

8.2.2 Pretreatment not only for regression

Classification and Regression Tree, and any other method could benefit of the variable selection pretreatment implied by our marginal model.

8.2.3 Improved programming

Even if it is written in C++, the algorithm could be optimized by a better usage of sparse matrices, memory usage optimization, and other small things that could reduce computational cost to be faster and allow to work with larger datasets (already works with thousands of covariates).

8.2.4 Missing values in classical methods

The full generative approach could be used to manage missing values without imputation for many classical methods.

8.2.5 Interpretation improvements

Ergonomy of the software could be improved to better fit industrial needs.

Bibliography

[Andrieu and Doucet, 1999] Andrieu, C. and Doucet, A. (1999). Joint bayesian model selection and estimation of noisy sinusoids via reversible jump mcmc. *Signal Processing, IEEE Transactions on*, 47(10):2667–2676.

[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.

[Celeux and Diebolt, 1986] Celeux, G. and Diebolt, J. (1986). L'algorithme sem: un algorithme d'apprentissage probabiliste: pour la reconnaissance de mélange de densités. *Revue de statistique appliquée*, 34(2):35–52.

[Chipman et al., 2001] Chipman, H., George, E., McCulloch, R., Clyde, M., Foster, D., and Stine, R. (2001). The practical implementation of bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134.

[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.*

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

[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.

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

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

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

[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.

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

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

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

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

[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.

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

[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.

[Timm, 2002] Timm, N. (2002). *Applied multivariate analysis*. Springer Verlag.

[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