

CoMPASS: **C**orrelation **M**odeling for **P**retreatment by **A**utomated **S**tructure **S**election

Clément THERY

June 18, 2014

To my sons

Contents

1	The industrial context	6
2	State of the art	7
2.1	Ordinary least squares and associated problems	7
2.2	Penalized models	7
2.2.1	Ridge regression	7
2.2.2	LASSO: Least Absolute Shrinkage and Selection Operator	8
2.2.3	Adaptive LASSO and Random LASSO	8
2.2.4	Elasticnet	8
2.2.5	OSCAR: Octogonal Shrinkage and Clustering Algorithm for Regression	8
2.3	Modeling the parameters	8
2.3.1	CLERE: CLusterwise Effect REgression	8
2.3.2	Spike and Slab	9
2.4	Multiple Equations	9
2.4.1	SEM and Path Analysis	9
2.4.2	SUR: Seemingly Unrelated Regression	9
2.4.3	SPRING: Structured selection of Primordial Relationships IN the General linear model	9
2.4.4	Selvarclust: Linear regression within covariates for clustering	9
I	Pretreatment for correlations	10
3	Decorrelating covariates by a generative model	11
3.1	Our proposal: modelisation of the correlations	11
3.2	A by-product model: marginal regression with decorrelated covariates	12
3.3	Strategy of use: pre-treatment before classical estimation/selection methods	12
3.4	Illustration of the tradeoff conveyed by the pre-treatment	13
4	Estimation of the Structure of subregression by MCMC	15
4.1	How to compare structures ?	15
4.1.1	Bayesian criterion for quality	15
4.1.2	Some indicators for proximity	15
4.2	Neighbourhood	15
4.2.1	Classical	15
4.2.2	Active relaxation of the constraints	15
4.3	The walk	15
4.4	Numerical results on simulated datasets	15
4.4.1	The datasets	15
4.4.2	Results on prediction	17

II	Further usage of the structure	23
5	Missing values	24
5.1	State of the art	24
5.2	How to manage missing values in the MCMC ?	24
5.2.1	Decomposition of the integrated likelihood	24
5.2.2	Alternative E step	27
5.3	SEM	27
5.3.1	SE step in details	28
5.3.2	Estimation of the coefficients in each regression	29
5.3.3	Weighted penalty	30
5.4	Missing values in the main regression	30
5.4.1	explicative	30
5.4.2	predictive	30
6	Taking back the residuals	31
6.1	The model	31
6.2	Properties	31
6.3	Consistency	31
6.3.1	Consistency Issues	31
6.4	Numerical results	32
7	CorReg: the R package	33
8	Conclusion and perspectives	34
	References	35
	Appendices	36
A	Graphs and CorReg	37
A.1	Matricial notations	37
A.2	Properties	37
B	Mixture models	38
B.1	Linear combination	38
B.2	Industrial examples	38

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: $\| \beta \|_2^2 = \sum_{i=1}^p (\beta_i)^2$, $\| \beta \|_1 = \sum_{i=1}^p |\beta_i|$ and $\| \beta \|_\infty = \max(|\beta_1|, \dots, |\beta_p|)$.

2.1 Ordinary least squares and associated problems

We note the linear regression model:

$$Y_{|X} = X\beta + \varepsilon \quad (2.1)$$

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

$$\hat{\beta} = (X'X)^{-1} X'Y \quad (2.2)$$

with variance matrix

$$\text{Var}(\hat{\beta}_{OLS}) = \sigma_Y^2 (X'X)^{-1} \quad (2.3)$$

and without any bias. Estimation of β requires the inversion of $X'X$ which will be ill-conditioned or even singular if some covariates depend linearly from each other. Conditionning of $X'X$ 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:

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

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 $\hat{\beta}$:

$$\hat{\beta} = \operatorname{argmin} \left\{ \| \mathbf{Y} - \mathbf{X}\beta \|_2^2 \right\} \text{ subject to } \| \beta \|_1 \leq \lambda \text{ with } \lambda > 0 \quad (2.5)$$

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:

$$\hat{\beta} = (1 + \lambda_2) \operatorname{argmin} \left\{ \| \mathbf{Y} - \mathbf{X}\beta \|_2^2 \right\}, \text{ subject to } (1 - \alpha) \| \beta \|_1 + \alpha \| \beta \|_2^2 \leq t \text{ for some } t \quad (2.6)$$

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).

$$\hat{\beta} = \operatorname{argmin}_{\beta} \| \mathbf{Y} - \mathbf{X}\beta \|_2^2 \text{ subject to } \sum_{j=1}^p |\beta_j| + c \sum_{j < k} \max(|\beta_j|, |\beta_k|) \leq \lambda \quad (2.7)$$

But OSCAR depends on two tuning parameters: c and 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 implicitly 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 yield reliable prediction with a smaller variance on $\hat{\beta}$.

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 $\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{x}_3 = \mathbf{x}_1 + \mathbf{x}_2 + \boldsymbol{\varepsilon}_3$ where $\boldsymbol{\varepsilon}_3 \sim \mathcal{N}(\mathbf{0}, \sigma_3^2 \mathbf{I}_n)$. We also define another couple $\mathbf{x}_4, \mathbf{x}_5$ of covariates that are *i.i.d.* with $(\mathbf{x}_1, \mathbf{x}_2)$ and two *scenarii* for \mathbf{Y} with $\boldsymbol{\beta} = (1, 1, 1, 1, 1)$ and $\sigma_Y \in \{10, 20\}$. It is clear that $\mathbf{X}'\mathbf{X}$ will become more ill-conditioned as σ_3 gets smaller.

3.1 Our proposal: modelisation of the correlations

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

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

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

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 \mathbf{X} implies the uncrossing rule $\mathbf{X}_r \cap \mathbf{X}_f$ *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 ($P(\mathbf{Y}|\mathbf{X})$) and a system of sub-regressions (without the response variable: $P(\mathbf{X}_r|\mathbf{X}_f, S)$). 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 \mathbf{Y} so it can be learnt independently, even with a larger dataset (if missing values in \mathbf{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: $\mathbf{X}_r = \mathbf{x}_3$, $\mathbf{X}_f = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5\}$, $p_r = 1$ and $\boldsymbol{\alpha}_3 = (1, 1, 0, 0)'$

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

Now we know $P(\mathbf{X}_r|\mathbf{X}_f, S)$ by the structure of sub-regressions, we are able to define a marginal regression model $P(\mathbf{Y}|\mathbf{X}_f, S)$ 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 $\mathbf{X} = [\mathbf{X}_f, \mathbf{X}_r]$ we can rewrite (2.1):

$$\mathbf{Y}_{|\mathbf{X}_f, \mathbf{X}_r, S} = \mathbf{X}_f \boldsymbol{\beta}_f + \mathbf{X}_r \boldsymbol{\beta}_r + \boldsymbol{\varepsilon}_Y \quad (3.2)$$

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

$$P(\mathbf{Y}|\mathbf{X}_f) = \int_{\mathbf{X}_r} P(\mathbf{Y}|\mathbf{X}_r, \mathbf{X}_f) P(\mathbf{X}_r|\mathbf{X}_f) d\mathbf{X} \quad (3.3)$$

$$\mathbf{Y}_{|\mathbf{X}_f, S} = \mathbf{X}_f (\boldsymbol{\beta}_f + \sum_{j \in I_r} \beta_j \boldsymbol{\alpha}_j) + \sum_{j \in I_r} \beta_j \boldsymbol{\varepsilon}_j + \boldsymbol{\varepsilon}_Y \quad (3.4)$$

$$= \mathbf{X}_f \boldsymbol{\beta}_f^* + \boldsymbol{\varepsilon}_Y^* \quad (3.5)$$

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 \mathbf{X}_f are independent so the matrix $\mathbf{X}_f' \mathbf{X}_f$ 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 $\mathbf{Y}_{|\mathbf{X}}$. 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 \mathbf{X}_f because of the correlations given by S . So we also get the model

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

for which OLS estimator of the coefficients may be biased.

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

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 $\mathbf{X}_r = \mathbf{0}$

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 $\hat{\boldsymbol{\beta}}_r^* = \mathbf{0}$ (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 \mathbf{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 \mathbf{X} defined in section 2.1 with the estimator obtained by the pre-treatment that is \mathbf{X}_f selection.

For the marginal regression model defined in (3.5) we have the OLS unbiased estimator of β^* :

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

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

$$\mathbb{E}[\hat{\beta}_f^* | \mathbf{X}_f] = \beta_f + \sum_{j \in I_r} \beta_j \alpha_j \text{ and } \mathbb{E}[\hat{\beta}_r^* | \mathbf{X}_f] = \mathbf{0} \quad (3.8)$$

with variance:

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

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 $\hat{\beta}$ is:

$$\text{MSE}(\hat{\beta} | \mathbf{X}) = \|\text{Bias}\|_2^2 + \text{Tr}(\text{Var}(\hat{\beta})) \quad (3.10)$$

$$\text{MSE}(\hat{\beta}_{OLS} | \mathbf{X}) = 0 + \sigma_Y^2 \text{Tr}((\mathbf{X}' \mathbf{X})^{-1}) \quad (3.11)$$

$$\text{MSE}(\hat{\beta}_{OLS}^* | \mathbf{X}) = \left\| \sum_{j \in I_r} \beta_j \alpha_j \right\|_2^2 + \|\beta_r\|_2^2 + (\sigma_Y^2 + \sum_{j \in I_2} \sigma_j^2 \beta_j^2) \text{Tr}((\mathbf{X}_f' \mathbf{X}_f)^{-1}) \quad (3.12)$$

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:

$$1 - \mathcal{R}^2 = \frac{\text{Var}(\varepsilon)_3}{\text{Var}(\mathbf{x}_3)} = \frac{\sigma_3^2}{\sigma_3^2 + 2} \quad (3.13)$$

It is clear in Figure 3.4 that the marginal model is more robust than OLS on \mathbf{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 $\text{Tr}(\text{Var}(\hat{\beta})) = \text{Tr}(\text{Var}(\hat{\beta}_f)) + \text{Tr}(\text{Var}(\hat{\beta}_r))$ for both models but for the marginal model $\text{Tr}(\text{Var}(\hat{\beta}_r)) = 0$. 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 4.4 and ??.



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 How to compare structures ?

4.1.1 Bayesian criterion for quality

4.1.2 Some indicators for proximity

The first criterion is $\psi(\mathbf{X}, S)$ 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 \mathbf{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 α_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}

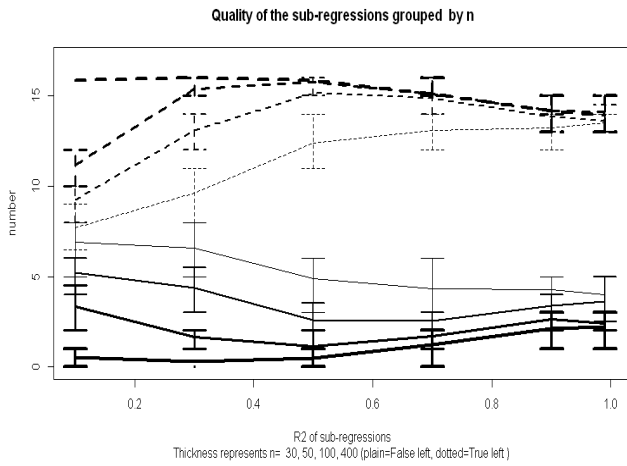


Figure 4.1: Quality of the subregressions found with classical BIC criterion

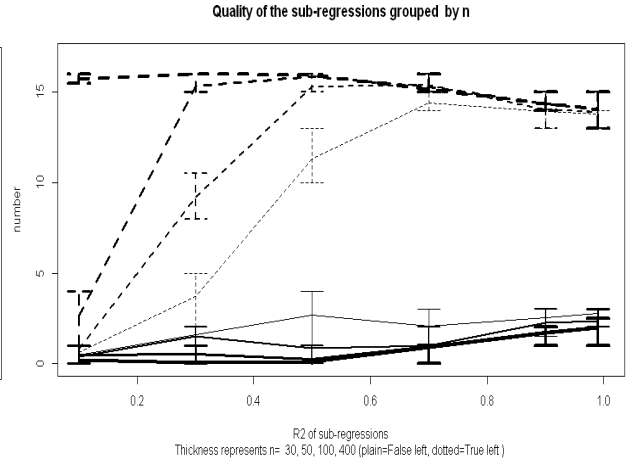


Figure 4.2: Quality of the subregressions found with our BIC_+ criterion

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 complexities between LASSO and CorReg+LASSO



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



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



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



Figure 4.10: Comparison of the complexities 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 ??.

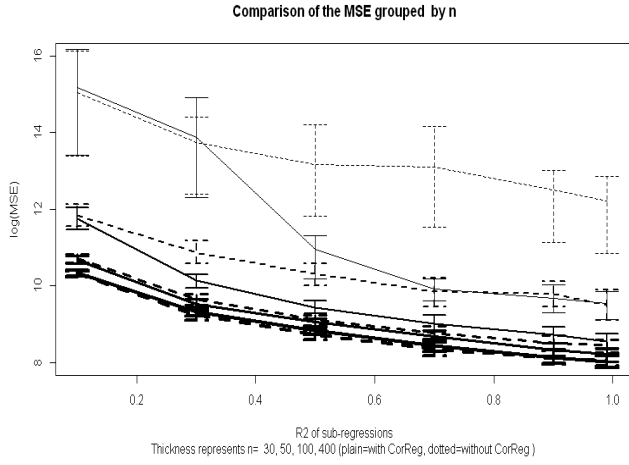


Figure 4.11: Comparison of the MSE between OLS and CorReg+OLS

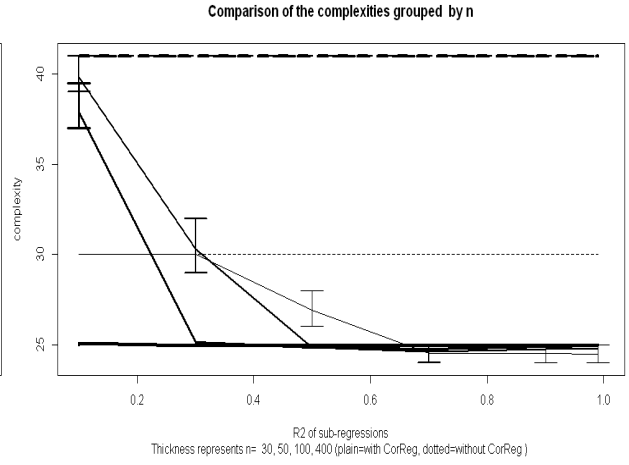


Figure 4.12: Comparison of the complexities between OLS and CorReg+OLS

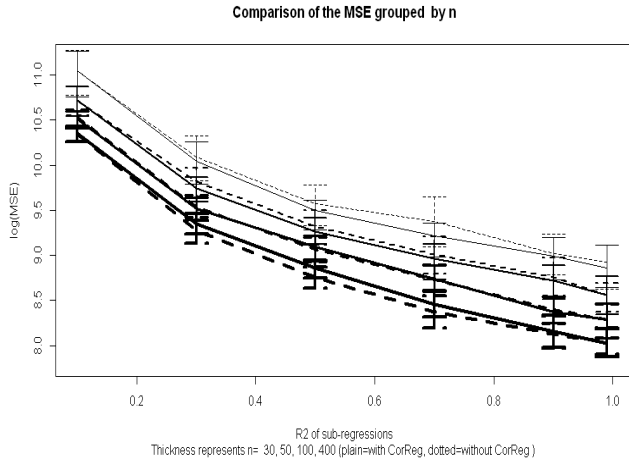


Figure 4.13: Comparison of the MSE between LASSO and CorReg+LASSO

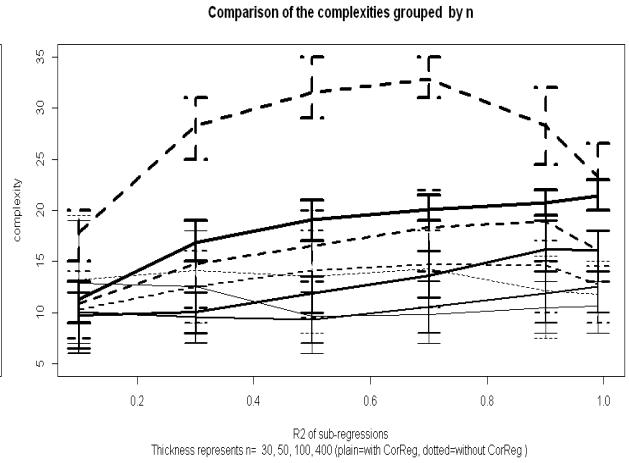


Figure 4.14: Comparison of the complexities between LASSO and CorReg+LASSO

We see that CorReg tends to give more parsimonious models and better predictions, even if the true model is not parsimonious. 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 chosen).

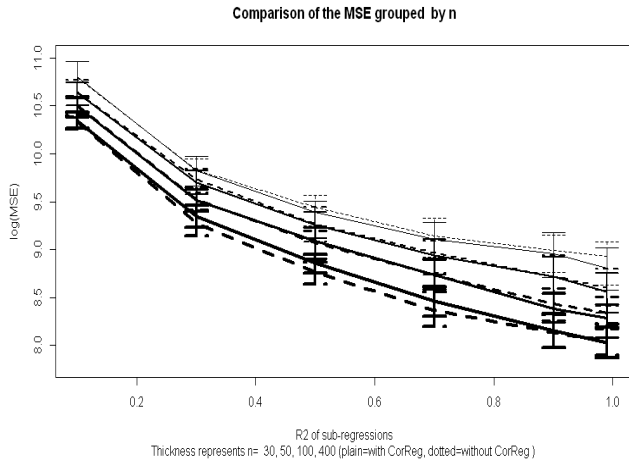


Figure 4.15: Comparison of the MSE between elasticnet and CorReg+elasticnet

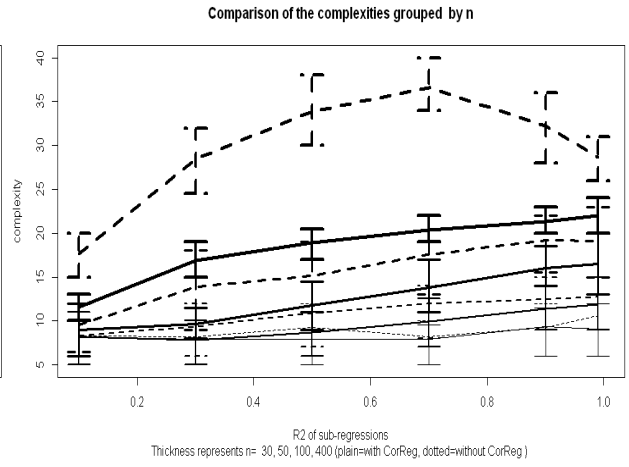


Figure 4.16: Comparison of the complexities between elasticnet and CorReg+elasticnet

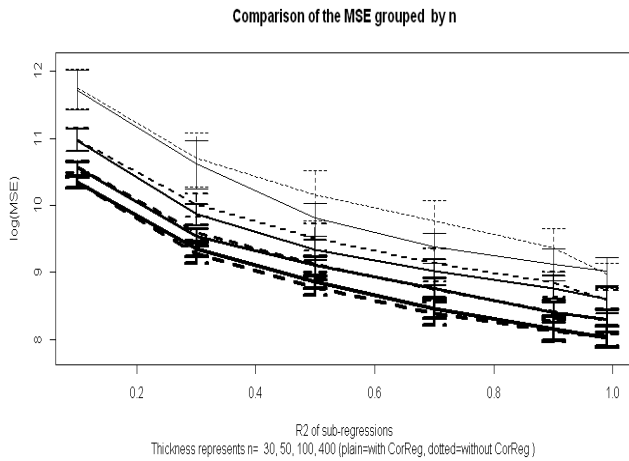


Figure 4.17: Comparison of the MSE between stepwise and CorReg+stepwise

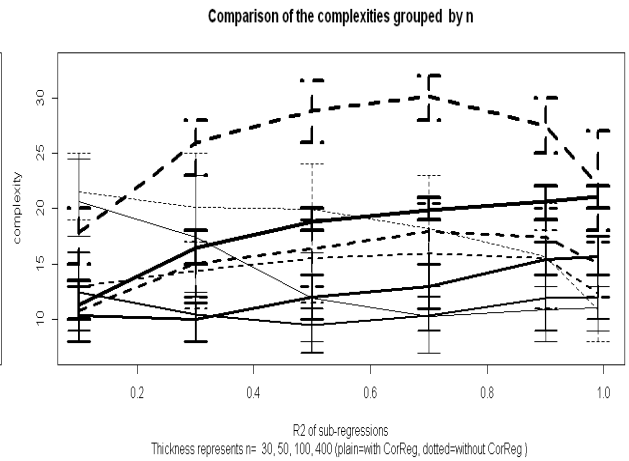


Figure 4.18: Comparison of the complexities between stepwise and CorReg+stepwise

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 X_r . The datasets used here were still those from ???. Depending only on X_r implies sparsity and impossibility to obtain the true model when using the true structure.

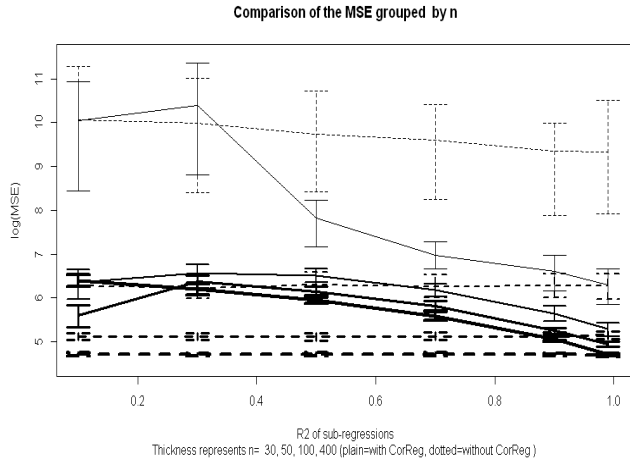


Figure 4.19: Comparison of the MSE between OLS and CorReg+OLS

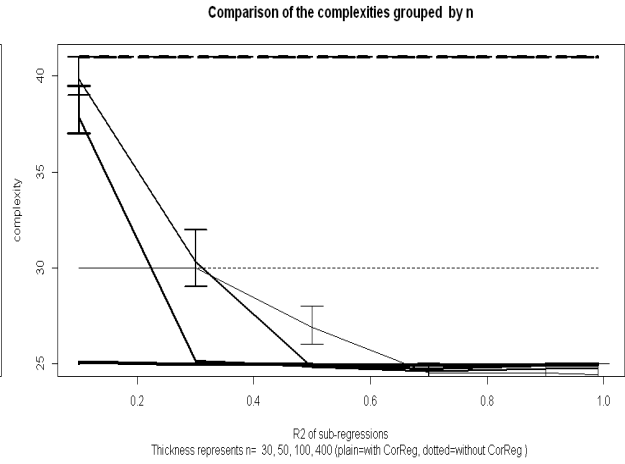


Figure 4.20: Comparison of the complexities between OLS and CorReg+OLS

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

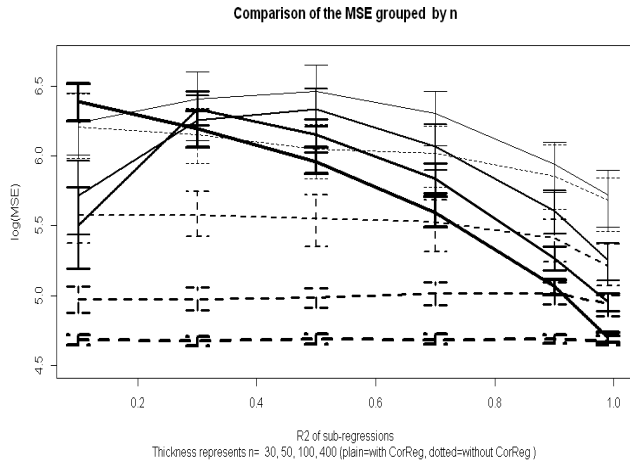


Figure 4.21: Comparison of the MSE between LASSO and CorReg+LASSO

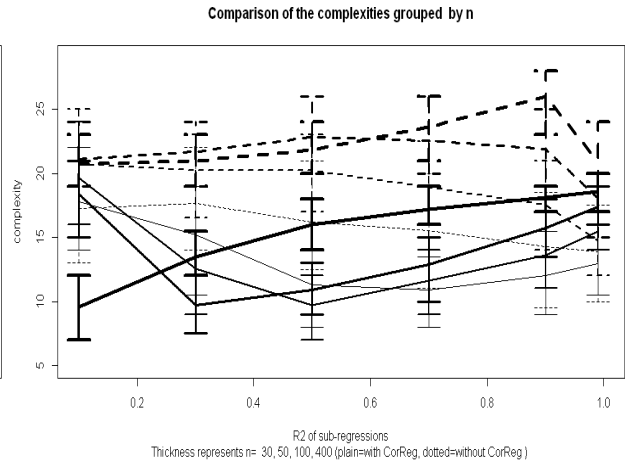


Figure 4.22: Comparison of the complexities between LASSO and CorReg+LASSO



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

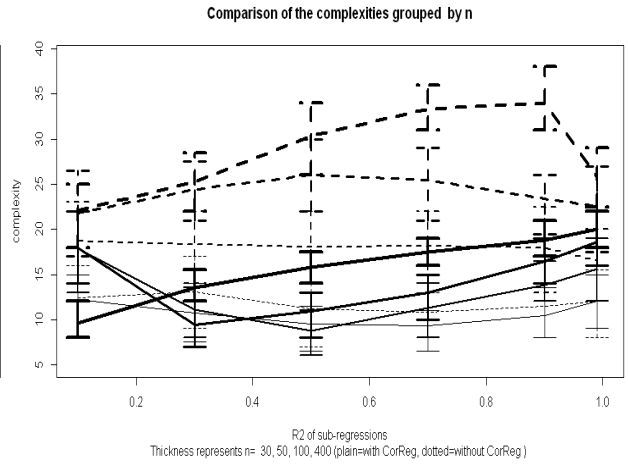


Figure 4.24: Comparison of the complexities between elasticnet and CorReg+elasticnet



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

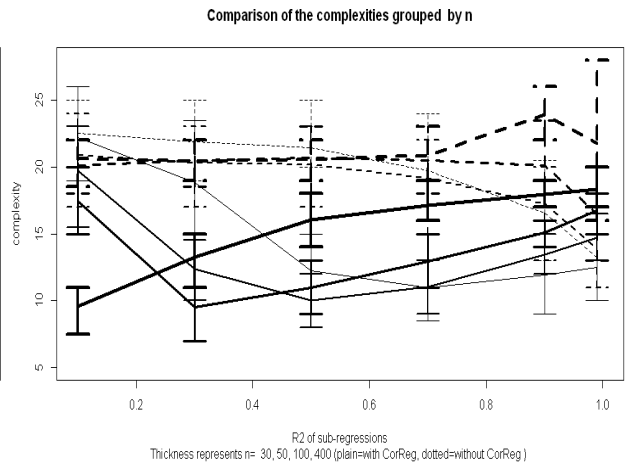


Figure 4.26: Comparison of the complexities between stepwise and CorReg+stepwise

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

5.1 State of the art

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

5.2 How to manage missing values in the MCMC ?

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

5.2.1 Decomposition of the integrated likelihood

We start with the complete likelihood

$$L(\alpha, \Theta, S; \mathbf{X}) = \prod_{i=1}^n f(\mathbf{X}_i) = \prod_{i=1}^n [f(\mathbf{X}_i^{I_r} | \mathbf{X}_i^{I_f}, \alpha, \Theta, S) f(\mathbf{X}_i^{I_f} | \alpha, \Theta, S)] \quad (5.1)$$

$$= \prod_{i=1}^n \left[\prod_{j \in I_r} f(x_{i,j} | \mathbf{X}_i^{I_f}, \alpha, \Theta, S) \prod_{j \notin I_r} f(x_{i,j} | \alpha, \Theta, S) \right] \quad (5.2)$$

$$= \prod_{i=1}^n \left[\prod_{j \in I_r} f(x_{i,j} | \mathbf{X}_i^{I_f^j}, \alpha, \Theta, S) \prod_{j \notin I_r} f(x_{i,j} | \alpha, \Theta, S) \right] \quad (5.3)$$

$$\mathcal{L}(\alpha, \Theta, S; \mathbf{X}) = \sum_{i=1}^n \left[\sum_{j \in I_r} \log \left(f(x_{i,j} | \mathbf{X}_i^{I_f^j}, \alpha, \Theta, S) \right) + \sum_{j \notin I_r} \log (f(x_{i,j} | \alpha, \Theta, S)) \right] \quad (5.4)$$

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 \mathbf{X}_M .

We know that \mathbf{X} is a Gaussian mixture with K components.

$$L(\boldsymbol{\alpha}, \Theta, S; \mathbf{X}_0) = \int_{\mathbf{X}_M} L(\boldsymbol{\alpha}, \Theta, S; \mathbf{X}) d\mathbf{X} \quad (5.5)$$

$$= \int_{\mathbf{X}_M} \sum_{k=1}^K \pi_k \phi_k(\mathbf{X}, \boldsymbol{\alpha}, \Theta, S) d\mathbf{X} \quad (5.6)$$

$$= \sum_{k=1}^K \pi_k \int_{\mathbf{X}_M} \phi_k(\mathbf{X}, \boldsymbol{\alpha}, \Theta, S) d\mathbf{X} \quad (5.7)$$

$$= \sum_{k=1}^K \pi_k \int_{\mathbf{X}_M} \prod_{i=1}^n \phi_k(\mathbf{X}_i, \boldsymbol{\alpha}, \Theta, S) d\mathbf{X} \quad (5.8)$$

$$= \sum_{k=1}^K \pi_k \prod_{i=1}^n \int_{\mathbf{X}_{i,M}} \phi_k(\mathbf{X}_i, \boldsymbol{\alpha}, \Theta, S) d\mathbf{X}_i \quad (5.9)$$

$$= \sum_{k=1}^K \pi_k \prod_{i=1}^n \phi_k(\mathbf{X}_{i,O}, \boldsymbol{\alpha}, \Theta, S) \quad (5.10)$$

$$= \sum_{k=1}^K \pi_k \phi_k(\mathbf{X}_O, \boldsymbol{\alpha}, \Theta, S) = f(\mathbf{X}_O, \boldsymbol{\alpha}, \Theta, S) \quad (5.11)$$

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

$$\int_{\mathbf{X}_M} f(x_{i,j} | \boldsymbol{\alpha}, \Theta, S) d\mathbf{X} = \sum_{k=1}^{k_j} \pi_{j,k} \Phi(x_{i,j} | \mu_{j,k}, \Sigma_{j,k}) \quad (5.12)$$

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

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

$$\int_{\mathbf{X}_M} f(x_{i,j} | \mathbf{X}_i^{I_f^j}, \boldsymbol{\alpha}, \Theta, S) d\mathbf{X} = \sum_{k=1}^{K_{ij}} \pi_{ij,k} \Phi(x_{i,j}, \mu_{ij,k}, \Sigma_{ij,k}) \text{ where} \quad (5.13)$$

$$\boldsymbol{\pi}_{ij} = \bigotimes_{\substack{l \in I_f^j \\ \mathbf{M}_{i,l}=1}} \boldsymbol{\pi}_l \text{ and } K_{ij} = |\boldsymbol{\pi}_{ij}|, \quad (5.14)$$

$$\boldsymbol{\mu}_{ij} = \sum_{\substack{l \in I_f^j \\ \mathbf{M}_{i,l}=0}} \alpha_{l,j} x_{i,l} + \bigoplus_{\substack{l \in I_f^j \\ \mathbf{M}_{i,l}=1}} \alpha_{l,j} \boldsymbol{\mu}_l \quad (5.15)$$

$$\boldsymbol{\Sigma}_{ij} = \sigma_j^2 + \bigoplus_{\substack{l \in I_f^j \\ \mathbf{M}_{i,l}=1}} \alpha_{i,l}^2 \boldsymbol{\Sigma}_l \quad (5.16)$$

This can be easily used for imputation of the missing values in \mathbf{X}^{I_r} knowing the parameters $\boldsymbol{\alpha}, \Theta$ and S . To estimate $\boldsymbol{\alpha}$ we use an EM algorithm. We start with an arbitrary value $\boldsymbol{\alpha}^{(0)}$, then: For the iteration h of the algorithm at the E step we want

$$E_{\mathbf{X}_M | \mathbf{X}_O, \boldsymbol{\alpha}, \Theta, S} [\mathcal{L}(\mathbf{X} | \boldsymbol{\alpha}^{(h)}, S, \Theta)] \quad (5.17)$$

So we get $\mathbf{X}_M^{(h)}$ the imputation for \mathbf{X}_M and then the M step simply is

$$\boldsymbol{\alpha}^{(h+1)} = \operatorname{argmax}_{\boldsymbol{\alpha}} E[\mathcal{L}(\mathbf{X} | \boldsymbol{\alpha}, S, \Theta)] \quad (5.18)$$

and we can use the same method as the one for classical case without missing values (OLS, SUR, *etc.*). And we continue until convergence ($\|\boldsymbol{\alpha}^{(h+1)} - \boldsymbol{\alpha}^{(h)}\| < \text{tol}$ where tol is the tolerance).

Imputation in \mathbf{X}^{I_r} is made according to (5.13) and we have orthogonality in \mathbf{X}^{I_f} :

$$\forall j \notin I_r, P(\mathbf{X}^j | \mathbf{X}^{I_r}, \mathbf{X}^{I_f} \setminus \mathbf{X}^j) = P(\mathbf{X}^j | \mathbf{X}^{I_r})$$

Moreover, $\forall j \notin I_r, \forall l \in I_r$ with $j \notin I_f^l, \mathbf{X}^j \perp \mathbf{X}^l$ so

$$\forall j \notin I_r, P(\mathbf{X}^j | \mathbf{X}^{I_r}, \mathbf{X}^{I_f} \setminus \mathbf{X}^j) = P(\mathbf{X}^j | \mathbf{X}^{I_r}) = P(\mathbf{X}^j | \mathbf{X}^{I_f^j})$$

where $I_r^j = \{i \in I_r | j \in I_f^i\} = \{i \in I_r | \alpha_{j,i} \neq 0\}$ So we will make imputations (for E step only) according to $P(\mathbf{X}^j | \mathbf{X}_O^{I_r^j})$

$\forall 1 \leq i \leq n$ To use formulas on conditional distribution for Gaussian multivariate distribution we first write $P(\mathbf{X}_i^j, \mathbf{X}_i^{I_r^j})$ which is a Gaussian mixture with K_{ij} components.

$$\forall j \notin I_r, \forall (l_1, l_2) \in I_r^j, P(x_{i,l_1} | x_{i,j}, x_{i,l_2}) = P(x_{i,l_1} | x_{i,j})$$

$$P(x_{i,j}, \mathbf{X}_{i,O}^{I_r^j}, \boldsymbol{\alpha}) = P(x_{i,j}) \prod_{\substack{l \in I_r^j \\ M_{i,l}=0}} P(x_{i,l} | x_{i,j}) \quad (5.19)$$

$$= \sum_{k=1}^{K_{ij}} \pi_{ij,k} \phi(x_{i,j}, \mathbf{X}_{i,O}^{I_r^j}, \boldsymbol{\mu}_{ij,k}, \boldsymbol{\Sigma}_{ij,k}) \text{ where} \quad (5.20)$$

$$\boldsymbol{\pi}_{ij} = \boldsymbol{\pi}_j \otimes \left[\bigotimes_{\substack{l \in I_r^j \\ M_{i,l}=0}} \boldsymbol{\pi}_{ijl} \right] = \boldsymbol{\pi}_j \otimes \bigotimes_{\substack{l \in I_r^j \\ M_{i,l}=0}} \left(\bigotimes_{\substack{h \in I_f^l \\ h \neq j}} \boldsymbol{\pi}_h \right) \text{ and } K_{ij} = |\boldsymbol{\pi}_{ij}|, \quad (5.21)$$

$$\boldsymbol{\mu}_{ij} = \boldsymbol{\mu}_j \times \left[\prod_{\substack{l \in I_r^j \\ M_{i,l}=0}} \boldsymbol{\mu}_{ijl} \right] = \boldsymbol{\mu}_j \times \prod_{\substack{l \in I_r^j \\ M_{i,l}=0}} \left[\alpha_{j,l} x_{i,j} + \bigoplus_{h \in I_f^l} \alpha_{h,l} \boldsymbol{\mu}_h \right] \quad (5.22)$$

$$\boldsymbol{\Sigma}_{ij} \text{ is the associated variance-covariance matrix} \quad (5.23)$$

Cartesian product and power in the expression of the mean.

Then we have

$$P(x_{i,j} | \mathbf{X}_{i,O}^{I_r^j}, \boldsymbol{\alpha}) = \sum_{k=1}^{K_{ij}} \pi_{ij,k} \phi(x_{i,j} | \mathbf{X}_{i,O}^{I_r^j}, \boldsymbol{\mu}_{ij,k}, \boldsymbol{\Sigma}_{ij,k}) \quad (5.24)$$

$$= \sum_{k=1}^{K_{ij}} \pi_{ij,k} \phi \left(x_{i,j}, \boldsymbol{\mu}_{ij,k | \mathbf{X}_{i,O}^{I_r^j}}, \boldsymbol{\Sigma}_{ij,k | \mathbf{X}_{i,O}^{I_r^j}} \right) \text{ with} \quad (5.25)$$

$$\boldsymbol{\mu}_{ij,k | \mathbf{X}_{i,O}^{I_r^j}} = \boldsymbol{\mu}_{j,k} + \text{cov}(x_{i,j,k}, \mathbf{X}_{i,O,k}^{I_r^j}) \text{var}(\mathbf{X}_{i,O,k}^{I_r^j})^{-1} (\mathbf{X}_{i,O}^{I_r^j} - \boldsymbol{\mu}_{\mathbf{X}_{i,O}^{I_r^j}}) \text{ and} \quad (5.26)$$

$$\boldsymbol{\Sigma}_{ij,k | \mathbf{X}_{i,O}^{I_r^j}} = \text{var}(x_{i,j,k}) - \text{cov}(x_{i,j,k}, \mathbf{X}_{i,O,k}^{I_r^j}) \text{var}(\mathbf{X}_{i,O,k}^{I_r^j})^{-1} \text{cov}(\mathbf{X}_{i,O,k}^{I_r^j}, x_{i,j,k}) \quad (5.27)$$

But we do not need to compute the variance because we only want $E_{\mathbf{X}_M | \mathbf{X}_O, \boldsymbol{\alpha}, \Theta, S} [\mathcal{L}(\mathbf{X} | \boldsymbol{\alpha}^{(h)}, S, \Theta)]$ so the mean is sufficient, we impute

$$\hat{x}_{i,j} = \frac{1}{K_{ij}} \sum_{k=1}^{K_{ij}} \pi_{ij,k} \boldsymbol{\mu}_{ij,k | \mathbf{X}_{i,O}^{I_r^j}} \quad (5.28)$$

We have

$$\forall l \in I_r^j, \forall 1 \leq k \leq K_{ij}, \text{cov}(x_{i,j,k}, \mathbf{X}_{i,O,k}^l) = \text{cov}(x_{i,j,k}, \sum_{h \in I_f^l} x_{i,h} \alpha_{h,l} + \varepsilon_{i,l}) = \alpha_{j,l} \sigma_{j,k}^2 \quad (5.29)$$

$$\forall l \in I_r^j, \forall 1 \leq k \leq K_{ij}, \text{var}(x_{i,l,k}) = \sigma_l^2 + \sum_{h \in I_f^l} \alpha_{h,l}^2 \sigma_{h,k}^2 \quad (5.30)$$

$$\forall l_1 \neq l_2 \in I_r^j, \forall 1 \leq k \leq K_{ij},$$

$$\text{cov}(x_{i,l_1,k}, x_{i,l_2,k}) = \text{cov}\left(\sum_{h \in I_f^{l_1}} x_{i,h,k} \alpha_{h,l_1} + \varepsilon_{i,l_1}, \sum_{h \in I_f^{l_2}} x_{i,h,k} \alpha_{h,l_2} + \varepsilon_{i,l_2}\right) \quad (5.31)$$

$$= \text{cov}\left(\sum_{h \in I_f^{l_1}} x_{i,h,k} \alpha_{h,l_1}, \sum_{h \in I_f^{l_2}} x_{i,h,k} \alpha_{h,l_2}\right) \quad (5.32)$$

$$= \sum_{h \in I_f^{l_1} \cap I_f^{l_2}} \sigma_{h,k}^2 \alpha_{h,l_1} \alpha_{h,l_2} \quad (5.33)$$

5.2.2 Alternative E step

If we can't (or don't want to) compute the E step described above, then we can use alternative imputation step for missing data based on α (and keep the alternate optimisation to find the best α).

$\forall x_{i,j} \in \mathbf{X}_M$ we have:

if $j \in I_r$, Equation(5.13) gives:

$$E[x_{i,j} | \alpha^{(h)}, \Theta, \mathbf{X}_O, S] = E\left[\sum_{k=1}^{k_{ij}} \pi_{ij,k} \Phi(x_{i,j} | \mu_{ij,k}, \Sigma_{ij,k}) | \alpha^{(h)}, \Theta, \mathbf{X}_O, S\right] \quad (5.34)$$

Let $r_{i,j} = \{l \in I_r | \alpha_{j,l} \neq 0, \mathbf{M}_{i,j} = 0\}$ 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:

$$E[x_{i,j} | \alpha^{(h)}, \Theta, \mathbf{X}_O, S] = \frac{1}{|r_{i,j}|} \sum_{k \in r_{i,j}} E_{|\alpha^{(h)}, \Theta, \mathbf{X}_O, S} \left[\frac{1}{\alpha_{j,k}} \left(x_{i,k} - \varepsilon_k(i) - \sum_{l \in I_f^k} x_{i,l} \alpha_{l,k} \right) \right] \quad (5.35)$$

$$= \frac{1}{|r_{i,j}|} \sum_{k \in r_{i,j}} E_{|\alpha^{(h)}, \Theta, \mathbf{X}_O, S} \left[\frac{1}{\alpha_{j,k}} \left(x_{i,k} - \sum_{l \in I_f^k} x_{i,l} \alpha_{l,k} \right) \right] \quad (5.36)$$

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

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

5.3 SEM

We use a SEM [Celeux and Diebolt, 1986] to estimate α . At iteration h ,

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

M step: We estimate

$$\alpha^{(h+1)} = \text{argmax}_{\alpha} E[\mathcal{L}(\mathbf{X} | \alpha, S, \Theta)] \quad (5.37)$$

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

5.3.1 SE step in details

We use a Gibbs sampling method to generate the missing values. \mathbf{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 \mathbf{X}_M are imputed by the marginal means. At each iteration of the Gibbs sampler: $\forall x_{i,j} \in \mathbf{X}_M^{I_r}$: $x_{i,j}$ is generated according to

$$P(x_{i,j} | \mathbf{X}_{i,O}, \mathbf{X}_{i,\bar{M}_{i,j}}, Z; \boldsymbol{\alpha}^{(h)}, \Theta, S) = P(x_{i,j} | \mathbf{X}_{i,O}, \mathbf{X}_{i,\bar{M}_{i,j}}; \boldsymbol{\alpha}^{(h)}, \Theta, S) \quad (5.38)$$

$$= P(x_{i,j} | \mathbf{X}_i^{I_f^j}; \boldsymbol{\alpha}^{(h)}, \Theta, S) = \mathcal{N}(\mathbf{X}_i^{I_f^j} \boldsymbol{\alpha}_{I_f^j}^{(h)}; \sigma_j^2) \quad (5.39)$$

We have $P(\mathbf{X} | Z) = \mathcal{N}(\boldsymbol{\mu}_{|Z}, \boldsymbol{\Sigma}_{|Z})$.

$\forall x_{i,j} \in \mathbf{X}_M^{I_f}$: $x_{i,j}$ is generated according to

$$P(x_{i,j} | \mathbf{X}_{i,O}, \mathbf{X}_{i,\bar{M}_{i,j}}, Z; \boldsymbol{\alpha}^{(h)}, \Theta, S) = P(x_{i,j} | \mathbf{X}_{i,\bar{j}}, Z_i; \boldsymbol{\alpha}^{(h)}, \Theta, S) \quad (5.40)$$

$$= \mathcal{N}(\mu_{j|Z_i,j} + \boldsymbol{\Sigma}_{j,X_{\bar{i}j}|Z_i} \boldsymbol{\Sigma}_{X_{\bar{i}j},X_{\bar{i}j}|Z_i}^{-1} (X_{\bar{i}j} - \boldsymbol{\mu}_{X_{\bar{i}j}|Z_i}); \sigma_{j|Z_i,j}^2 - \boldsymbol{\Sigma}_{j,X_{\bar{i}j}|Z_i} \boldsymbol{\Sigma}_{X_{\bar{i}j},X_{\bar{i}j}|Z_i}^{-1} \boldsymbol{\Sigma}_{j,X_{\bar{i}j}|Z_i}') \quad (5.41)$$

With $\forall j \in I_r$

$$\text{var}_{|Z_i}(x_{i,j}) = \sigma_{j|Z_i,j}^2 + \sum_{l \in I_f^j} \alpha_{l,j}^2 \sigma_{l|Z_i,l}^2 = \sigma_j^2 + \sum_{l \in I_f^j} \alpha_{l,j}^2 \sigma_{l|Z_i,l}^2 \quad (5.42)$$

$\forall j \notin I_r$

$$\text{var}_{|Z_i}(x_{i,j}) = \sigma_{j|Z_i,j}^2 \quad (5.43)$$

$\forall j_1 \in I_r, j_2 \in I_r, I_f^{j_1} \cap I_f^{j_2} \neq \emptyset$

$$\text{cov}_{|Z_i}(x_{i,j_1}, x_{i,j_2}) = \sum_{k \in I_f^{j_1} \cap I_f^{j_2}} \alpha_{k,j_1} \alpha_{k,j_2} \text{var}_{|Z_i}(x_{k|Z_i,k}) = \sum_{k \in I_f^{j_1} \cap I_f^{j_2}} \alpha_{k,j_1} \alpha_{k,j_2} \sigma_{k|Z_i,k}^2 \quad (5.44)$$

$\forall j_1 \in I_r, j_2 \in I_r, I_f^{j_1} \cap I_f^{j_2} = \emptyset$

$$\text{cov}_{|Z_i}(x_{i,j_1}, x_{i,j_2}) = 0 \quad (5.45)$$

$\forall j_1 \in I_f, j_2 \in I_f$

$$\text{cov}_{|Z_i}(x_{i,j_1}, x_{i,j_2}) = 0 \quad (5.46)$$

$\forall j_1 \in I_r, j_2 \in I_f^{j_1}$

$$\text{cov}_{|Z_i}(x_{i,j_1}, x_{i,j_2}) = \alpha_{j_2,j_1} \sigma_{j_2|Z_i,j_2}^2 \quad (5.47)$$

$\forall j_1 \in I_r, j_2 \notin I_f^{j_1} \cup I_r$

$$\text{cov}_{|Z_i}(x_{i,j_1}, x_{i,j_2}) = 0 \quad (5.48)$$

We see that the 0 in the variance-covariance matrix does not depend on Z 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.

Then, $\forall Z_{i,j}$

We see that $Z_{i,j}$ are not used if there is no missing values in \mathbf{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.

5.3.2 Estimation of the coefficients in each regression

Estimating the α_j with missing values is just estimating independent regressions with missing values. We have seen in equation (5.13) that we know the expression of this density for a given the α_j . So it's just about maximizing the likelihood of this density on the α_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):

$$\mathbf{X}^{(h)} = E[\mathbf{X}|\mathbf{X}_O, \alpha^{(h)}, \varepsilon, \Theta, S] \quad (5.49)$$

step M:

$$\alpha^{(h+1)} = \operatorname{argmax}_{\alpha}(\mathcal{L}(\mathbf{X}^{(h)}, \alpha, \varepsilon, \Theta, S)) \text{ by OLS} \quad (5.50)$$

But estimation of the α_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 α with full sub-matrices of \mathbf{X}_f . When relying on too much missing values, $\hat{\alpha}$ 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 α_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 α^j only depends on individuals not missing in \mathbf{X}^j (individuals are *iid*). So we work with a restriction of \mathbf{X} for each α^j . Thus in this section, to simplify the notation, we will consider no missing values in \mathbf{X}_r but in fact we work with restrictions.

The EM algorithm can be written here: we start with some $\Theta^{(0)} = (\alpha, \varepsilon)$ 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

$$\mathbf{X}^{(h)} = E(\mathbf{X}|\mathbf{X}_{\bar{M}}, \Theta^{(h)}) \text{ so it simply is} \quad (5.51)$$

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

$$x_{i,j}^{(h)} = E(x_{i,j}|\mathbf{X}_{\bar{M}}, \Theta^{(h)}) = \sum_{k=1}^{k_{ij}} \pi_{ij,k} \mu_{ij,k}^{(h)} \quad (5.52)$$

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

$$E(\mathbf{X}|\Theta) = \mathbf{X}^{(h)} \text{ done by OLS} \quad (5.53)$$

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 $\bar{\mathbf{X}}_{i,j} = (\mathbf{X}_{\bar{M}} \cap \mathbf{X}_i \setminus \mathbf{X}^j)$

$$P(\mathbf{X}_{fi}^M, \mathbf{X}_{fi}^{\bar{M}}, \mathbf{X}_{ri}^{\bar{M}}|\Theta) = P(\mathbf{X}_{fi}^M|\mathbf{X}_{fi}^{\bar{M}}, \mathbf{X}_{ri}^{\bar{M}}, \Theta)P(\mathbf{X}_{fi}^{\bar{M}}, \mathbf{X}_{ri}^{\bar{M}}|\Theta) \quad (5.54)$$

$$= P(\mathbf{X}_{ri}^{\bar{M}}|\mathbf{X}_{fi}^M, \mathbf{X}_{fi}^{\bar{M}}, \Theta)P(\mathbf{X}_{fi}^M, \mathbf{X}_{fi}^{\bar{M}}|\Theta) \quad (5.55)$$

$$P(\mathbf{X}_{fi}^M|\mathbf{X}_{fi}^{\bar{M}}, \mathbf{X}_{ri}^{\bar{M}}, \Theta) = \frac{P(\mathbf{X}_{ri}^{\bar{M}}|\mathbf{X}_{fi}^M, \mathbf{X}_{fi}^{\bar{M}}, \Theta)P(\mathbf{X}_{fi}^M, \mathbf{X}_{fi}^{\bar{M}}|\Theta)}{P(\mathbf{X}_{fi}^{\bar{M}}, \mathbf{X}_{ri}^{\bar{M}}|\Theta)} \quad (5.56)$$

$$= \frac{P(\mathbf{X}_{ri}^O|\mathbf{X}_{fi}^M, \mathbf{X}_{fi}^{\bar{M}}, \Theta)P(\mathbf{X}_{fi}^M|\Theta)P(\mathbf{X}_{fi}^{\bar{M}}|\Theta)}{P(\mathbf{X}_{ri}^{\bar{M}}|\mathbf{X}_{fi}^{\bar{M}}, \Theta)P(\mathbf{X}_{fi}^{\bar{M}}|\Theta)} \quad (5.57)$$

$$= \frac{P(\mathbf{X}_{ri}^{\bar{M}}|\mathbf{X}_{fi}^M, \mathbf{X}_{fi}^{\bar{M}}, \Theta)P(\mathbf{X}_{fi}^M|\Theta)}{P(\mathbf{X}_{ri}^{\bar{M}}|\mathbf{X}_{fi}^{\bar{M}}, \Theta)} \quad (5.58)$$

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

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

5.4 Missing values in the main regression

5.4.1 explicative

The reduced model (explicative one) is just a linear regression without structure so estimating β is like estimating the α_j and the same methods can be used. An EM algorithm would be preferred 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 X^j by their means, we can use when $j \in I_r$ the conditionnal mean $E(X^j|X_f, \alpha^j)$ defined in 5.52.

5.4.2 predictive

If there are missing values in $X^j \in X_r$ a new possibility appears. Knowing S and α_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 6

Taking back the residuals

6.1 The model

6.2 Properties

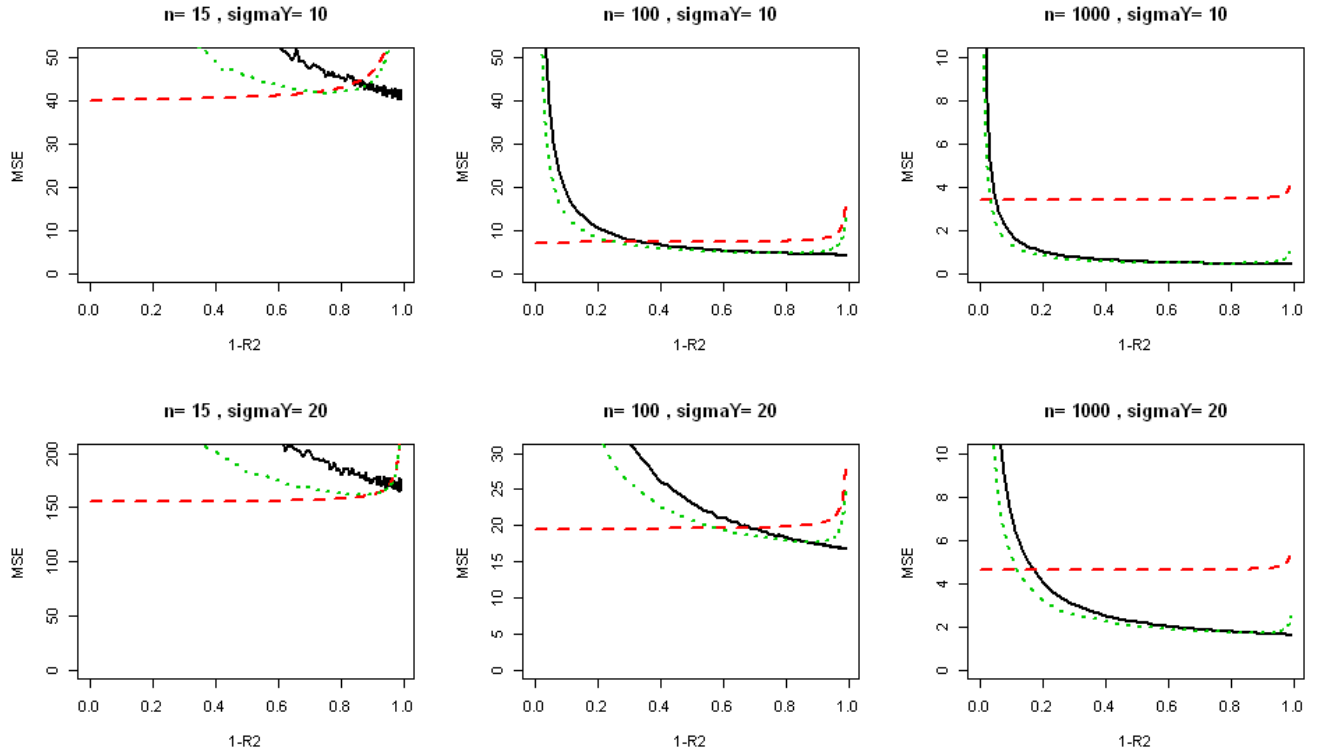


Figure 6.1: MSE of OLS (plain black) and CorReg marginal (red dashed) and CorReg full (green dotted) estimators for varying $(1 - R^2)$ of the sub-regression, n and σ_Y .

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. Here $p = 3$ and $n = 1000$. We define $\mathbf{X}_f, \mathbf{X}_r, \varepsilon_Y, \varepsilon_X i.i.d. \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ 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 with 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 interpret 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.

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.
- [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.
- [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*.
- [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.
- [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.
- [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.

- [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